Amazon Web Services

AWS RoboMaker

Run, scale, and automate robotics simulation.

How it works

AWS RoboMaker is a cloud-based simulation service that enables robotics developers to run, scale, and automate simulation without managing any infrastructure.

Use cases

Automated regression testing

Automate testing within a continuous integration and continuous delivery (CI/CD) pipeline.

Train reinforcement learning models

Train reinforcement learning models with high volumes of iterative trials.

Multi-robot testing

Connect multiple concurrent simulations to your fleet management software for testing.

AWS RoboMaker features

AWS RoboMaker makes it easy for developers to build robotic applications, and enables developers to run, scale, and automate simulation without managing any infrastructure.

Simulation

Building and testing applications for autonomous robots is challenging, complicated and resource intensive. Traditionally, teams of developers write code to address a wide array of deployment scenarios, the code is integrated, and then the application scenarios are tested on robotics hardware in physical environments. This manual process of development and testing consumes staff time, requires expensive hardware, slows the release cycle for application updates, and cannot be easily scaled.

With simulation, developers can test applications in virtual environments, or worlds, enabling them to increase testing coverage, reduce code errors, and accelerate development velocity. One of the most important benefits of using simulation is the ability to test and train robots in a large number and diverse set of worlds. Scaling simulation unlocks the ability to conduct robust regression testing, reinforcement learning, and synthetic data generation. However, for robotics companies, building simulation worlds, managing simulation infrastructure, and scaling testing is expensive and requires specialized skills. This often limits the use of simulation among robotics companies, and limits the benefits of automated testing at scale.

Simulation with AWS RoboMaker

AWS RoboMaker is a fully managed service that enables developers, QAs, and DevOps engineers to easily create simulation worlds and run simulation jobs without provisioning or managing any infrastructure. RoboMaker makes simulation at scale affordable and accessible to all robotics companies by providing tools for developers to test and iterate code in 3D virtual environments. The service supports large-scale and parallel simulations, and automatically scales based on the complexity of the scenarios being tested. With RoboMaker simulation, robotics companies can make robotics application testing and machine learning faster, less expensive, and more robust.

Run large-scale, parallel simulations

With the RoboMaker batch simulation API, developers can easily launch a large batch of simulations with a single API call. Running large-scale simulations through an API makes it easier to access compute power, allowing developers to increase the complexity, scale, and frequency of their tests, which speeds up the development lifecycle and makes testing more robust.

Use any simulator

RoboMaker Simulation can be used to run the robot software and simulator of choice, including ROS, custom robot applications, Gazebo, Unity, Unreal, and Nvidia Isaac-based simulations. RoboMaker enables you to run your simulations in the cloud without provisioning, configuring, or managing any infrastructure.

Simulation WorldForge

Build simulation worlds at scale

RoboMaker WorldForge makes it possible to automatically create hundreds of pre-defined, randomized simulation worlds that mimic real-world conditions without engineering investment or managing world generation infrastructure. Without a simulation world generator, building a

simulation world is costly, time consuming, and requires specialized skills in 3D modeling and knowledge of simulation engines. WorldForge makes it easy to increase the scale, speed, and variance of simulation, improving the quality of production code and accelerating time to market. Today, WorldForge provides worlds for indoor home environments, complete with configurable floor plans and furnishings.

What Is AWS RoboMaker?

AWS RoboMaker is a cloud-based simulation service with which robotics developers can run, scale, and automate simulation without managing any infrastructure. With AWS RoboMaker, robotics developers can cost-effectively scale and automate simulation workloads, run large-scale and parallel simulations with a single API call, and create user-defined, randomized 3D virtual environments. Using the simulation service, you can speed application testing and create hundreds of new worlds from templates you define.

AWS RoboMaker is capable of automated testing within a continuous integration and continuous delivery (CI/CD) pipeline, training reinforcement models with high volumes of iterative trials and connecting multiple concurrent simulations to your fleet management software for testing. When combined with AWS machine learning, monitoring, and analytics services, robots can stream data, navigate, communicate, comprehend, and learn.

AWS RoboMaker FAQs

Simulation

Q: What is AWS RoboMaker Simulation?

A: AWS RoboMaker Simulation is a fully managed service that enables you to easily create simulation worlds and run simulation jobs without provisioning or managing any infrastructure.

Q: What can I do with RoboMaker Simulation?

A: There are multiple use cases for RoboMaker Simulation. You can use RoboMaker Simulation to test robotics applications in an automated manner and to speed up iterative development. You can use it to generate synthetic image or lidar data for algorithm development. You can also use RoboMaker Simulation to train machine learning models using reinforcement learning.

Q: What are the key capabilities of RoboMaker Simulation?

A: RoboMaker Simulation provides fully managed compute infrastructure for running simulations at any scale. RoboMake Simulation WorldForge makes it possible to automatically create hundreds of pre-defined, randomized simulation worlds that replicate real-world conditions without engineering investment or managing world generation infrastructure.

Q: What simulation engines does RoboMaker Simulation run support?

A: RoboMaker Simulation supports any simulation engine of your choice. You can connect your container-based simulation application to Robomaker and run simulations without provisioning, configuring, or managing any infrastructure. The service supports simulation world creation in WorldForge for Gazebo-compatible world formats.

Q: What is a robot application in RoboMaker?

A: A robot application in RoboMaker refers to the robot middleware (such as ROS) based application that you will run on a physical robot. You need to build your robot application source code to X86 architecture in order to work with RoboMaker Simulation.

Q: What is a simulation application in RoboMaker?

A: A simulation application includes a 3D simulation world and simulator plugins that control the movement of a robot in a simulation world. Similar to a robot application, you need to build your simulation application source code to X86 architecture in order to work with RoboMaker Simulation.

Q: How do I get started with RoboMaker Simulation run?

A: You get started with RoboMaker Simulation run by creating a robot application and/or the simulation application. The applications link to your code stored in an Amazon S3 bucket. You can then create a simulation job to execute your simulation application and your robot application on RoboMaker’s fully managed infrastructure.

Q: What is a simulation job?

A: A simulation job is the main service resource in RoboMaker Simulation run. You run your simulation workloads on RoboMaker by creating a simulation job, which automatically provisions and manages the compute infrastructure for you.

Q: What is a simulation job batch?

A: A simulation job batch is a collection of simulation job requests. You can create a new simulation job batch using AWS RoboMaker's API. A simulation job batch allows you to submit multiple simulation job requests through a single API call and provides queuing capability for all submitted requests.

Q: What tools do I use to build my robot application and simulation application?

A: You can use the container tools that you are already familiar with to build and package your code into Open Container Initiative (OCI) compliant images and store in Amazon Elastic Conatiner Registry (ECR). You can use those images to create a RoboMaker robot application and simulation application.

Q: What is application versioning?

A: RoboMaker supports versioning of robot applications and simulation applications so you can control which implementation your robots and simulations use. A version is a numbered snapshot of your robot application or simulation application you can create for use in different parts of your development workflow, such as development, beta deployment, and production.

Q: Can I delete a particular version of a robot application or simulation application?

A: Yes, you can delete a particular version of an application or all the versions of an application.

Q: Does RoboMaker store my robot application and simulation application?

A: No, RoboMaker does not store your robot application or simulation application in the service. You upload your applications to your S3 bucket and refer to the S3 object path during robot application and simulation application creation.

Q: What is the failure behavior of a simulation job?

A: You can configure the failure behavior of your simulation job to be fail or continue. In fail mode, your simulation job will fail, and the underlying infrastructure will be released if there is any failure during the simulation job run. In continue mode, your simulation job will fail, but the underlying infrastructure will be kept if there is any failure during the simulation job run. The continue mode allows you to further troubleshoot a simulation job failure by using the command line tool. In continue mode, RoboMaker Simulation charges will continue to incur until you terminate the job.

Q: When do I restart a simulation job?

A: Restarting a simulation job can be used for quick iterative development. Restarting a simulation job is faster than creating a new simulation job because it doesn't require resource provisioning and configuration behind the scenes. During robotics application development, you can use the restart simulation job feature to quickly test your code as you iterate through code changes.

Q: How is restarting a simulation job different from cloning a simulation job?

A: Restarting a simulation job is used for quick iteration test and the updated simulation job runs on the same infrastructure resources provisioned and managed by RoboMaker. Cloning a simulation job uses the same configurations of the existing job to create a new job. It triggers infrastructure resource provisioning and configuration behind the scenes, which might take a few minutes.

Q: Why do I need to provide an AWS AWS Identity and Access Management (IAM) role in a simulation job?

A: The IAM role provides RoboMaker Simulation access to the resources you specified in your simulation job. For example, your S3 bucket for RoboMaker to write simulation logs into.

Q: What is simulation duration?

A: Simulation duration defines how long the simulation job should run until it gets terminated. The duration is based on wall clock time versus simulation time.

Q: Does my simulation job run in real-time?

A: RoboMaker Simulation attempts to run your simulation job as fast as possible. You can also use the "real-time update rate" and "max step size" settings in your Gazebo world configuration to control how fast the simulation job should run.

Q: How do I get charged for RoboMaker Simulation?

A: You are charged for simulation WorldForge based on the number of simulation worlds you generate and export. You are charged for simulation run based on the duration of your simulation job and the CPU/Memory resources used by your simulation job.

Q: Does RoboMaker Simulation run charge based on wall clock time or simulation time?

A: Your simulation job can run faster, slower than real time, or at real time. RoboMaker Simulation charges based on wall clock time so the speed of the simulation time doesn't impact how you are charged.

Simulation WorldForge

Q: What is a simulation world?

A: A simulation world is the virtual model of an environment such as a house. A simulation world is required for running simulations, and it includes the geometry, visual, and physics information of a modeled environment.

Q: Why do I need WorldForge?

A: Building a simulation world can be difficult, costly and time consuming, and it also requires specialized skills in 3D modeling and simulation engines. While a few companies have the time and resources to create a single simulation world, building enough worlds to effectively scale simulation for workloads like regression testing and reinforcement learning is prohibitively difficult and costly.

RoboMaker WordForge makes it possible to automatically create hundreds of pre-defined, randomized simulation worlds that replicate real-world conditions without engineering investment, specialized design skills or managing world generation infrastructure.

Q: How do I get started with simulation WorldForge?

A: You get started with simulation WorldForge by creating a world template, either from scratch or from one of the sample templates that RoboMaker provides. With a world template, you can then generate one or more simulation worlds by creating a world generation job. If you want to use the generated worlds in a RoboMaker Simulation job or in a local environment, you can do so by creating a world export job, which exports a specified world to your Amazon S3 bucket as a ROS package and in a .zip format.

Q: What is a simulation world template?

A: A simulation world template defines the specifications of simulation worlds. For example, you can define the floorplan, rooms, and furnishings in a world template for generating a set of indoor residential worlds. A world template doesn't precisely define a single simulation world. Instead, a world template can generate multiple simulation worlds with certain randomness such as random room sizes and furniture. The specifications of a world template defines the boundaries of such randomness.

Q: What is a world generation job?

A: A world generation job generates one or more simulation worlds from a world template. When creating a world generation job, you can specify the number of world variations across two dimensions – floor plans and interiors. For example, 2 floor plans with 2 interior variations per floor plan result in 4 (2*2) unique worlds.

Q: What is a world export job?

A: A world export job exports a generated world as a ROS package and in a .zip format to your Amazon S3 bucket. You create a world export job when you need to use the generated world with a local environment or the generated worlds with RoboMaker Simulation run.

AWS IoT Core

Easily and securely connect devices to the cloud.

How it works

AWS IoT Core lets you connect billions of IoT devices and route trillions of messages to AWS services without managing infrastructure.

Message Broker

Securely transmit messages to and from all of your IoT devices and applications with low latency and high throughput.

Mirror Device State

With Device Shadow, store the latest state of a connected device so it can be read or set at any time, making the device appear to your applications as if it were online all the time.

Built-in Alexa

The Alexa Voice Service (AVS) Integration for AWS IoT Core introduces a new virtual Alexa device in the cloud. Use a new set of AWS IoT-reserved MQTT topics to transfer audio messages between devices.

LoRaWAN Devices

Set up and manage a private LoRaWAN network by connecting LoRaWAN devices and gateways to the AWS cloud—all without developing or operating a LoRaWAN Network Server (LNS).

Use cases

Monitor and manage industrial operations

Build industrial IoT applications for predictive quality, maintenance, and remote operation monitoring.

Build differentiated consumer products

Create connected applications for home automation, home security and monitoring, and home networking.

Innovate with automotive data

Develop solutions for connected, autonomous, shared, and electric vehicle (EV) applications.

Develop safety products

Design commercial applications for traffic monitoring, public safety, and health monitoring.

AWS IoT Core features

AWS IoT Core enables you to connect devices to AWS Services and other devices, secure data and interactions, process and act upon device data, enables applications to interact with devices even when they are offline and that allows you to produce low-cost Alexa built-in devices.

Key features

AWS IoT Device SDK

The AWS IoT Device SDK helps you easily and quickly connect your hardware device or your mobile application to AWS IoT Core. The AWS IoT Device SDK enables your devices to connect, authenticate, and exchange messages with AWS IoT Core using the MQTT, HTTP, or WebSockets protocols. The AWS IoT Device SDK supports C, JavaScript, and Arduino, and includes the client libraries, the developer guide, and the porting guide for manufacturers. You can also use an open source alternative or write your own SDK.

Device Advisor

Device Advisor is a fully managed cloud-based test capability for validating IoT devices during development. It provides pre-built tests that helps developers to validate their IoT devices for reliable and secure connectivity with AWS IoT Core. By using Device Advisor, developers can test if their IoT devices can reliably interoperate with AWS IoT Core and follow security best practices. Developers can identify and resolve the most common device software issues during development before they deploy their devices in production. Device Advisor also provides a signed qualification report which can be used by hardware partners to qualify their devices for inclusion in the AWS Partner Device Catalog.

Device Gateway

The Device Gateway serves as the entry point for IoT devices connecting to AWS. The Device Gateway manages all active device connections and implements semantics for multiple protocols to ensure that devices are able to securely and efficiently communicate with AWS IoT Core. Currently the Device Gateway supports the MQTT, WebSockets, and HTTP 1.1 protocols. For devices that connect using MQTT or WebSockets the Device Gateway will maintain long

lived, bidirectional connections, enabling these devices to send and receive messages at any time with low latency. The Device Gateway is fully managed and scales automatically to support over a billion devices without requiring you to manage any infrastructure. For customers migrating to AWS IoT, the Device Gateway offers capabilities to transition infrastructures with minimal impact to existing architectures and IoT devices.

Message Broker

The Message Broker is a high throughput pub/sub message broker that securely transmits messages to and from all of your IoT devices and applications with low latency. The flexible nature of the Message Broker's topic structure allows you to send messages to, or receive messages from, as many devices as you would like. It supports messaging patterns ranging from one-to-one command and control messaging, to one-to-one million (or more!) broadcast notification systems and everything in between. In addition, you can set up fine grained access controls that enable you to manage the permissions of individual connections at the topic level, ensuring that your devices and applications will only send and receive the data that you want them to. The Message Broker is a fully managed service, so no matter how you choose to use it, it will scale automatically with your message volume without requiring you to run any infrastructure.

Authentication and Authorization

AWS IoT Core provides mutual authentication and encryption at all points of connection, so that data is never exchanged between devices and AWS IoT Core without a proven identity. AWS IoT Core supports the AWS method of authentication (called 'SigV4'), X.509 certificate based authentication, and customer created token based authentication (through custom authorizers.) Connections using HTTP can use any of these methods, while connections using MQTT use certificate based authentication, and connections using WebSockets can use SigV4 or custom authorizers. With AWS IoT Core you can use AWS IoT Core generated certificates, as well as those signed by your preferred Certificate Authority (CA). You can map your choice of policies to each certificate, so that you can authorize devices or applications to have access, or change your mind and revoke access altogether without ever touching the device.

You can create, deploy and manage certificates and policies for the devices from the console or using the API. Those device certificates can be provisioned, activated and associated with the relevant IoT policies that are configured using AWS IoT Core. This allows you to instantly revoke access for an individual device if you choose to do so. AWS IoT Core also supports connections from users' mobile apps using Amazon Cognito, which takes care of all the steps necessary to create a unique identifier for your app's users and retrieve temporary, limited-privilege AWS credentials. AWS IoT Core can also provide temporary AWS credentials after a device has authenticated with an X.509 certificate, so that the device can more easily access other AWS services such as DynamoDB or S3.

Fleet provisioning for AWS IoT Core makes it easy to onboard any number of manufactured devices to the cloud at scale. It provides an end-to-end managed device onboarding experience that sets up devices with unique digital identities and performs the device-side and cloud-side configuration needed for each device to connect and operate with AWS IoT automatically upon its first connection to AWS IoT Core. Fleet provisioning is offered at no additional charge to customers. To learn more about Fleet Provisioning, read Fleet provisioning documentation here.

Registry

The Registry establishes an identity for devices and tracks metadata such as the devices' attributes and capabilities. The Registry assigns a unique identity to each device that is consistently formatted regardless of the type of device or how it connects. It also supports

metadata that describes the capabilities of a device, for example whether a sensor reports temperature, and if the data are Fahrenheit or Celsius.

The Registry lets you store metadata about your devices at no additional charge, and metadata in the Registry does not expire as long as you access or update your registry entry at least once every 7 years.

Device Shadow

With AWS IoT Core, you can create a persistent, virtual version, or Device Shadow, of each device that includes the device's latest state so that applications or other devices can read messages and interact with the device. The Device Shadow persists the last reported state and desired future state of each device even when the device is offline. You can retrieve the last reported state of a device or set a desired future state through the API or using the rules engine.

The Device Shadow makes it easier to build applications that interact with your devices by providing always available REST APIs. In addition, applications can set the desired future state of a device without accounting for the devices current state. AWS IoT Core will compare the difference between the desired and last reported state, and command the device to make up the difference.

The AWS IoT Device SDK makes it easy for your device to synchronize its state with its Device Shadow, and to respond to desired future states set via the Device Shadow.

The Device Shadow lets you store the state of your devices for up to a year for free. The Device Shadow persist forever if you update them at least once per year, otherwise they expire.

Rules Engine

The Rules Engine makes it possible to build IoT applications that gather, process, analyze and act on data generated by connected devices at global scale without having to manage any infrastructure. The Rules Engine evaluates inbound messages published into AWS IoT Core and transforms and delivers them to another device or a cloud service, based on business rules you define. A rule can apply to data from one or many devices, and it can take one or many actions in parallel.

The Rules Engine can also route messages to AWS endpoints including AWS IoT Analytics, AWS IoT Events, AWS Lambda, Amazon Kinesis, Amazon S3, Amazon DynamoDB, Amazon CloudWatch, Amazon Simple Notification Service (SNS), Amazon Simple Queue Service (SQS), Amazon Elasticsearch Service, and AWS Step Functions. External endpoints can be reached using AWS Lambda, Amazon Kinesis, Amazon SNS, and Rules Engine's native HTTP action.

You can author rules within the management console or write rules using a SQL-like syntax. Rules can be authored to behave differently depending upon the content of the message. For example, if a temperature reading exceeds a certain threshold it could trigger a rule to transmit data to AWS Lambda. Rules can also be authored to take into account other data in the cloud, such as data from other devices. For example you could say take an action if this temperature is more than 15% higher than the average of 5 other devices.

The Rules Engine provides dozens of available functions that can be used to transform your data, and it's possible to create infinitely more via AWS Lambda. For example, if you're dealing with a wide range of values you could take the average of incoming numbers. Rules can also trigger the execution of your Java, Node.js or Python code in AWS Lambda, giving you maximum flexibility and power to process device data.

Alexa Voice Service (AVS) Integration

Alexa Built-in is a category of devices created with the Alexa Voice Service (AVS) that have a microphone and speaker. You can talk to these products directly with the wake word “Alexa,” and receive voice responses and content instantly.

With the AVS Integration, you can cost-effectively scale to hundreds of millions of Alexa Built-in devices and accelerate your time to market with differentiated, voice-forward products. The AVS Integration for AWS IoT Core shifts media retrieval, audio decoding, audio mixing, and state management from a physical device to a new virtual Alexa Built-in device in the cloud. This lowers the cost of producing Alexa Built-in devices by up to 50% by reducing the hardware requirements from 50MB to 1MB of RAM and from ARM Cortex ‘A’ class microprocessors to ARM Cortex ‘M’ class microcontrollers.

Offloading compute and memory intensive workloads to the cloud and reducing the on-device requirements for integrating AVS makes it possible to bring Alexa to resource constrained, low-cost products such as light switches, thermostats, and small appliances. With new categories of Alexa Built-in devices available on the market, end users can now experience Alexa in new parts of their home, office, or hotel rooms for a truly ambient experience, where they talk directly to their surroundings rather than to an Alexa Family of Devices.

Our partners make it easy for you to get started by launching hardware development kits for AVS for AWS IoT. Development kits are enabled by real time operating system for microcontrollers like FreeRTOS and include out-of-the box connectivity to AWS IoT. Kits also remove the need to develop your voice application from scratch by including AVS qualified Audio Algorithms for Far-Field voice pickup, Echo Cancellation, and Alexa Wake Word as well as AVS for AWS IoT application code. Using the feature application code, you can quickly prototype a device and port the implementation to your chosen MCU design for testing and device production when you're ready.

AVS Integration is available as a new reserved topic in AWS IoT Core. Messages sent to and from the device to the AVS for AWS IoT reserved topics are free.

AWS IoT Core for LoRaWAN

AWS IoT Core for LoRaWAN enables customers to connect wireless devices that use low-power, long-range wide area network (LoRaWAN) technology. Using AWS IoT Core, customers can now setup a private LoRaWAN network by connecting their own LoRaWAN devices and gateways to the AWS Cloud - without developing or operating a LoRaWAN Network Server (LNS). This eliminates the undifferentiated development work and operational burden of managing an LNS and associated infrastructure, accelerating the network set-up time.

AWS IoT Core for LoRaWAN includes support for open-source LoRaWAN gateway-LNS communication protocol called Basic Station, which means no writing or testing of custom gateway software is required to connect LoRaWAN gateways to AWS IoT Core. Once LoRaWAN gateways and devices are connected, device data is automatically routed to AWS IoT Core Rules Engine accelerating IoT application development.

Developers can write rules, like simple SQL queries to transform and act on the device data, raising alerts, or routing it to other AWS services like Amazon S3 using the AWS IoT Core Rules Engine. From the console, engineers can also query metrics for connected devices and gateways to troubleshoot connectivity issues. With pay-as-you-go pricing and no monthly commitments, customers can connect and scale LoRaWAN device fleets reliably, and build applications with AWS services quickly and efficiently.

AWS IoT Partners make it easy for you to get started by providing AWS IoT Core for LoRaWAN qualified gateways that connect to AWS IoT Core out of the box, without any need to modify embedded software. Search the AWS Partner Device Catalog to find gateways that are qualified for use with AWS IoT Core for LoRaWAN.

Amazon Sidewalk Integration

Amazon Sidewalk is a shared network that helps connected devices work better through improved connectivity options. Operated by Amazon at no charge to customers, Sidewalk can help simplify new device setup, extend the low-bandwidth working range of devices, and help devices stay online even if they are outside the range of their home Wi-Fi.

Amazon Sidewalk Integration enables you to easily onboard your Sidewalk device fleets to AWS IoT Core. Amazon Sidewalk has been designed to support a wide range of customer devices like pet or valuables location trackers, to smart home security and lighting controllers, to remote diagnostics for home appliances and tools.

LoRaWAN is a mark used under license from the LoRa Alliance.

What is AWS IoT?

AWS IoT provides the cloud services that connect your IoT devices to other devices and AWS cloud services. AWS IoT provides device software that can help you integrate your IoT devices into AWS IoT-based solutions. If your devices can connect to AWS IoT, AWS IoT can connect them to the cloud services that AWS provides.

AWS IoT lets you select the most appropriate and up-to-date technologies for your solution. To help you manage and support your IoT devices in the field, AWS IoT Core supports these protocols:

- MQTT (Message Queuing and Telemetry Transport)
- MQTT over WSS (Websockets Secure)
- HTTPS (Hypertext Transfer Protocol - Secure)

- LoRaWAN (Long Range Wide Area Network)

The AWS IoT Core message broker supports devices and clients that use MQTT and MQTT over WSS protocols to publish and subscribe to messages. It also supports devices and clients that use the HTTPS protocol to publish messages.

AWS IoT Core for LoRaWAN helps you connect and manage wireless LoRaWAN (low-power long-range Wide Area Network) devices. AWS IoT Core for LoRaWAN replaces the need for you to develop and operate a LoRaWAN Network Server (LNS).

If you don't require AWS IoT features such as device communications, rules, or jobs, see AWS Messaging for information about other AWS IoT messaging services that might better fit your requirements.

AWS IoT Core FAQs

General

Q. What is AWS IoT Core?

AWS IoT Core is a managed cloud platform that lets connected devices easily and securely interact with cloud applications and other devices. AWS IoT Core can support billions of devices and trillions of messages, and can process and route those messages to AWS endpoints and to other devices reliably and securely. With AWS IoT Core, your applications can keep track of and communicate with all your devices, all the time, even when they aren't connected.

Q. What does AWS IoT Core offer?

Connectivity between devices and the AWS cloud. First, with AWS IoT Core you can communicate with connected devices securely, with low latency and with low overhead. The communication can scale to as many devices as you want. AWS IoT Core supports standard communication protocols (HTTP, MQTT, and WebSockets and LoRaWAN are supported currently). Communication is secured using TLS.

Processing data sent from connected devices. Secondly, with AWS IoT Core you can continuously ingest, filter, transform, and route the data streamed from connected devices. You can take actions based on the data and route it for further processing and analytics.

Application interaction with connected devices. Finally, AWS IoT Core accelerates IoT application development. It serves as an easy to use interface for applications running in the cloud and on mobile devices to access data sent from connected devices, and send data and commands back to the devices.

Q. How does AWS IoT Core work?

Connected devices, such as sensors, actuators, embedded devices, smart appliances, and wearable devices, connect to AWS IoT Core over HTTPS, WebSockets, or secure MQTT or LoRaWAN. Included in AWS IoT Core is a Device Gateway that allows secure, low-latency, low-

overhead, bi-directional communication between connected devices and your cloud and mobile applications.

AWS IoT Core also contains a Rules Engine which enables continuous processing of data sent by connected devices. You can configure rules to filter and transform the data. You also configure rules to route the data to other AWS services such as DynamoDB, Kinesis, Lambda, SNS, SQS, CloudWatch, Elasticsearch Service with built-in Kibana integration, as well as to non-AWS services, via Lambda for further processing, storage, or analytics.

There is also a Registry where you can register and keep track of devices connected to AWS IoT Core, or devices that may connect in the future. The Device Shadow in AWS IoT Core enables cloud and mobile applications to query data sent from devices and send commands to devices, using a simple REST API, while letting AWS IoT Core handle the underlying communication with the devices. The Device Shadow accelerates application development by providing a uniform interface to devices, even when they use one of the several IoT communication and security protocols with which the applications may not be compatible. The Device Shadow also accelerates application development by providing an always available interface to devices even when the connected devices are constrained by intermittent connectivity, limited bandwidth, limited computing ability or limited power.

Communication with AWS IoT Core is secure. The service requires all of its clients (connected devices, server applications, mobile applications, or human users) to use strong authentication (X.509 certificates, AWS IAM credentials, or 3rd party authentication via AWS Cognito). All communication is encrypted. AWS IoT Core also offers fine-grained authorization to isolate and secure communication among authenticated clients.

Q. What is 2lemetry and how does it relate to AWS IoT?

2lemetry was acquired by AWS in 2015, and their capabilities provided foundational elements such as the MQTT Message Broker and the Rules Engine for AWS IoT Core.

Q. In which regions is AWS IoT Core available?

See the AWS Region Table for the current list of regions for AWS IoT Core.

You can use AWS IoT Core regardless of your geographic location, as long as you have access to one of the above AWS regions.

Q. How do I get started with using AWS IoT Core?

Use the AWS IoT Console or refer to the Quickstart section of our developer guide to test drive the AWS IoT Core in minutes.

Also, take a look at the AWS-powered Starter Kits provided by our partners.

Q. Which languages does the AWS IoT Console support?

The AWS IoT Console supports English, French, Japanese, Korean, Simplified Chinese, German, Portuguese, Spanish, Italian and Traditional Chinese.

Q. How can I switch the console's language?

Click on the language at the bottom left corner of the console to pick the language. The language selection will persist throughout the consoles of different AWS services.

Q. What are the ways for accessing AWS IoT Core?

You can use the AWS Management Console, the AWS SDKs, the AWS CLI, and the AWS IoT Core APIs. Connected devices can use the AWS IoT Device SDKs to simplify the communication with AWS IoT Core.

The AWS IoT Core APIs and commands are largely divided into control plane operations and data plane operations. The control plane operations enable you to do tasks such as configuring security, registering devices, configuring rules for routing data, and setting up logging. The data plane operations enable you to ingest data from connected devices into AWS IoT Core with low latency and high throughput rate at a large scale.

Q. What communication and authentication protocols does AWS IoT Core support?

For control plane operations, AWS IoT Core supports HTTPS. For data plane operations, AWS IoT Core supports HTTPS, WebSockets, and secure MQTT – a protocol often used in IoT scenarios.

HTTPS and WebSockets requests sent to AWS IoT Core are authenticated using AWS IAM or AWS Cognito, both of which support the AWS SigV4 authentication. If you are using the AWS SDKs or the AWS CLI, the SigV4 authentication is taken care of for you under the hood. HTTPS requests can also be authenticated using X.509 certificates. MQTT messages to AWS IoT Core are authenticated using X.509 certificates.

With AWS IoT Core you can use AWS IoT Core generated certificates, as well as those signed by your preferred Certificate Authority (CA).

Q. Can devices that are NOT directly connected to the Internet access AWS IoT Core?

Yes, via a physical hub. Devices connected to a private IP network and devices using non-IP radio protocols such as ZigBee or Bluetooth LE can access AWS IoT Core as long as they have a physical hub as an intermediary between them and AWS IoT Core for communication and security.

Q. How should applications access AWS IoT Core?

Applications connecting to AWS IoT Core largely fall in two categories: 1. companion apps and 2. server applications. Companion apps are mobile or client-side browser applications that interact with connected devices via the cloud. A mobile app that lets a consumer remotely unlock a smart lock in the consumer’s house is an example of a companion app. Server applications are designed to monitor and control a large number of connected devices at once. An example of a server application would be a fleet management website that plots thousands of trucks on a map in real-time.

AWS IoT Core enables both companion apps and server applications to access connected devices via uniform, RESTful APIs. Applications also have the option to use pub/sub to communicate directly with the connected devices.

Typically the companion apps would authenticate using end-user identities which are managed either by your own identity store or a third party identity provider such as Facebook and Login with Amazon. For companion apps, use Amazon Cognito, which integrates with several identity providers. Cognito identities can be authorized to access AWS IoT Core, and their access can be restricted only to the resources relevant to them. For example, as a connected washing machine manufacturer, you can authorize your consumers to access your AWS IoT Core information pertaining only to their individual washing machines.

Server applications (such as a mapping application running on Amazon EC2) can use IAM roles to access AWS IoT Core.
Q. Can I get a history of AWS IoT Core API calls made on my account for security analysis and operational troubleshooting purposes?

Yes, to receive a history of AWS IoT Core API calls made on your account, you simply turn on CloudTrail in the AWS Management Console.

Q. How do I send feedback?

To send feedback, click on the “Feedback” link in the footer bar of the console.

Features

Q: What is the Device Gateway?

The Device Gateway forms the backbone of communication between connected devices and the cloud capabilities such as the Rules Engine, Device Shadow, and other AWS and 3rd-party services.

The Device Gateway supports the pub/sub messaging pattern, which enables scalable, low-latency, and low-overhead communication. It is particularly useful for IoT scenarios where billions of devices are expected to communicate frequently and with minimal delay. Pub/sub involves clients publishing messages on logical communication channels called ‘topics’ and clients subscribing to topics to receive messages. The device gateway enables the communication between publishers and subscribers. Traditionally, organizations have had to provision, operate, scale, and maintain their own servers as device gateways to take advantage of pub/sub. AWS IoT Core has eliminated this barrier by providing the Device Gateway.

The Device Gateway scales automatically with your usage, without any operational overhead for you. AWS IoT Core supports secure communication with the device gateway, AWS-account level isolation, as well as fine-grained authorization within an AWS account. The device gateway currently supports publish and subscribe over secure MQTT and WebSockets, as well as publish over HTTPS.

Q. What is MQTT?

MQTT is a lightweight pub/sub protocol, designed to minimize network bandwidth and device resource requirements. MQTT also supports secure communication using TLS. MQTT is often used in IoT use cases. MQTT v3.1.1 is an OASIS standard, and the Device Gateway supports most of the MQTT specification.

Q. What is the Rules Engine?

The Rules Engine enables continuous processing of inbound data from devices connected to AWS IoT Core. You can configure rules in the Rules Engine in an intuitive, SQL-like syntax to automatically filter and transform inbound data. You can further configure rules to route data from AWS IoT Core to several other AWS services as well as your own or 3rd party services.

Here are just a few example use cases of rules:

- Filtering and transforming incoming messages and storing them as time series data in DynamoDB.
- Sending a push notification via SNS when the data from a sensor crosses a certain threshold.
- Saving a firmware file to S3
- Processing messages simultaneously from a multitude of devices using Kinesis
- Invoke Lambda to do custom processing on incoming data
- Sending a command to a group of devices with an automated republish

Q. How are the rules defined and triggered?

An AWS IoT Core rule consists of two main parts:

A SQL statement that specifies the pub/sub topics to apply the rule on, data transformation to perform, if any, and the condition under which the rule should be executed. The rule is applied on every message published on the specified topics.

An actions list that defines the actions to take when the rule is executed, that is, when an incoming message matches the condition specified in the rule.

Rule definitions use a JSON-based schema. You can directly edit the JSON or use the rules editor in the AWS Management Console.

As an example, here is a rule for saving temperature data from a sensor to DynamoDB whenever the temperature is above 50:

```
{
   "sql": "SELECT * from 'iot/tempSensors/#' WHERE temp > 50",
   "description": "Rule to save sensor data when temperature is about 50",
   "actions": [
    {
         "dynamoDB": {
         "tableName": "HighTempTable",
         "roleArn": "arn:aws:iam::your-aws-account-id:role/dynamoPut",
         "hashKeyField": "key",
         "hashKeyValue": "${topic(3)}",
         "rangeKeyField": "timestamp",
         "rangeKeyValue": "${timestamp()}"
         }
      }
   ]
}
```

Sensors in this example are publishing on their topics under “iot/tempSensors/”. The first line of the rule defines the SQL SELECT statement used to query on the “iot/tempSensors/#” topic. It contains a WHERE clause that extracts the value of a ‘temp’ field in the message’s payload and checks if it passes the condition ‘greater than 50’. If the condition is met, the data is stored in the specified DynamoDB table. The example uses built-in functions for tasks such as traversing the message payload and getting current time.

Q. Where can I learn more about rules?

You can learn more about rule here Core Rules documentation

Q. What is the Registry and what should I use it for?

IoT scenarios can range from a small number of mission-critical devices to large fleets of devices. The Registry allows you to organize and track those devices. You can maintain a logical handle in the Registry for every device you are connecting to AWS IoT Core. Each device in the Registry can be uniquely identified and can have metadata such as model numbers, support contact, and certificates associated with it. You can search for connected devices in the Registry based on the metadata.

Q. What is a Thing Type?

Thing Types allow you to effectively manage your catalogue of devices by defining common characteristics for devices that belong to the same device category. In addition, a Thing associated with a Thing Type can now have up to 50 attributes including 3 searchable attributes.

Q. What is Simplified Permission Management?

This feature allows you to easily manage permission policies for a large number of devices by using variables that reference Registry or X.509 certificate properties. The integration of Registry and Certificate properties with device policies offers the benefits listed below:

- You can now reference Registry properties in device permission policies. Referencing device properties defined in the Registry allows your policies to reflect any changes made in the Registry. For example, by referencing the Thing Attribute named “building-address” as a variable in the policy, devices will automatically inherit a new set of permissions when they move buildings.
- You can share a single generic policy for multiple devices. A generic policy can be shared among the same category of devices instead of creating a unique policy per device. For example, a policy that references the “serial-number” as a variable, can be attached to all the devices of the same model. When devices of the same serial number connect, policy variables will be automatically substituted by their serial-number.

Q. What is the Device Shadow?

The Device Shadow enables cloud and mobile applications to easily interact with the connected devices registered in AWS IoT Core. The Device Shadow in AWS IoT Core contains properties of a connected device. You can define any set of properties applicable to your use case. For example, for a smart light bulb, you might define ‘on-or-off’, ‘color’, and ‘brightness’ as the

properties. The connected device is expected to report the actual values of those properties, which are stored in the Device Shadow. Applications get and update the properties simply by using a RESTful API provided by AWS IoT Core. AWS IoT Core and the Device SDKs take care of synchronizing property values between the connected device and its Device Shadow in AWS IoT Core.

Q. Do I have to use the Registry and the Device Shadow?

You can have applications communicate directly to the connected devices using the Device Gateway and/or the Rules Engine in AWS IoT Core. However, we recommend using the Registry and the Device Shadow since they offer richer and more structured development and management experience that lets you focus on the unique value you want to create for your customers rather than having to focus on the underlying communication and synchronization between the connected devices and the cloud.
Q. What is the lifecycle of a device and its Device Shadow in AWS IoT Core?

- You register a device (such as a light bulb) in the Registry.
- You program connected device to publish a set of its property values or 'state ("I am ON and my color is RED") to the AWS IoT Core service.
- The last reported state is stored in the Device Shadow in AWS IoT Core.
- An application (such as a mobile app controlling the light bulb) uses a RESTful API to query AWS IoT Core for the last reported state of the light bulb, without the complexity of communicating directly with the light bulb.
- When a user wants to change the state (such as turning the light bulb from ON to OFF), the application uses a RESTful API to request an update, i.e. sets a 'desired' state for the device in AWS IoT Core. AWS IoT Core takes care of synchronizing the desired state to the device.
- The application gets notified when the connected device updates its state to the desired state.

Q. Where can I learn more about the Registry and the Device Shadow?

For more information on the Registry, see the Registry documentation. For more information on the Device Shadow, see the Device Shadow documentation.

Q. Can I configure fine-grained authorization in AWS IoT Core?

Yes. Similar to other AWS services, in AWS IoT Core you have fine-grained control over the set of API actions each identity is authorized to invoke. In addition, you have fine-grained control over the pub/sub topics that an identity can publish or subscribe to, as well as over the devices and the Device Shadow in the Registry that an identity can access.

Q. Where can I learn more about Security and Access Control in AWS IoT Core?

For more information, see AWS IoT Core Security and Identity.

Q. What is Just-in-time registration of certificates?

Just-in-time registration (JITR) of device certificates expands on the "Use Your Own Certificate" feature launched in April 2016 by simplifying the process of enrolling devices with AWS IoT Core. Prior to support for JITR, the device enrollment process required two steps: first, registering the Certificate Authority (CA) certificate to AWS IoT Core, then individually registering the device certificates that were signed by the CA. Now, with JITR you can complete the second step by auto-registering device certificates when devices connect to AWS IoT Core for the first time. This saves time spent on registering device certificates and allows devices to remain off-line during the manufacturing process. To further automate IoT device provisioning, you can create an AWS IoT Core rule with a Lambda action that activates the certificates and attaches policies.

Q. What is the AWS IoT Device SDK?

The AWS IoT Device SDKs simplify and accelerate the development of code running on connected devices (micro-controllers, sensors, actuators, smart appliances, wearable devices, etc.). First, devices can optimize the memory, power, and network bandwidth consumption by using the Device SDKs. At the same time, Device SDKs enable highly secure, low-latency, and low-overhead communication with built-in TLS, WebSockets, and MQTT support. The Device SDKs also accelerate IoT application development by supporting higher level abstractions such as synchronizing the state of a device with its Device Shadow in AWS IoT Core.

AWS IoT Device SDKs are freely available as open-source projects. For more details visit our Device SDK page.

Q. Which programming languages and hardware platforms does the AWS IoT Device SDK support?

AWS currently offers the AWS IoT Device SDKs for C and Node.js languages, as well as for the Arduino Yún platform.

In addition, several hardware manufacturers have partnered with AWS to make the AWS IoT Device SDKs available on their respective platforms.

Lastly, AWS IoT Device SDKs are open-source. You can port them to the languages and hardware platforms of your choice if they are not supported already.

Q: Should I use AWS IoT Device SDK or the AWS SDKs?

The AWS IoT Device SDK complements the AWS SDKs. IoT projects often involve code running on micro-controllers and other resource-constrained devices. However, IoT projects often include application running in the cloud and on mobile devices that interact with the micro-controllers/resource-constrained devices. AWS IoT Device SDKs are designed to be used on the micro-controllers/resource-constrained devices, while the AWS SDKs are designed for cloud and mobile applications.

Alexa Voice Service Integration

Q. Why should I use the AVS Integration for AWS IoT?

Until now, producing an Alexa Built-in device required on-device memory and compute to be at least 50MB RAM and ARM Cortex 'A' class microprocessors, increasing the engineering bill of materials (eBOM) and MSRP. Additionally, retrieving, buffering, decoding, and mixing audio on devices can be complex and time consuming. The high production cost and complexity makes

it difficult for device makers to quickly go to market with differentiated, voice-enabled experiences on resource-constrained IoT devices.

The AVS Integration lowers the Alexa Built-in cost by up to 50% by offloading compute & memory intensive workloads to the cloud. This reduces the hardware requirements of integrating AVS on a device from 50MB to 1MB of RAM and from ARM Cortex 'A' class microprocessors to ARM Cortex 'M' class microcontrollers and brings Alexa to ubiquitous products such as light switches, thermostats, and small appliances. In addition to the simplified device side AVS integration, device makers using the AVS Integration also have access to simple and cost-effective ongoing device maintenance and management, enhanced device security, and analytics services via the scalable AWS IoT Core infrastructure.

With new categories of low-price Alexa Built-in devices available on the market, end users can now experience Alexa in new parts of their home, office, or hotel rooms for a truly ambient experience, where they talk directly to their surroundings rather than to an Alexa Family of Devices.

Q. How do I use the Alexa Voice Service (AVS) Integration?

Learn how to create low-cost Alexa Built-in devices with the AVS Integration for AWS IoT Core Getting Started Guide.

Q. How is AVS for IoT different from traditional AVS?

AVS Integration supports Device arbitration, Dialog, Multi-turn dialog, Timers, Alarms, Reminders, Flash Briefing, Routines, Alexa Announce, eBooks, and Skills. It does not support high-quality music playback, Whole Home Audio, Alexa Calling, Spotify, Bluetooth, and rich multi-modal displays.

Q. What types of devices can I build with AVS?

The AVS Integration for AWS IoT is a great solution for device makers producing low-cost, resource-constrained devices (including light switches, light bulbs, home hubs, home appliances and more) that want to allow their customers to talk to these products directly with the wake word "Alexa," and receive voice responses and content instantly. These devices will have Built-in microphones and speakers that are capable of playing back dialog, alerts, and the news but are not adequate to support high-quality music playback. Device makers who want full music playback with richer Alexa Music Playback capabilities such as high fidelity (>128kbps) music streaming, Spotify, synchronized music streaming over multiple speakers, should continue to build these devices using the existing Alexa Built-in solutions.

Q. Do I get the Alexa Built-in badge using the AVS Integration for AWS IoT Core?

Similar to other Alexa Built in products, products built with the AVS Integration will need to pass the Alexa Voice Service product certification process comprising of Amazon-managed testing of security, acoustic performance, user experience, and functional testing to earn the Amazon Certified Alexa Built-in badge.

Q. What AWS regions will AVS Integration for IoT Core be available in at launch?

The AVS Integration for AWS IoT Core is available in all AWS regions where AWS IoT Core is available other than China (Beijing and Ningxia), Asia Pacific (Hong Kong) and Middle East (Bahrain). See the AWS Region Table for the current list of regions for AWS IoT Core.

AWS IoT Core for LoRaWAN

Q: Where can I find the Basic Station source code, if required for AWS IoT Core for LoRaWAN?

Basic Station software is maintained and distributed by Semtech via their Github repository.

Q: Which private LoRaWAN network components are owned and managed by AWS IoT Core vs the customer?

Devices: You own and connect your choice of LoRaWAN devices to AWS IoT Core. You can buy any LoRa device or sensor compliant with LoRa 1.0.3 or 1.1 specification (without any need to develop or update software).

Gateways: You own and connect your choice of LoRaWAN gateways to AWS IoT Core. AWS IoT Core enables two choices: you can easily select and purchase AWS qualified gateways from Amazon Partner Device Catalog that connect with AWS IoT Core out of the box. Otherwise, you can connect a brownfield (legacy) or off-the-shelf (e.g., Raspberry Pi) gateway by updating its firmware to support open source Basic Station protocol. You can find instructions to update the firmware in our getting started guide.

LoRaWAN Network Server (LNS): AWS IoT Core owns and provides a fully-managed LNS. With a few simple steps on the AWS IoT console, you can register your chosen hardware and visualize messages from your connected LoRaWAN devices.

Cloud Application: You own and develop your cloud application. Once your devices are connected to AWS IoT Core, you can start developing an application or solution by routing device messages to AWS services.

Network management: AWS IoT Core provides the network management features that customers use and configure to operate their LoRaWAN networks. Using AWS IoT Management Console, you can query the connected/disconnected devices and gateways, device traffic attributes and gateway stats of the last message received. In addition, using APIs, you can remotely update the LoRaWAN gateway firmware.

Amazon Sidewalk Integration

Q: What is Amazon Sidewalk?

Amazon Sidewalk is a shared network that helps devices work better. Operated by Amazon at no charge to customers, Sidewalk can help simplify new device setup, extend the low-bandwidth working range of devices, and help devices stay online even if they are outside the range of their home Wi-Fi. In the future, Sidewalk will support a range of experiences from using Sidewalk-enabled devices to help find pets or valuables, to smart security and lighting, to diagnostics for home appliances and tools.

Q: What is Sidewalk Developer Service (SDS)?

SDS is a collection of approved silicon chipsets along with reference development boards, device software development kits (SDKs), device provisioning tools, technical documentation, and a cloud integration (see Amazon Sidewalk Integration). Together these tools help device manufacturers develop, and easily launch devices that work on Amazon Sidewalk. See SDS Console for more info.

Q: What is Amazon Sidewalk Integration for AWS IoT Core?

Amazon Sidewalk Integration enables you to easily onboard your Sidewalk device fleet to AWS

IoT Core. Amazon Sidewalk Integration ensures your devices are authenticated properly to work on Amazon Sidewalk, and that your device data will automatically route to AWS IoT Core so that you can further develop your business application(s) using AWS services.

LoRaWAN is a mark used under license from the LoRa Alliance.

AWS IoT Core Device Advisor

Q. Who should use Device Advisor?

Developers at device manufacturers should use Device Advisor to test their devices against pre-built test scenarios to verify reliable and secure connectivity to AWS IoT Core. Device Advisor provides a test endpoint in the AWS cloud, which device manufacturers can immediately use to test their devices, saving time and cost of development and testing. The test setup also provides detailed logs for each test, enabling faster troubleshooting of device software issues. Device Advisor also provides test coverage for complex test scenarios, enabling customers to discover and fix issues during their device software development. This results in more reliable performance and lower maintenance costs for device fleets after deployment.

Also, with IoT Device Advisor, hardware partners can self-test their devices, download signed qualification reports and submit the reports to APN to get their devices listed in the AWS Partner Device Catalog.

Q. How do I use the Device Advisor?

Any device that has been built to connect to AWS IoT Core can take advantage of Device Advisor. Developers at device manufacturers can access Device Advisor from the AWS IoT Core console or by using the AWS SDK. Once developers are ready to test their devices, they can register the devices with AWS IoT Core and configure the device software with the Device Advisor end point. They can then choose and execute the pre-built tests with a few simple clicks in the IoT Core Console and instantly get the test results along with detailed logs.

Q. What tests are provided by Device Advisor?

See the test cases section in the Device Advisor for details on the pre-built tests supported.

Q. Is there a cost to use Device Advisor?

Device Advisor is free to use. However, developers will be responsible for any costs associated with AWS usage as part of the testing (e.g. AWS IoT Core, Amazon CloudWatch usage). The AWS resource usage as part of testing will be visible to developers in their AWS account and charges for these will apply to the developers' AWS bill.

AWS IoT Device Client

Q. What is the AWS IoT Device Client?

The AWS IoT Device Client is a free, open-source, and modular device-side reference implementation. It allows you to easily connect your devices to AWS IoT Core, and access AWS IoT Device Management and AWS IoT Device Defender features by default. It is written in C++ and can be compiled and installed on Embedded Linux-based IoT devices.

Q. What AWS IoT features does the AWS IoT Device Client enable on my device?

The current release of the AWS IoT Device Client enables you to connect your IoT device to AWS IoT Core, access the Jobs and Secure Tunneling features of AWS IoT Device Management, and access the Rules Detect feature of AWS IoT Device Defender.

Q. How do I use the AWS IoT Device Client?

To get started on your device, download the AWS IoT Device Client source code from GitHub and learn more using the readme.

Q. What hardware and software environments are currently supported in the AWS IoT Device Client?

The IoT Device Client currently works by default with microprocessor based IoT devices running an x86_64 or ARM architectures, and common Linux software environments (Debian, Ubuntu, RHEL). This covers most commonly available devices including for example, the latest Raspberry Pi. If you are looking for a different device, visit the AWS Partner Device Catalog.

Q. What are the minimum hardware specifications to run the AWS IoT Device Client?

We recommend using the IoT Device Client on a microprocessor based device with 100MHz+ compute and 64MB+ memory. If you wish to write your own purpose-built device code for more constrained devices, or microcontroller based devices, we offer updated feature libraries for Fleet Provisioning, Jobs, and AWS IoT Device Defender in the latest version of the AWS IoT Device SDK for Embedded C on GitHub.

AWS Ground Station

AWS Ground Station is a fully managed service that lets you control satellite communications, process data, and scale your operations without having to worry about building or managing your own ground station infrastructure. Satellites are used for a wide variety of use cases, including weather forecasting, surface imaging, communications, and video broadcasts. Ground stations form the core of global satellite networks. With AWS Ground Station, you have direct access to AWS services and the AWS Global Infrastructure including a low-latency global fiber network. For example, you can use Amazon S3 to store the downloaded data, Amazon Kinesis Data Streams for managing data ingestion from satellites, and Amazon SageMaker for building custom machine learning applications that apply to your data sets. You can save up to 80% on the cost of your ground station operations by paying only for the actual antenna time used, and relying on the global footprint of ground stations to download data when and where you need it. There are no long-term commitments, and you gain the ability to rapidly scale your satellite communications on-demand when your business needs it.

How it works

AWS Ground Station Features

Schedule satellites and download data using AWS services

You can use the AWS Ground Station console to identify the satellites you need to communicate with and schedule “Contacts” with the satellite, where each Contact consists of a selected satellite, start and end time, and the ground location. After scheduling your Contacts, you can launch Amazon EC2 instances to run each portion of the Contact. You can launch a Command EC2 instance to receive operational telemetry from your satellite and transmit commands up to the satellite to schedule future activities. You can also launch a Downlink EC2 instance to receive receive bulk mission data from your satellite in near real-time or receive the data in an Amazon S3 Bucket asynchronously. The EC2 instance will communicate with AWS Ground Station’s antenna gateway over an elastic network interface (ENI) connection in Amazon VPC that exists for the duration of the contact.

Fully managed global ground station network integrated with AWS Global Infrastructure

AWS Ground Station antennas are located within fully managed AWS ground station locations, and are interconnected via Amazon’s low-latency, highly reliable, scalable and secure global network backbone. Data downlinked and stored in one AWS Region can be sent to other AWS Regions over the global network, so it can be further processed.

Graphical AWS Ground Station console

AWS Ground Station provides an easy to use graphical console that allows you to reserve contacts and antenna time for your satellite communications. You can review, cancel, and reschedule contact reservations up to 15 minutes prior to scheduled antenna times.

Direct access to AWS services

AWS Ground Station provides our satellite antennas direct access to AWS services for faster, simpler and more cost-effective storage and processing of downloaded data. This allows you to reduce data processing and analysis times for use cases like weather prediction or natural disaster imagery from hours to minutes or seconds. This also enables you to quickly create business rules and workflows to organize, structure, route the satellite data before it can be analyzed and incorporated into key applications such as imaging analysis and weather forecasting. Key AWS services include Amazon EC2, Amazon S3, Amazon VPC, Amazon Rekognition, Amazon SageMaker, and Amazon Kinesis Data Streams.

Support most common satellites and communication frequencies

You can connect with any satellite in low Earth orbit (LEO) and medium Earth orbit (MEO) operating in X-band and S-band frequencies, including: S-band uplink and downlink, X-band narrowband and wideband downlink.

What Is AWS Ground Station?

AWS Ground Station is a fully managed service that enables you to control satellite communications, process satellite data, and scale your satellite operations. This means that you no longer have to build or manage your own ground station infrastructure.

AWS Ground Station enables you to focus on innovating and rapidly experimenting with new applications that ingest satellite data and dynamically scale your server and storage use, rather than spend resources on operating and maintaining your own ground stations.

AWS Ground Station FAQs

General

Q. What is AWS Ground Station?

AWS Ground Station is a fully managed service that will enable customers to easily command, control, and downlink data from satellites. You can schedule access to AWS Ground Station antennas on a per-minute basis and pay only for your time used. When using AWS Ground Station, you can ingest data from the satellite, monitor the satellite health and status, and transmit commands to change the satellite's operations. Incoming data can be streamed to an Amazon EC2 instance or an Amazon S3 Bucket, where it can then be stored or processed using other AWS services.

Q. What are the key benefits of AWS Ground Station?

AWS Ground Station is the industry's first satellite Ground Station as a Service located within the AWS Global Infrastructure footprint to offer on-demand, elastic access to ground station satellite antennas without long-term contracts. Traditionally, you needed to make significant capital investments to build ground stations for these communications, including the costs for servers and storage to process incoming data. This limits your ability to respond quickly to new business opportunities or significant events (such as major weather events) and requires you to operate and maintain a global network of ground antennas. With AWS Ground Station, AWS manages the ground station infrastructure, so you can focus on innovating and rapidly experimenting with new applications that process, analyze, and distribute satellite data. You can easily integrate this data with other AWS services, either in the same region or in another AWS Region using Amazon's international, high-capacity backbone network.

Q. How do I get started using AWS Ground Station?

To get started, navigate to the AWS Ground Station console in the AWS Management Console. Here, you identify the satellites you need to communicate with and schedule "Contacts" with the satellite. Each Contact consists of a selected satellite, start and end time, and the ground location. You can review confirmed Contact times in your console and cancel or reschedule up to 15 minutes prior to the scheduled contact time. After scheduling a Contact, use the AWS Ground Station EC2 AMI to launch EC2 instances that will uplink and downlink data during the contact or receive downlinked data in an Amazon S3 bucket. Depending on your mission, you may need Command instances to receive operational telemetry from the satellite and transmit changes to the satellite's planned future activities, and Downlink instances to receive bulk mission data from the satellite. These instances will communicate with the AWS Ground Station antenna gateway using an ENI connection that exists between the instances and the satellite antenna for the duration of the contact.

Q. Can I schedule time on an antenna in one location (e.g. Oregon) while I am working from another location (e.g. Sydney)?

Yes, AWS Ground Station is a global network of antenna systems that are available to users around the world. In fact, it will be common for many users to schedule time on antennas at every location in the global network.

Q: Where are Ground Station antennas located?

AWS Ground Station continues to expand the service to AWS Regions and locations around the world.

Features

Q. What is a Contact?

A Contact is a reservation to communicate with a specific satellite from an identified ground location between certain times.

Q. What happens when my scheduled Contact time arrives?

Prior to the scheduled contact time, use the Amazon EC2 AMI to launch the required instances to communicate with the satellite or create an Amazon S3 Bucket to receive downlinked data. Just prior to the scheduled contact, your EC2 instances will be able to establish connections to the AWS Ground Station antenna gateway over an ENI connection. Once the contact begins, your EC2 instances will begin sending and receiving data from the satellite. If you choose to receive downlinked data to an Amazon S3 Bucket, the data will be delivered to your Amazon S3 bucket without the need for a Downlink instance to receive the data.

Q. What happens if my requested Contact cannot be reserved?

In the event your preferred Contact cannot be granted (for example, due to existing reservations for antenna time at the chosen location), the AWS Ground Station Management Console will provide available alternatives for your review.

Q. How does AWS Ground Station make sure nobody else commands my satellite?

AWS Ground Station uses multiple measures to protect satellites from unauthorized contact. Prior to allowing contact with a satellite for the first time, AWS onboards and identifies the satellite owner and associates the satellite and satellite owner with a designated customer account. The customer is then able to use that account with AWS Ground Station to schedule satellite contacts or authorize other AWS accounts to schedule contacts. Access to the AWS Ground Station antenna gateway is limited to the authorized EC2 instances for the period of the contact. Customers are in complete control of the encryption keys used to authorize and encrypt data to the satellite. Amazon S3 data is written to their Amazon S3 bucket and stored using Server Side Encryption (SSE).

Q. What types of satellites can AWS Ground Station communicate with?

Low Earth Orbit (LEO), Non-Geostationary Earth Orbit (NGSO), and Medium Earth Orbit (MEO).

Q. What radio frequencies/bands can AWS Ground Station use to communicate with satellites?

AWS Ground Station antennas are capable of using common frequency bands to communicate with satellites, including: S- and X-.

Q. How fast is AWS Ground Station?

AWS Ground Station supports narrowband uplink speeds up to 54 MHz and downlink speeds up to 500MHz.

Q. How far in advance can I schedule Contacts?

Reserved Minute Contacts may be reserved up to 21 days in advance and rescheduled up to 1 day prior to a scheduled contact. On demand contacts can be scheduled as much as 7 days and as little as 15 minutes in advance and cannot be rescheduled.

Q. How does AWS Ground Station relate to/work with other AWS products?

It's easy to use AWS Ground Station with other AWS services to process and store satellite-originated data, either in the region local to each AWS Ground Station antenna or in another AWS region (using Amazon's international backbone network.) Data can be stored locally on EC2 instances using Amazon Elastic Block Store (EBS), in a shared filesystem using Amazon Elastic File System (EFS), or in Amazon S3. In S3, you can configure lifecycle policies to automatically migrate older, less-frequently-accessed data to less expensive storage classes, including S3 Infrequent Access and Amazon Glacier.

You can use Amazon Kinesis Data Streams to fully manage data ingestion and provide consistent APIs for integrating data analysis into your applications. It provides seamless integration with Amazon Rekognition, enabling automatic recognition of objects (such as cars or airplanes). You can also use Amazon SageMaker to build custom machine learning applications that apply to your data.

Q: What transmit and receive operational frequencies does Ground Station support?

Existing Ground Station antenna systems are capable of supporting the following frequencies:

- S-Band transmit from 2025 to 2120 MHz
- S-Band receive from 2200 to 2300 MHz
- X-Band receive from 7750 to 8400 MHz

Amazon Braket

Accelerate quantum computing research.

How it works

Amazon Braket is a fully managed quantum computing service designed to help speed up scientific research and software development for quantum computing.

Use cases

Research quantum computing algorithms

Accelerate scientific discovery with tools for algorithm development and support from the AWS Cloud Credit for Research Program.

Test different quantum hardware

Push the boundaries of quantum hardware research with easy access to superconducting, trapped ion, neutral atom, and photonic devices.

Build quantum software faster

Bring software for quantum computing to market rapidly with Amazon Braket's software development kit (SDK), simple pricing, and workflow management.

Develop open-source software

Contribute new quantum applications and influence software features, tools, or plug-ins that would make development simpler.

Amazon Braket Features

Amazon Braket is a fully managed quantum computing service that helps you get started with quantum computing and accelerate your research. Amazon Braket provides everything you need to build, test, and run quantum algorithms on AWS. It includes access to different types of quantum computers, a unified development environment, a choice of classical circuit simulators, and fully managed execution of hybrid quantum-classical algorithms.

Developer Tools

Hardware-agnostic developer framework

To simplify the process of designing and executing quantum algorithms, you can use the Amazon Braket SDK. It has been designed to be technology agnostic, removing the need to code against different quantum programming tools for each type of quantum hardware. The SDK supports a unified developer framework that enables you to build quantum algorithms and run them on any compatible quantum hardware or circuit simulator provided through the Amazon Braket service. As new quantum technologies emerge and are added to the Amazon Braket service, you can be confident that your development experience will remain consistent and that your existing designs and quantum algorithms can be tested on these new systems.

Fully managed executions of quantum-classical algorithms with Hybrid Jobs

Amazon Braket Hybrid Jobs simplifies the process of setting up, monitoring, and running hybrid quantum-classical algorithms. After you provide your algorithm script and select the QPU or simulator to run on, Amazon Braket spins up the classical compute, executes the algorithm, and releases the resources once the job is completed. You can define custom metrics for your algorithms, which are automatically logged by Amazon CloudWatch and displayed in real-time in the Amazon Braket console as the algorithm runs. This gives you live insights into how your algorithm is progressing, so you can make adjustments to your algorithm if needed. Most importantly, Hybrid Jobs provides prioritized access to your chosen QPU to help your algorithm execute quickly and predictably, enabling you to improve the quality and reproducibility of results.

Develop variational quantum algorithms with PennyLane

Amazon Braket natively supports PennyLane, an open source software framework built around the concept of quantum differentiable programming, to help you build and run hybrid quantum-classical, or variational, algorithms. This approach enables you to train quantum circuits in the same way that you would train a machine learning neural network to find solutions to computational problems in quantum chemistry, quantum machine learning, and optimization. PennyLane is performance-optimized for Amazon Braket and provides interfaces to familiar machine learning tools, including PyTorch and TensorFlow, to make training quantum circuits fast, easy, and intuitive.

Fully managed Jupyter notebooks

You have the choice of using your own development environment or fully managed Jupyter notebooks in Amazon Braket to build your quantum algorithms and manage experiments. Amazon Braket makes it easy to create notebooks with a single click. You can select the notebook instance type to match your performance requirements and configure security settings such as encryption for stored data. Amazon Braket notebooks come pre-configured with a suite of quantum computing developer tools, including the Amazon Braket SDK and PennyLane to help you get started quickly.

Pre-built algorithms and tutorials

Amazon Braket notebooks come pre-installed with the Amazon Braket SDK, tutorials and a selection of pre-built algorithms to give you everything you need to get started in a single place. Use them to familiarize yourself with the recommended steps to build and execute quantum algorithms using Amazon Braket.

Simulators

Choice of simulation tools

With Amazon Braket, you have a choice of four circuit simulators to run and test quantum algorithms. These include the local simulator that is included in the Amazon Braket SDK and three on-demand simulators. The local simulator can run on a laptop or within an Amazon Braket managed notebook and supports simulation of quantum circuits with and without noise. The on-demand simulators are SV1, a general-purpose state vector simulator; DM1, a density matrix simulator that supports noise modeling; and TN1, a tensor network simulator that specializes in certain larger scale structured quantum circuits.

Consistent experience

You can run a circuit on Amazon Braket simulators with a single API call. A request to run your algorithm on a simulator works in the same way as a request to run on quantum hardware; by changing a single line of code, you can change from running on a simulator to an actual quantum computer.

Choice of result types

You can choose different result types for your simulation tasks, including individual samples, custom observables, individual amplitudes, or the full state vector. Amazon Braket simulators can calculate exact results, or return measurement samples emulating the behavior of quantum computers.

Quantum Computers

Simplified access to quantum computers

Amazon Braket provides secure access to a variety of quantum computing technologies. There is no upfront commitment or contract to sign, and you pay only for what you use through your AWS bill.

Choice of quantum processing units (QPUs)

Amazon Braket provides access to superconducting, trapped ion, neutral-atom, and photonic quantum computers. Following the gate-based quantum computing paradigm, you can access superconducting quantum processors from Oxford Quantum Circuits and Rigetti, and trapped-

ion technology from IonQ. You can also access photonic quantum computers that can be used to study protocols such as Guassian Boson Sampling using the latest QPUs from Xanadu. Finally, you can simulate equilibrium behavior and non-equilibrium dynamics of Hamiltonians of interest to areas of physics such as condensed-matter and high-energy physics, and study optimization problems using the latest neutral-atom processor from QuEra. This helps you test different technologies, compare the compute performance of different machines for the problem that you are trying to solve, and choose the hardware system that is best suited to your application. Please visit the Hardware Providers page to learn more about the QPUs that are available on Amazon Braket.

Expert Help

Amazon Quantum Solutions Lab

Today, quantum computing is still in its infancy, and there are still many unknowns and challenges to solve. The Amazon Quantum Solutions Lab can help. It is a collaborative research and professional services program staffed with quantum computing experts who can assist you to more effectively explore quantum computing and assess the current performance of this nascent technology. Additionally, you can work with our qualified technology and consulting partners in the AWS Partner Network (APN) that specialize in applications for quantum computing and can help you address your specific requirements.

Management and Security

Management Console

As a native AWS service, Amazon Braket is accessible through the AWS Management Console, a centralized and easy to use web interface for Amazon Web Services, which provides you with a secure login using your AWS account or AWS Identity and Access Management (IAM) credentials. You can use the console to manage and monitor your Amazon Braket resources, such as notebooks and tasks, and access detailed information about quantum circuit simulators and QPUs.

User access management, security, and monitoring

Amazon Braket is integrated with AWS services such as Amazon CloudWatch, AWS CloudTrail, Amazon EventBridge, and AWS IAM to enable the monitoring of workloads, generate notifications when your tasks are completed, and manage access controls and permissions. Your simulation and quantum task results are delivered to your preferred Amazon Simple Storage Service (S3) bucket for storage and analysis, giving you full control over your data.

What is Amazon Braket?

Amazon Braket is a fully managed AWS service that helps researchers, scientists, and developers get started with quantum computing. Quantum computing has the potential to solve computational problems that are beyond the reach of classical computers because it harnesses the laws of quantum mechanics to process information in new ways.

Gaining access to quantum computing hardware can be expensive and inconvenient. Limited access makes it difficult to run algorithms, optimize designs, evaluate the current state of the

technology, and plan for when to invest your resources for maximum benefit. Braket helps you overcome these challenges.

Braket offers a single point of access to a variety of quantum computing technologies. With Braket, you can:

- Explore and design quantum and hybrid algorithms.
- Test algorithms on different quantum circuit simulators.
- Run algorithms on different types of quantum computers.
- Create proof of concept applications.

Defining quantum problems and programming quantum computers to solve them requires a new set of skills. To help you gain these skills, Braket offers different environments to simulate and run your quantum algorithms. You can find the approach that best suits your requirements and get started quickly with a set of example environments called *notebooks*.

Braket development has three stages—build, test, and run:

Build - Braket provides fully managed Jupyter notebook environments that make it easy to get started. Braket notebooks are pre-installed with sample algorithms, resources, and developer tools, including the Amazon Braket SDK. With the Amazon Braket SDK, you can build quantum algorithms and then test and run them on different quantum computers and simulators by changing a single line of code.

Test - Braket provides access to fully managed, high-performance quantum circuit simulators. You can test and validate your circuits. Braket handles all the underlying software components and Amazon Elastic Compute Cloud (Amazon EC2) clusters to take away the burden of simulating quantum circuits on classical high performance computing (HPC) infrastructure.

Run - Braket provides secure, on-demand access to different types of quantum computers. You have access to gate-based quantum computers from IonQ, OQC, Rigetti, and Xanadu as well as an analog Hamiltonian simulator from QuEra. You also have no upfront commitment, and no need to procure access through individual providers.

About quantum computing and Braket

Quantum computing is in its early developmental stage. It's important to understand that no universal, fault-tolerant quantum computer exists at present. Therefore, certain types of quantum hardware are better suited for each use case and it's crucial to have access to a variety of computing hardware. Braket offers a variety of hardware through third-party providers.

Existing quantum hardware is limited due to noise, which introduces errors. The industry is in the Noisy Intermediate Scale Quantum (NISQ) era. In the NISQ era, quantum computing devices are too noisy to sustain pure quantum algorithms, such as *Shor's algorithm* or *Grover's algorithm*. Until better quantum error correction is available, the most practical quantum computing requires the combination of classical (traditional) computing resources with quantum computers to create hybrid algorithms. Braket helps you work with *hybrid quantum algorithms.*

In hybrid quantum algorithms, quantum processing units (QPUs) are used as co-processors for CPUs, thus speeding up specific calculations in a classical algorithm. These algorithms utilize iterative processing, in which computation moves between classical and quantum computers. For example, current applications of quantum computing in chemistry, optimization, and machine learning are based on *variational quantum algorithms*, which are a type of *hybrid quantum algorithm*. In variational quantum algorithms, classical optimization routines adjust the parameters of a parameterized quantum circuit iteratively, much in the same way the weights of a neural network are adjusted iteratively based on the error in a machine learning training set. Braket offers access to the PennyLane open source software library, which assists you with *variational quantum algorithms.*

Quantum computing is gaining traction for computations in four main areas:

- **Number theory**—including factoring and cryptography (for example, *Shor's algorithm* is a primary quantum method for number theory computations)
- **Optimization**—including constraint satisfaction, solving linear systems, and machine learning
- **Oracular computing**—including search, hidden subgroups, and order finding (for example, *Grover's algorithm* is a primary quantum method for oracular computations)
- **Simulation**—including direct simulation, knot invariants, and quantum approximate optimization algorithm (QAOA) applications

Applications for these categories of computations can be found in financial services, biotechnology, manufacturing, and pharmaceuticals, to name a few. Braket offers capabilities and example notebooks that can already be applied to many proof of concept problems in addition to certain practical problems.

Amazon Braket FAQs

General

Q: What is Amazon Braket?

Amazon Braket is a fully managed service that helps you get started with quantum computing.

Q: What can I do with Amazon Braket?

With Amazon Braket, you can learn how to program quantum computers and explore potential applications. You can design your own quantum algorithms from scratch or choose from a set of pre-built algorithms. Amazon Braket provides an SDK that you can run locally on a laptop, or in Amazon Braket's fully managed notebook environment. The SDK includes a quantum circuit simulator. The Amazon Braket service also provides fully managed quantum circuit simulators that enable you to run your algorithms on AWS managed infrastructure to validate and test your implementation. When you are ready, you can run your algorithm on Amazon Braket using different quantum computers, or quantum processing units (QPUs), from our hardware providers.

Q: How does Amazon Braket integrate with other AWS services?

Amazon Braket provides integrations with Amazon CloudWatch, Amazon EventBridge, AWS Identity and Access Management (IAM), and AWS CloudTrail for monitoring, event-based processing, user access management, and logs. Your simulation and quantum compute results will be stored in Amazon Simple Storage Service (S3) in your account.

Q: Why should our company be thinking about quantum computing today?

Quantum computing is an early stage technology but its long-term impact promises to be transformational for many industries. Developing quantum algorithms and designing useful quantum applications require new skills and potentially radically different approaches. Building this expertise will take time and requires access to quantum technologies and programming tools. Amazon Braket and the Amazon Quantum Solutions Lab help organizations assess the state of current technologies, identify how they might impact their business, and prepare for the future.

Q: Why is the service named "Braket"?

We named our service after the bra-ket notation, a standard notation in quantum mechanics. It was introduced by Paul Dirac in 1939 to describe the state of quantum systems, and it is also known as the Dirac notation.

Q: Can I perform academic research on Amazon Braket?

Yes. Scientists at universities around the world perform research on Amazon Braket. You can get started in the Amazon Braket console, our Github repository, or request funding for use of Amazon Braket through the AWS Cloud Credit for Research program. In the application process,

Developer Tools

Q: What is the Amazon Braket SDK?

The Amazon Braket Software Development Kit (SDK) is a technology-agnostic developer framework that allows you to develop quantum algorithms and run them on different quantum computing hardware and simulators through the Amazon Braket service. The SDK helps you track and monitor quantum tasks submitted to Amazon Braket and evaluate the results. The Amazon Braket SDK includes a local quantum circuit simulator that you can use to test your algorithms.

Q: How can I access the Amazon Braket SDK?

Amazon Braket provides fully managed Jupyter notebooks that come pre-installed with the Amazon Braket SDK and example tutorials that help you get started quickly. The Amazon

Braket SDK is open sourced so you can use Amazon Braket from any local integrated development environment (IDE) of your choice.

Q. What is pulse-level access?

Today's quantum computers are noisy, and researchers often need access to the lowest level of control over the hardware in order to study noise or cross-talk, develop new and more robust gates, devise error mitigation schemes, and explore novel quantum algorithms. With pulse-control, you can manipulate the low-level, analog signals or pulses that control the qubits of a quantum processor.

Q. Does the Amazon Braket SDK support pulse-level programming?

Yes. You can program quantum circuits on Rigetti Computing and Oxford Quantum Circuits quantum computers directly using gates, pulses, or a combination thereof. You can also selectively insert blocks of pulse instructions in specific areas of the program to focus on optimizing individual operations and fine tune performance.

Q: What is PennyLane?

PennyLane is an open source software library for variational quantum computing that integrates with Amazon Braket. Variational quantum computing is a paradigm that utilizes hybrid quantum-classical algorithms to iteratively find solutions to computational problems in a variety of domains, such as chemistry, optimization, and quantum machine learning. Built around the concept of quantum differentiable programming, PennyLane allows you to train quantum circuits the same way as neural networks. It provides interfaces to popular machine learning libraries, including PyTorch and TensorFlow, to make training your quantum algorithms easy and intuitive. You can learn more about PennyLane at https://pennylane.ai, and read our developer guide here.

Q: Why should I use PennyLane on Amazon Braket?

Near-term quantum computing applications in chemistry, optimization and quantum machine learning are based on variational quantum algorithms that utilize iterative processing between classical and quantum computers. PennyLane makes it easy to get started and build variational and quantum machine learning algorithms on Amazon Braket. It allows you to use familiar tools from machine learning to build and train your algorithms. PennyLane provides a chemistry library, qchem, which you can use to map a computational chemistry problem to a quantum computing formulation with a few lines of code.

Amazon Braket helps you to innovate faster with PennyLane. When testing and fine-tuning your algorithms, our fully managed, high-performance on-demand simulators speed up training by 10x or more, compared to simulating your algorithms locally. To speed up your hybrid quantum algorithms, you can now leverage high performance embedded simulators from PennyLane, such as the lightning.gpu simulator accelerated by NVIDIA's cuQuantum SDK for GPU based workloads. These simulators come with methods such as the adjoint method for gradient computation which reduce the number of circuits required to compute and gradient and can be used for fast iterative experimentation and prototyping.

Q: How can I access PennyLane?

Amazon Braket notebooks come pre-configured with PennyLane, and our tutorial notebooks help you get started quickly. Alternatively, you can install the Amazon Braket PennyLane plugin for any IDE of your choice. The plugin is open source and can be downloaded from here. You can find the PennyLane documentation at https://pennylane.ai.

Q: What is OpenQASM?

OpenQASM is an open source intermediate representation (IR) for quantum computing programs. You can run OpenQASM programs on all gate-based Braket devices either through the Amazon Braket SDK or by directly submitting them to the Braket API. AWS has joined the OpenQASM steering council to help build an open, hardware-agnostic, and unified specification for gate-based quantum programs.

Simulators

Q: Why do I want to simulate my algorithm?

Quantum circuit simulators run on classical computers. With simulators, you can test your quantum algorithms at a lower cost than using quantum hardware, and without having to wait to access specific quantum machines. Simulation is a convenient way to quickly debug quantum circuits and to troubleshoot and optimize algorithms before progressing to run them on quantum hardware. Classical simulation is also essential to verify the results of near-term quantum computing hardware and study the effects of noise.

Q: What simulators does Amazon Braket offer?

Amazon Braket offers you a choice of four quantum circuit simulators: the local simulator in the SDK, and three on-demand simulators: SV1, a general purpose quantum circuit simulator, DM1 that enables you to simulate the effect of noise on your circuits, and TN1, a high performance tensor network simulator. With these options, you can choose the approach that best suits your requirements.

Q: What is the local simulator?

The local simulator is included in the Amazon Braket SDK at no cost. It can run on your laptop, or within an Amazon Braket managed notebook. You can use it for quick validation of circuit designs. It is well suited for small and medium-scale simulations – up to 25 qubits without noise, or up to 12 qubits with noise, depending on your hardware.

Q: What is the SV1 simulator?

SV1 is a fully managed, high-performance state vector simulator for quantum circuits up to 34 qubits. As a state vector simulator, it takes the full wave function of the quantum state and applies the operations of the circuit to calculate the result. After you have designed and debugged your quantum algorithm using the local simulator in the Amazon Braket SDK, you can use SV1 for scaled testing and research. SV1 automatically scales classical compute resources so you can run up to 35 simulations in parallel.

Q: What is the DM1 simulator?

DM1 is a fully managed, density matrix simulator that enables you to investigate the effects of realistic noise on your quantum algorithms. This can help you develop error-mitigation strategies to get more accurate results from today's quantum computing devices.
DM1 supports the simulation of circuits up to 17 qubits. It can run up to 35 simulations in parallel, to speed up your experiments. For rapid prototyping and debugging before using DM1, you can use the local noise simulator in the Amazon Braket SDK.

Q: What is the TN1 simulator?

TN1 is a fully managed, high-performance tensor network simulator that is used for structured quantum circuits up to 50 qubits in size. A tensor network simulator encodes quantum circuits

into a structured graph to find the best way to compute the outcome of the circuit. TN1 is well suited for simulation of sparse circuits, circuits with local gates, and circuits with inherent structure.

Q: How do I choose between the Amazon Braket on-demand simulators SV1, TN1 and DM1?

SV1 is a general purpose simulator based on state vector technology. It provides predictable execution and high performance for universal circuits up to 34 qubits.

DM1 is specifically designed to support noise modeling. If you need to study your algorithms under the effects of various types of noise, use DM1.

TN1 is a specialized simulator for certain types of quantum circuits with up to 50 qubits. Consider it for sparse circuits, circuits with local gates and other circuits with inherent structure. Other circuit types, such as those with all-to-all connectivity between qubits, are often better suited to SV1.

Q: Why would I want to simulate noise in my circuits?

Today's quantum devices are inherently noisy. Every executed operation has a chance of introducing an error. Consequently, the results obtained from a quantum computer generally differ from what is ideally expected. DM1 enables you to study the robustness of your algorithms under the effects of realistic noise, and build error mitigation strategies that help to get more accurate results with today's quantum computing devices.

Q: Can I run a noiseless circuit on the DM1 simulator?

DM1 can simulate circuits without noise. However, for best performance, we recommend using SV1 for large-scale simulations of noise-free circuits.

Q: Do I have to pick an instance type to run a simulation?

Not if you are using an Amazon Braket on-demand simulator. When using SV1, TN1, or DM1, Amazon Braket manages the software and infrastructure for you. You just need to provide the circuit to run.

If you are running the local simulator in the SDK on your Amazon Braket managed notebook, it will run on the Amazon instance you have already specified for your notebook.

Q: How do I know if I can run a circuit on TN1?

As long as your circuit is within the qubit number and circuit depth limits described here, TN1 will attempt to simulate it. In contrast to SV1, however, it is not possible to give an accurate estimate of the runtime based on the qubit number and circuit depth alone. During the so-called "rehearsal phase", TN1 will first try to identify an efficient computational path for your circuit, and estimate the runtime of the next stage, the "contraction phase". If the estimated contraction time exceeds the TN1 limit, TN1 will not attempt contraction and you pay only for the time spent in the rehearsal phase. To learn more, visit the technical documentation.

Q: Do I have to program or design algorithms differently to use a simulator?

No, with Amazon Braket, you can direct the same quantum circuit to run on any simulators and gate-based quantum hardware available on the service by changing a few lines of code.

Q: Do you offer simulators for annealing problems?

On the AWS Marketplace you can find offerings that use the state-of-the-art classical approaches to combinatorial optimization problems such as Meta Analytics and the Toshiba SBM.

Quantum Computers

Q: How do I access quantum computers with Amazon Braket?

Running your circuit design on an actual quantum processing unit (QPU) is easy. Once you have created your circuit or problem graph in the Amazon Braket SDK, you can submit your task from within a managed Jupyter notebook or any IDE of your choice, such as PyCharm.

Q: How is running a task on a QPU different from running on a simulator?

The steps to run a quantum task on a QPU are the same as running on a simulator, you simply chose the back-end, or *device*, when making API calls within the Amazon Braket SDK. They are both compute operations for which you can request different back-ends, or *devices* through API calls within the Amazon Braket SDK. The choice of device includes the various simulators and quantum computers that are available through the service. Switching from one device to another is as easy as changing a single line of code. Nonetheless, simulators are always available, whereas QPU resources might require wait time.

Q: How do I choose which quantum computers to use?

Some types of quantum computers are particularly well suited to solving specific sets of problems. There are many factors that determine which type of machine will meet your needs, such as qubit count, qubit fidelity (error rate), qubit connectivity, coherence time, and cost. Full specifications of the quantum computers are provided in the Amazon Braket console.

Q: What quantum computers does Amazon Braket support?

Please click here to learn more about the hardware providers of Amazon Braket.

Q: What is Analog Hamiltonian Simulation?

Analog Hamiltonian Simulation or AHS, is a different paradigm of quantum computing, distinct from universal, gate-based quantum computing. Devices capable of AHS are special purpose QPUs designed to solve a limited set of problems, represented by a Hamiltonian. AHS allows users to specify a Hamiltonian of interest, and the quantum computer tunes parameters in a way such that it emulates the continuous time-evolution of a quantum state under that Hamiltonian. As these Hamiltonians are directly implemented on the system, AHS systems do not suffer from the overhead required to formulate algorithms in terms of circuits and gates, and thus can already simulate systems with hundreds of qubits, that are prohibitive to simulate classically. Amazon Braket supports AHS via the QuEra QPUs.

Q: Where can I find system and performance information on Rigetti QPUs?

Visit Rigetti's QPU page for system and performance information on the Rigetti QPUs, including gate fidelities and coherence times.

Q: Where can I find best practice recommendations for the IonQ QPU?

Visit the IonQ Best Practices webpage for information on the IonQ QPU's topology, gates, and best practices.

Q: How does the QuEra QPU work?

The QuEra system is a programmable QPU consisting of Rydberg atoms arranged in a two-dimensional customizable arrangement of optical tweezers. The qubits in this device consist of neutral Rubidium-87 atoms with the two states being the ground state of the atom, and one of the highly excited Rydberg state. The Rydberg-Rydberg interaction between atoms decays as the sixth power of the distance between the atoms, giving rise to an effective spin Hamiltonian with local interactions. Additionally, laser fields can tune a transverse magnetic field, to give rise to interesting spin dynamics that can produce novel ground and non-equilibrium states. Researchers accessing this device through Braket will be able to program the geometry of the two-dimensional atom arrangement, and vary the strength of the longitudinal and transverse magnetic fields in a time-dependent manner, all using the Braket SDK. This will produce an effective Hamiltonian of interest, the continuous time evolution of which can be studied on the device. For more information, refer to QuEra's best practices here.

Q: Do my quantum tasks start running immediately on a QPU, or do I have to wait?

Quantum computing is a nascent technology, and quantum computers remain a scarce resource. Different types of quantum computers have different operational characteristics and levels of availability and so process tasks at different rates. If the QPU you select is online and not currently being used, your task will be processed immediately; otherwise it will be put in queue. As the QPU becomes available the tasks in the queue are processed in the order they were received. To notify you when your task has been completed Amazon Braket sends status change events to Amazon EventBridge. You can create a rule in EventBridge to specify an action to take, such as using the Amazon Simple Notification Service (SNS), which can send alerts to you through SMS, or other methods such as email, HTTPs, AWS Lambda or Amazon SQS.

Q: Do I need to compile my circuit before I run them on QPUs?

No, not necessarily. Amazon Braket automatically compiles your code for you when you run it. However you have the option on Rigetti, OQC and IonQ devices to run your circuit as is, without compiler modifications using verbatim compilation. On Rigetti, you can additionally define only specific blocks of code to run as is, without any intervening compiler passes. To learn more, see our documentation on Verbatim Compilation.

Q: What is the benefit of verbatim compilation?

Quantum circuit compilation transforms a quantum circuit to a compiled circuit, which undergoes qubit allocation, mapping to native gates, and optimization. However, compiler gate optimization can be problematic for researchers and quantum algorithm specialists who are developing benchmarking or error-mitigation circuits, as compiler optimizations remove or reorder gates and redundant components. With verbatim compilation, users can specify parts of circuits or entire circuits to run as is, without any compiler modifications.

Q: Can I access quantum computers on the AWS Marketplace?

Yes, you can access quantum hardware via the AWS Marketplace such as D-Wave's annealing devices.

Hybrid Jobs

Q: What is the Hybrid Jobs feature?

Hybrid Jobs make the execution of hybrid quantum-classical workloads easier, faster, and more predictable. With this feature, you only have to provide your algorithm script or container, and AWS will spin up the requested resources, run the algorithm, and release the resources after completion so you only pay for what you use. The Hybrid Jobs feature also provides live

insights into algorithm metrics, so you can see how your algorithm is progressing. Most importantly, jobs have priority access to the target QPU so execution can be faster, more predictable and less affected by other users' workloads.

Q: Why should I use Hybrid Jobs?

Braket Hybrid Jobs provide three main benefits. First, it simplifies running hybrid quantum-classical algorithms. Many quantum researchers are often new to cloud computing and do not want to set up and manage their compute environment before running their hybrid algorithm. With Hybrid Jobs, you only need to specify your preferred compute instance – or use the default. Braket Hybrid Jobs will wait for the target QPU to become available, spin up the classical resources, run the workload in pre-built container environments, return the results to Amazon S3, and finally release the compute resources.

Second, Hybrid Jobs provide live insights into running algorithms. You can define custom algorithm metrics as part of your algorithm that will be automatically logged by Amazon CloudWatch and displayed in the Amazon Braket console. With this, you can track the progress of your algorithms.

Third, Amazon Braket Hybrid Jobs provide better performance than running hybrid algorithms from your own environment. During the entire time that your job is running, it has priority access to the selected QPU. This means tasks executed on that device as part of your job will be executed ahead of other tasks that may be queued up on the device. This results in shorter and more predictable run times for hybrid algorithms, and ultimately better results by reducing the detrimental effects of slowly changing device characteristics ('device drift') on algorithm performance.

Q: Which quantum computers can I use with Hybrid Jobs?

You can use any of the available QPUs on Amazon Braket with Hybrid Jobs.

Q: Which simulators can I use with Hybrid Jobs?

You can use any of the available Amazon Braket on-demand simulators (SV1, DM1, TN1), embedded simulators based on the PennyLane lightning plug-in, or a custom simulator embedded as a container to hybrid jobs. For the embedded simulators or the custom simulator, you can choose one or multiple CPU and GPU instances to run your hybrid workload.

Q: Why should I use embedded simulators with Hybrid Jobs?

Embedded simulators are a set of high-performance simulators that are directly embedded in the same container as your application code to avoid latencies associated with round-trips between a fully-managed, on-demand simulator, such as SV1, and your containerized classical code. Embedded simulators support advanced features such as the adjoint method for gradient computation which reduce the number of circuits needed to compute a gradient. Today, Amazon Braket supports embedded simulators from PennyLane, such as the lightning.gpu simulator, which is accelerated with NVIDIA cuQuantum SDK, specifically designed for running quantum circuit simulation on high-performance GPUs.

Q: Can I bring my own simulator to Amazon Braket Hybrid Jobs?

Yes, you can bring your own simulator library to Amazon Braket Hybrid Jobs by embedding the simulator and its dependencies into a container. You can then pass code to the container as an entry point and execute the code as an Amazon Braket Hybrid Job on CPU or GPU instances. Amazon Braket handles spinning up the resources for the duration of your job and you only pay for what you use.

Q: Do I have to pick an instance type to run a Hybrid Job?

No, by default the jobs container runs on a single ml.m5.xlarge instance type. If you are running a hybrid algorithm using an Amazon Braket on-demand simulator (SV1, TN1, DM1) or a QPU, then Amazon Braket manages the software and infrastructure for you. If you are running a hybrid algorithm using the embedded simulators from PennyLane or a custom simulator packaged as a container, then you can select one or more CPU or GPU instance types on which to run the job. Amazon Braket manages the set up of the underlying infrastructure and releases the resources once the job is complete so you only pay for what you use.

Q: How do I choose between the the embedded state-vector simulator from Penny Lane and the SV1 simulator when running hybrid jobs?

Today, the embedded state-vector simulator from PennyLane that comes pre-installed with Amazon Braket's Hybrid Jobs container can be used for variational algorithms which can benefit from methods such as back-propagation or the adjoint method for gradient computation. Examples of these algorithms are quantum machine learning (QML), quantum adiabatic approximate algorithm (QAOA) or variational quantum eigensolver (VQE) . With embedded simulators you also have the choice to use GPU instances, if your algorithm can benefit from GPU based acceleration, and can fit in the GPU memory. This is generally the case for variational algorithms and QML algorithms with intermediate qubit counts (< 30). Otherwise, consider using the SV1 on-demand simulator. As the adjoint method does not support non-zero shots today, consider using SV1 for any workload where the number of shots is greater than zero. Note that the embedded simulator is only supported as part of the Hybrid Jobs feature, while SV1 supports both standalone tasks and hybrid jobs.

Q: How do I choose between the different embedded simulators from PennyLane?

The PennyLane lightning.gpu simulator can be used for hybrid algorithms such as QML, QAOA or VQE, provided the problem size is small enough to fit within the GPU memory. The lightning.qubit CPU-based simulator can be used for algorithms that are memory-intensive, and cannot fit in GPU memory, such as variational algorithms with high qubit counts (29+ qubits). Note that your costs will differ depending on whether you use a CPU or GPU instance type.

Q: How am I charged for the use of Hybrid Jobs?

There are two components to pricing for the Hybrid Jobs feature: charges for the use of a classical job instance, and charges for the use of quantum computers or quantum circuit simulators. First, you will be charged for the duration the job is running, based on the job instance you use. The Hybrid Jobs feature uses the ml.m5.xlarge instance as the default, or you can choose a different instance type when creating a job. You also have the option to add extra data storage within the compute instance at an additional cost. Second, you will be charged for the execution of quantum tasks that are created as part of your job and run on the quantum computers or circuit simulators of your choice. If you are using one of Amazon Braket's on-demand (SV1, DM1, TN1) simulators or a quantum computer for a portion of your Hybrid Job, then you will be charged for the execution of quantum tasks that are created as part of your job. Pricing for these tasks is the same whether they are run as part of a hybrid job or not. Please refer to the "Quantum Computers" and "Simulators" tabs on the Amazon Braket pricing page. If you are using an embedded simulator such as the lightning simulator which comes pre-installed with the managed Hybrid Jobs containers on Amazon Braket, or a simulator of your choice embedded as a custom container, then you only pay for the classical CPU or GPU resources you use for the duration of the job, based on the pricing table below.

Q: How can I get started with Hybrid Jobs?

You can get started by visiting the Amazon Braket Jobs User Guide section of Braket documentation. Amazon Braket hybrid example notebooks provide tutorials on how to get started with Jobs, and run different types of hybrid algorithms. These examples come pre-installed on Amazon Braket notebooks to help you get started quickly. You can also review hybrid algorithm examples with the PennyLane plug-in in the Amazon Braket examples repository.

Security

Q: Does my data leave the AWS environment when I am using Amazon Braket Services?

Yes, QPUs on Amazon Braket are hosted by our third-party quantum hardware providers. If you use Amazon Braket to access quantum computers, your circuit and associated metadata will be sent to and processed by the hardware providers outside of facilities operated by AWS. Your content is anonymized so that only content necessary to process the quantum task is sent to the quantum hardware providers. AWS account information is not transmitted to them. All data is encrypted at rest and in transit, and is only decrypted for processing. In addition, Amazon Braket hardware providers are not permitted to store or use your content for purposes other than processing your task. Once the circuit completes, the results are returned to Amazon Braket and stored in your Amazon S3 bucket. The security of Amazon Braket third-party quantum hardware providers is periodically audited to ensure standards of network security, access control, data protection, and physical security are met.

Q: Where will my results be stored?

Your results will be stored in Amazon S3. In addition to providing the results of the execution, Amazon Braket also publishes event logs and performance metrics such as completion status and execution time to Amazon CloudWatch.

Q: Can I use Amazon Braket in my Amazon Virtual Private Cloud (Amazon VPC)?

Amazon Braket is integrated with AWS PrivateLink, so you can access Amazon Braket from within your Amazon Virtual Private Cloud (Amazon VPC) without requiring the traffic to traverse across the Internet. This reduces exposure to security threats from Internet-based attacks and the risk of sensitive data leakage.

Quantum Solutions Lab

Q: What’s the Quantum Solutions Lab (QSL)?

The Amazon Quantum Solutions Lab is a collaborative research and professional services program staffed with quantum computing experts who can assist you to more effectively explore quantum computing and work to overcome the challenges that arise with this nascent technology. Please visit the Quantum Solutions Lab webpage to get started.

Q: How can I engage with the QSL?

You can request information about engagements with the QSL and our partners by submitting this form, and by working through your AWS account manager.

Q: What is the cost of a QSL engagement?

The cost of QSL engagements vary depending on the length of the engagement and nature of your needs. Please reach out to your account manager for more details.

Q: What's the typical duration of a QSL engagement?

Quantum Solutions Lab engagements typically last from 6 to 12 months.

Q: Do I need to travel to the Lab to participate?

The entire process can be done remotely, if needed, which is likely during the current pandemic. However, typically we meet in person to kick off the engagements and determine a working cadence. After that, we will visit your site as needed and have regular checkpoints using video conferencing, while collaborating remotely on a regular basis.

AWS Center for Quantum Computing

Q: What's AWS Center for Quantum Computing?

The AWS Center for Quantum Computing is a research program that brings together researchers and engineers from Amazon and academic institutions that are leaders in the field of quantum computing. Together they collaborate on near-term applications, error-correction schemes, hardware architectures, and programming models to explore the development of quantum technologies. We established the AWS Center for Quantum Computing on the campus of California Institute of Technology (Caltech). Today, the Center collaborates with researchers at Caltech, Stanford University, Harvard University, Massachusetts Institute of Technology, and the University of Chicago through the Amazon Scholars program.

Q: What research has the AWS Center for Quantum Computing published?

The AWS Center for Quantum Computing team regularly publishes research and presents scientific papers at conferences such as QIP, APS and IEEE QCE on quantum hardware, algorithms, error correction and other domains. Notable research includes a paper on "Designing a fault-tolerant quantum computer based on Schrödinger-cat qubits." For other research publications, please see our Amazon.Science Quantum Technologies research area page.

AWS Elastic Compute Cloud (EC2)

Secure and resizable compute capacity for virtually any workload.

Amazon Elastic Compute Cloud (Amazon EC2) offers the broadest and deepest compute platform, with over 500 instances and choice of the latest processor, storage, networking, operating system, and purchase model to help you best match the needs of your workload. We are the first major cloud provider that supports Intel, AMD, and Arm processors, the only cloud with on-demand EC2 Mac instances, and the only cloud with 400 Gbps Ethernet networking. We offer the best price performance for machine learning training, as well as the lowest cost per inference instances in the cloud. More SAP, high performance computing (HPC), ML, and Windows workloads run on AWS than any other cloud.

Use cases

Run cloud-native and enterprise applications

Amazon EC2 delivers secure, reliable, high-performance, and cost-effective compute infrastructure to meet demanding business needs.

Scale for HPC applications

Access the on-demand infrastructure and capacity you need to run HPC applications faster and cost-effectively.

Develop for Apple platforms

Build, test, and sign on-demand macOS workloads. Access environments in minutes, dynamically scale capacity as needed, and benefit from AWS's pay-as-you-go pricing.

Train and deploy ML applications

Amazon EC2 delivers the broadest choice of compute, networking (up to 400 Gbps), and storage services purpose-built to optimize price performance for ML projects.

Amazon EC2 features

Amazon EC2 provides the broadest and deepest instance choice to match your workload's needs. General purpose, compute optimized, memory optimized, storage optimized, and accelerated computing instance types are available that provide the optimal compute, memory, storage, and networking balance for your workloads. Processors from Intel, AMD, NVIDIA and AWS power these instance types and provide additional performance and cost optimizations. Local storage and enhanced networking options available with instance types further help optimize performance for workloads that are disk or network I/O bound. Many instance types also offer bare metal instances that provide your applications with direct access to the processor and memory of the underlying server for running in non-virtualized environments or for applications where you want to use your own hypervisor. To find the right instance for your workload, visit the EC2 instance types page. You can also use the AWS Compute Optimizer to get recommendations on optimal AWS Compute resources for your workloads to reduce costs and improve performance.

Global Infrastructure

Multiple Locations

Amazon EC2 provides the ability to place instances in multiple locations. Amazon EC2 locations are composed of Regions and Availability Zones. Availability Zones are distinct locations that are engineered to be insulated from failures in other Availability Zones and provide inexpensive, low latency network connectivity to other Availability Zones in the same Region. By launching instances in separate Availability Zones, you can protect your applications from failure of a single location. Regions consist of one or more Availability Zones and are geographically dispersed. The Amazon EC2 Service Level Agreement commitment is 99.99% availability for each Amazon EC2 Region.

High Precision Time with Amazon Time Sync Service

The Amazon Time Sync Service provides a highly accurate, reliable and available time source to AWS services including EC2 instances.

Choice of operating systems and software

Amazon Machine Images (AMIs) are preconfigured with an ever-growing list of operating systems, including Microsoft Windows and Linux distributions such as Amazon Linux 2, Ubuntu, Red Hat Enterprise Linux, CentOS, SUSE and Debian. We work with our partners and community to provide you with the most choice possible. The AWS Marketplace features a wide selection of commercial and free software from well-known vendors, designed to run on your EC2 instances.

Cost and Capacity Optimization

Pay for What You Use

With per-second billing, you only pay for what you use. It takes the cost of unused minutes and seconds in an hour off of the bill, so you can focus on improving your applications instead of maximizing usage to the hour.

Scale Seamlessly with Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling allows you to automatically scale your Amazon EC2 capacity up or down according to conditions you define. You can use the dynamic and predictive scaling policies within EC2 Auto Scaling to add or remove EC2 instances. Predictive scaling uses machine learning to proactively allocate instances based on anticipated demand, and dynamic scaling allows you to scale compute based on defined metrics. With EC2 Auto Scaling, you can ensure that the number of Amazon EC2 instances you're using scales up seamlessly during demand spikes to maintain performance, and scales down automatically during demand lulls to minimize costs.

Optimize Compute Performance and Cost with Amazon EC2 Fleet

With a single API call, Amazon EC2 Fleet lets you provision compute capacity across EC2 instance types, Availability Zones, and purchase models to help optimize scale, performance and cost. Read FAQs and this AWS blog to learn more. You can also access EC2 Fleet capabilities via Amazon EC2 Auto Scaling to provision and automatically scale compute capacity across EC2 instance types, Availability Zones, and purchase options in a single Auto Scaling Group.

Optimized CPU Configurations

The Optimize CPUs feature gives you greater control of your Amazon EC2 instances on two fronts. First, you can specify a custom number of vCPUs when launching new instances to save on vCPU-based licensing costs. Second, you can disable Intel Hyper-Threading Technology (Intel HT Technology) for workloads that perform well with single-threaded CPUs, such as certain high-performance computing (HPC) applications.

Pause and Resume Your Instances

You can hibernate your Amazon EC2 instances backed by Amazon EBS, and resume them from this state at a later time. Applications that take a while to bootstrap and persist state into memory (RAM) can benefit from this feature.

Storage

Optimal storage for every workload

Different Amazon EC2 workloads can have vastly different storage requirements. Beyond the built-in instance storage, we also offer Amazon Elastic Block Store (Amazon EBS) and Amazon Elastic File System (Amazon EFS) to suit other cloud storage workload requirements. Amazon EBS provides persistent, highly available, consistent, low-latency block storage volumes for use with Amazon EC2 instances, while Amazon EFS provides simple, scalable, persistent, fully managed cloud file storage for shared access.

Networking

High Packet-Per-Second Performance and Low Latency with Enhanced Networking
Enhanced Networking enables you to get significantly higher packet per second (PPS) performance, lower network jitter and lower latencies. This feature uses a network virtualization stack that provides higher I/O performance and lower CPU utilization compared to traditional

implementations. For instructions on how to enable Enhanced Networking on EC2 instances, see the Enhanced Networking on Linux and Enhanced Networking on Windows tutorials.

Run High Levels of Inter-Node Communications with Elastic Fabric Adapter

Elastic Fabric Adapter (EFA) is a network interface for Amazon EC2 instances that enables customers to run applications requiring high levels of inter-instance communications, like machine learning, computational fluid dynamics, weather modeling, and reservoir simulation, at scale on AWS. EFA is available as an optional EC2 networking feature that you can enable on any supported EC2 instance at no additional cost.

Manage Dynamic Cloud Computing Services with Elastic IP Addresses

Elastic IP addresses are static IP addresses designed for dynamic cloud computing. An Elastic IP address is associated with your account, not with a particular instance, and you control that address until you choose to explicitly release it. Unlike traditional static IP addresses, however, Elastic IP addresses allow you to mask instance or Availability Zone failures by programmatically remapping your public IP addresses to any instance in your account. You can also optionally configure the reverse DNS record of any of your Elastic IP addresses by filling out this form.

High Throughput and Low Latency with High Performance Computing (HPC) Clusters

Customers with complex computational workloads such as tightly coupled parallel processes, or with applications sensitive to network performance, can achieve the same high compute and network performance provided by custom-built infrastructure while benefiting from the elasticity, flexibility and cost advantages of Amazon EC2. Cluster Compute, Cluster GPU, and High Memory Cluster instances have been specifically engineered to provide high-performance network capability and can be programmatically launched into clusters – allowing applications to get the low-latency network performance required for tightly coupled, node-to-node communication. Cluster instances also provide significantly increased throughput making them well suited for customer applications that need to perform network-intensive operations.

Access Services Hosted on AWS Easily and Securely with AWS PrivateLink

AWS PrivateLink is a purpose-built technology designed for customers to access Amazon services in a highly performant and highly available manner, while keeping all the network traffic within the AWS network.

Operating Systems and Software

Amazon Machine Images (AMIs) are preconfigured with an ever-growing list of operating systems, including Microsoft Windows and Linux distributions such as Amazon Linux 2, Ubuntu, Red Hat Enterprise Linux, CentOS, SUSE and Debian. We work with our partners and community to provide you with the most choice possible. The AWS Marketplace features a wide selection of commercial and free software from well-known vendors, designed to run on your EC2 instances.

Maintenance

AWS regularly performs routine hardware, software, power, and network maintenance with minimal disruption across all EC2 instance types. This is achieved by a combination of technologies and methods across the entire AWS Global infrastructure, such as live update and live migration as well as redundant and concurrently maintainable systems. Non-intrusive maintenance technologies such as live update and live migration do not require instances to be

stopped or rebooted. Customers are not required to take any action prior to, during or after live migration or live update. These technologies help improve application uptime and reduce your operational effort. Amazon EC2 uses live update to deploy software to servers quickly with minimal impact to customer instances. Live update ensures that customers' workloads run on servers with software that is up-to-date with security patches, new instance features and performance improvements. Amazon EC2 uses live migration when running instances need to be moved from one server to another for hardware maintenance or to optimize placement of instances or to dynamically manage CPU resources. Amazon EC2 has been expanding the scope and coverage of non-intrusive maintenance technologies over the years so that scheduled maintenance events are a fallback option rather than the primary means of enabling routine maintenance.

What is Amazon EC2?

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) Cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic.

Related services

You can provision Amazon EC2 resources, such as instances and volumes, directly using Amazon EC2. You can also provision Amazon EC2 resources using other services in AWS.

Amazon EC2 Auto Scaling User Guide

- AWS CloudFormation User Guide
- AWS Elastic Beanstalk Developer Guide
- AWS OpsWorks User Guide

To automatically distribute incoming application traffic across multiple instances, use Elastic Load Balancing.

To get a managed relational database in the cloud, use Amazon Relational Database Service (Amazon RDS) to launch a database instance. Although you can set up a database on an EC2 instance, Amazon RDS offers the advantage of handling your database management tasks, such as patching the software, backing up, and storing the backups.

To make it easier to manage Docker containers on a cluster of EC2 instances, use Amazon Elastic Container Service (Amazon ECS).

To monitor basic statistics for your instances and Amazon EBS volumes, use Amazon CloudWatch.

To detect potentially unauthorized or malicious use of your EC2 instances, use Amazon GuardDuty.

PCI DSS compliance

Amazon EC2 supports the processing, storage, and transmission of credit card data by a merchant or service provider, and has been validated as being compliant with Payment Card Industry (PCI) Data Security Standard (DSS).

Amazon EC2 FAQs

General

Overview

Q: What is Amazon Elastic Compute Cloud (Amazon EC2)?

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Q: What can I do with Amazon EC2?

Just as Amazon Simple Storage Service (Amazon S3) enables storage in the cloud, Amazon EC2 enables “compute” in the cloud. Amazon EC2’s simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon’s proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.

Q: How can I get started with Amazon EC2?

To sign up for Amazon EC2, click the “Sign up for This Web Service” button on the Amazon EC2 detail page. You must have an AWS account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon EC2 sign-up process. After signing up, please refer to the Amazon EC2 documentation, which includes our Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for Amazon EC2?

Amazon EC2 registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a couple of minutes and involves receiving a phone call during the registration process and entering a PIN number using the phone key pad.

Q: What can developers now do that they could not before?

Until now, small developers did not have the capital to acquire massive compute resources and ensure they had the capacity they needed to handle unexpected spikes in load. Amazon EC2 enables any developer to leverage Amazon’s own benefits of massive scale with no upfront investment or performance compromises. Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure they have the compute capacity they need to meet their business requirements.

The “Elastic” nature of the service allows developers to instantly scale to meet spikes in traffic or demand. When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Q: How do I run systems in the Amazon EC2 environment?

Once you have set up your account and select or create your AMIs, you are ready to boot your instance. You can start your AMI on any number of On-Demand instances by using the RunInstances API call. You simply need to indicate how many instances you wish to launch. If you wish to run more than your On-Demand quota, complete the Amazon EC2 instance request form.

If Amazon EC2 is able to fulfill your request, RunInstances will return success, and we will start launching your instances. You can check on the status of your instances using the DescribeInstances API call. You can also programmatically terminate any number of your instances using the TerminateInstances API call.

If you have a running instance using an Amazon EBS boot partition, you can also use the StopInstances API call to release the compute resources but preserve the data on the boot partition. You can use the StartInstances API when you are ready to restart the associated instance with the Amazon EBS boot partition.

In addition, you have the option to use Spot Instances to reduce your computing costs when you have flexibility in when your applications can run. Read more about Spot Instances for a more detailed explanation on how Spot Instances work.

If you prefer, you can also perform all these actions from the AWS Management Console or through the command line using our command line tools, which have been implemented with this web service API.

Q: What is the difference between using the local instance store and Amazon Elastic Block Store (Amazon EBS) for the root device?

When you launch your Amazon EC2 instances you have the ability to store your root device data on Amazon EBS or the local instance store. By using Amazon EBS, data on the root device will persist independently from the lifetime of the instance. This enables you to stop and restart the instance at a subsequent time, which is similar to shutting down your laptop and restarting it when you need it again.

Alternatively, the local instance store only persists during the life of the instance. This is an inexpensive way to launch instances where data is not stored to the root device. For example, some customers use this option to run large web sites where each instance is a clone to handle web traffic.

Q: How quickly will systems be running?

It typically takes less than 10 minutes from the issue of the RunInstances call to the point where all requested instances begin their boot sequences. This time depends on a number of factors including: the size of your AMI, the number of instances you are launching, and how recently you have launched that AMI. Images launched for the first time may take slightly longer to boot.

Q: How do I load and store my systems with Amazon EC2?

Amazon EC2 allows you to set up and configure everything about your instances from your operating system up to your applications. An Amazon Machine Image (AMI) is simply a packaged-up environment that includes all the necessary bits to set up and boot your instance. Your AMIs are your unit of deployment. You might have just one AMI or you might compose your system out of several building block AMIs (e.g., webservers, appservers, and databases). Amazon EC2 provides a number of tools to make creating an AMI easy. Once you create a custom AMI, you will need to bundle it. If you are bundling an image with a root device backed by Amazon EBS, you can simply use the bundle command in the AWS Management Console. If you are bundling an image with a boot partition on the instance store, then you will need to use the AMI Tools to upload it to Amazon S3. Amazon EC2 uses Amazon EBS and Amazon S3 to provide reliable, scalable storage of your AMIs so that we can boot them when you ask us to do so.

Or, if you want, you don't have to set up your own AMI from scratch. You can choose from a number of globally available AMIs that provide useful instances. For example, if you just want a simple Linux server, you can choose one of the standard Linux distribution AMIs.

Q: How do I access my systems?

The RunInstances call that initiates execution of your application stack will return a set of DNS names, one for each system that is being booted. This name can be used to access the system exactly as you would if it were in your own data center. You own that machine while your operating system stack is executing on it.

Q: Is Amazon EC2 used in conjunction with Amazon S3?

Yes, Amazon EC2 is used jointly with Amazon S3 for instances with root devices backed by local instance storage. By using Amazon S3, developers have access to the same highly scalable, reliable, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. In order to execute systems in the Amazon EC2 environment, developers use the tools provided to load their AMIs into Amazon S3 and to move them between Amazon S3 and Amazon EC2.

We expect developers to find the combination of Amazon EC2 and Amazon S3 to be very useful. Amazon EC2 provides cheap, scalable compute in the cloud while Amazon S3 allows users to store their data reliably.

Q: How many instances can I run in Amazon EC2?

You are limited to running On-Demand Instances per your vCPU-based On-Demand Instance limit, purchasing 20 Reserved Instances, and requesting Spot Instances per your dynamic Spot limit per region. New AWS accounts may start with limits that are lower than the limits described here.

If you need more instances, complete the Amazon EC2 limit increase request form with your use case, and your limit increase will be considered. Limit increases are tied to the region they were requested for.

Q: Are there any limitations in sending email from Amazon EC2 instances?

Yes. In order to maintain the quality of Amazon EC2 addresses for sending email, we enforce default limits on the amount of email that can be sent from EC2 accounts. If you wish to send

larger amounts of email from EC2, you can apply to have these limits removed from your account by filling out this form.

Q: How quickly can I scale my capacity both up and down?

Amazon EC2 provides a truly elastic computing environment. Amazon EC2 enables you to increase or decrease capacity within minutes, not hours or days. You can commission one, hundreds or even thousands of server instances simultaneously. When you need more instances, you simply call RunInstances, and Amazon EC2 will typically set up your new instances in a matter of minutes. Of course, because this is all controlled with web service APIs, your application can automatically scale itself up and down depending on its needs.

Q: What operating system environments are supported?

Amazon EC2 currently supports a variety of operating systems including: Amazon Linux, Ubuntu, Windows Server, Red Hat Enterprise Linux, SUSE Linux Enterprise Server, openSUSE Leap, Fedora, Fedora CoreOS, Debian, CentOS, Gentoo Linux, Oracle Linux, and FreeBSD. We are looking for ways to expand it to other platforms.

Q: Does Amazon EC2 use ECC memory?

In our experience, ECC memory is necessary for server infrastructure, and all the hardware underlying Amazon EC2 uses ECC memory.

Q: How is this service different than a plain hosting service?

Traditional hosting services generally provide a pre-configured resource for a fixed amount of time and at a predetermined cost. Amazon EC2 differs fundamentally in the flexibility, control and significant cost savings it offers developers, allowing them to treat Amazon EC2 as their own personal data center with the benefit of Amazon.com's robust infrastructure.

When computing requirements unexpectedly change (up or down), Amazon EC2 can instantly respond, meaning that developers have the ability to control how many resources are in use at any given point in time. In contrast, traditional hosting services generally provide a fixed number of resources for a fixed amount of time, meaning that users have a limited ability to easily respond when their usage is rapidly changing, unpredictable, or is known to experience large peaks at various intervals.

Secondly, many hosting services don't provide full control over the compute resources being provided. Using Amazon EC2, developers can choose not only to initiate or shut down instances at any time, they can completely customize the configuration of their instances to suit their needs – and change it at any time. Most hosting services cater more towards groups of users with similar system requirements, and so offer limited ability to change these.

Finally, with Amazon EC2 developers enjoy the benefit of paying only for their actual resource consumption – and at very low rates. Most hosting services require users to pay a fixed, upfront fee irrespective of their actual computing power used, and so users risk overbuying resources to compensate for the inability to quickly scale up resources within a short time frame.

EC2 On-Demand Instance limits

Q: What is changing?

Amazon EC2 is transitioning On-Demand Instance limits from the current instance count-based limits to the new vCPU-based limits to simplify the limit management experience for AWS

customers. Usage toward the vCPU-based limit is measured in terms of number of vCPUs (virtual central processing units) for the Amazon EC2 Instance Types to launch any combination of instance types that meet your application needs.

Q: What are vCPU-based limits?

You are limited to running one or more On-Demand Instances in an AWS account, and Amazon EC2 measures usage towards each limit based on the total number of vCPUs (virtual central processing unit) that are assigned to the running On-Demand instances in your AWS account. The following table shows the number of vCPUs for each instance size. The vCPU mapping for some instance types may differ; see Amazon EC2 Instance Types for details.

Instance Size	vCPUs
nano	1
micro	1
small	1
medium	1
large	2
xlarge	4
2xlarge	8
3xlarge	12
4xlarge	16
8xlarge	32
9xlarge	36
10xlarge	40
12xlarge	48
16xlarge	64
18xlarge	72
24xlarge	96
32xlarge	128

Q: How many On-Demand instances can I run in Amazon EC2?

There are five vCPU-based instance limits; each defines the amount of capacity you can use of a given instance family. All usage of instances in a given family, regardless of generation, size, or configuration variant (e.g. disk, processor type), will accrue towards the family's total vCPU limit, listed in the table below. New AWS accounts may start with limits that are lower than the limits described here.

On-Demand Instance Limit Name	Default vCPU Limit
Running On-Demand Standard (A, C, D, H, I, M, R, T, Z) instances	1152 vCPUs
Running On-Demand F instances	128 vCPUs
Running On-Demand G instances	128 vCPUs

Running On-Demand Inf instances	128 vCPUs
Running On-Demand P instances	128 vCPUs
Running On-Demand X instances	128 vCPUs

Q: Are these On-Demand Instance vCPU-based limits regional?

Yes, the On-Demand Instance limits for an AWS account are set on a per-region basis.

Q: Will these limits change over time?

Yes, limits can change over time. Amazon EC2 is constantly monitoring your usage within each region and your limits are raised automatically based on your use of EC2.

Q: How can I request a limit increase?

Even though EC2 automatically increases your On-Demand Instance limits based on your usage, if needed you can request a limit increase from the Limits Page on Amazon EC2 console, the Amazon EC2 service page on the Service Quotas console, or the Service Quotas API/CLI.

Q: How can I calculate my new vCPU limit?

You can find the vCPU mapping for each of the Amazon EC2 Instance Types or use the simplified vCPU Calculator to compute the total vCPU limit requirements for your AWS account.

Q: Do vCPU limits apply when purchasing Reserved Instances or requesting Spot Instances?

No, the vCPU-based limits only apply to running On-Demand instances and Spot Instances.

Q: How can I view my current On-Demand Instance limits?

You can find your current On-Demand Instance limits on the EC2 Service Limits page in the Amazon EC2 console, or from the Service Quotas console and APIs.

Q: Will this affect running instances?

No, opting into vCPU-based limits will not affect any running instances.

Q: Can I still launch the same number of instances?

Yes, the vCPU-based instance limits allow you to launch at least the same number of instances as count-based instance limits.

Q: Will I be able to view instance usage against these limits?

With the Amazon CloudWatch metrics integration, you can view EC2 usage against limits in the Service Quotas console. Service Quotas also enables customers to use CloudWatch for configuring alarms to warn customers of approaching limits. In addition, you can continue to track and inspect your instance usage in Trusted Advisor and Limit Monitor.

Q: Will I still be able to use the DescribeAccountAttributes API?

With the vCPU limits, we no longer have total instance limits governing the usage. Hence the DescribeAccountAttributes API will no longer return the max-instances value. Instead you

can now use the Service Quotas APIs to retrieve information about EC2 limits. You can find more information about the Service Quotas APIs in the AWS documentation.

Q: Will the vCPU limits have an impact on my monthly bill?

No. EC2 usage is still calculated either by the hour or the second, depending on which AMI you're running and the instance type and size you've launched.

Q: Will vCPU limits be available in all Regions?

vCPU-based instance limits are available in all commercial AWS Regions.

Changes to EC2 SMTP endpoint policy

Q. What is changing?

Starting Jan-27 2020, Amazon Elastic Compute Cloud (EC2) will begin rolling out a change to restrict email traffic over port 25 by default to protect customers and other recipients from spam and email abuse. Port 25 is typically used as the default SMTP port to send emails. AWS accounts that have requested and had Port 25 throttles removed in the past will not be impacted by this change.

Q. I have a valid use-case for sending emails to port 25 from EC2. How can I have these port 25 restrictions removed?

If you have a valid use-case for sending emails to port 25 (SMTP) from EC2, please submit a Request to Remove Email Sending Limitations to have these restrictions lifted. You can alternately send emails using a different port, or leverage an existing authenticated email relay service such as Amazon Simple Email Service (SES).

Instance types

Accelerated Computing instances

Q: What are Accelerated Computing instances?

Accelerated Computing instance family is a family of instances that use hardware accelerators, or co-processors, to perform some functions, such as floating-point number calculation and graphics processing, more efficiently than is possible in software running on CPUs. Amazon EC2 provides three types of Accelerated Computing instances – GPU compute instances for general-purpose computing, GPU graphics instances for graphics intensive applications, and FPGA programmable hardware compute instances for advanced scientific workloads.

Q. When should I use GPU Graphics and Compute instances?

GPU instances work best for applications with massive parallelism such as workloads using thousands of threads. Graphics processing is an example with huge computational requirements, where each of the tasks is relatively small, the set of operations performed form a pipeline, and the throughput of this pipeline is more important than the latency of the individual operations. To be able to build applications that exploit this level of parallelism, one needs GPU device specific knowledge by understanding how to program against various graphics APIs (DirectX, OpenGL) or GPU compute programming models (CUDA, OpenCL).

Q: What applications can benefit from P4d?

Some of the applications that we expect customers to use P4d for are machine learning (ML) workloads like natural language understanding, perception model training for autonomous vehicles, image classification, object detection and recommendation engines. The increased GPU performance can significantly reduce the time to train and the additional GPU memory will help customers train larger, more complex models. HPC customers can use P4's increased processing performance and GPU memory for seismic analysis, drug discovery, DNA sequencing, and insurance risk modeling.

Q: How do P4d instances compare to P3 instances?

P4 instances feature NVIDIA's latest generation A100 Tensor Core GPUs to provide on average 2.5X increase in TFLOP performance over the previous generation V100 along with 2.5X the GPU memory. P4 instances feature Cascade Lake Intel CPU that has 24C per socket and an additional instruction set for vector neural network instructions. P4 instances will have 1.5X the total system memory and 4X the networking throughput of P3dn or 16x compared to P3.16xl. Another key difference is that the NVSwitch GPU interconnect throughput will double what was possible on P3 so each GPU can communicate with every other GPU at the same 600GB/s bidirectional throughput and with single-hop latency. This allows application development to consider multiple GPUs and memories as a single large GPU and a unified pool of memory. P4d instances are also deployed in tightly coupled hyperscale clusters, called EC2 UltraClusters, that enable you to run the most complex multi-node ML training and HPC applications.

Q: What are EC2 UltraClusters and how can I get access?

P4d instances are deployed in hyperscale clusters called EC2 UltraClusters. Each EC2 UltraCluster is comprised of more than 4,000 NVIDIA A100 Tensor Core GPUs, Petabit-scale networking, and scalable low latency storage with FSx for Lustre. Each EC2 UltraCluster is one of the world's top supercomputers. Anyone can easily spin up P4d instances in EC2 SuperClusters.

Q: Will AMIs I used on P3 and P3dn work on P4?

The P4 AMIs will need new NVIDIA drivers for the A100 GPUs and a newer version of the ENA driver installed. P4 instances are powered by Nitro System and they require AMIs with NVMe and ENA driver installed. P4 also comes with new Intel Cascade Lake CPUs, which come with an updated instruction set, so we recommend using the latest distributions of ML frameworks, which take advantage of these new instruction sets for data pre-processing.

Q: How are P3 instances different from G3 instances?

P3 instances are the next-generation of EC2 general-purpose GPU computing instances, powered by up to 8 of the latest-generation NVIDIA Tesla V100 GPUs. These new instances significantly improve performance and scalability, and add many new features, including new Streaming Multiprocessor (SM) architecture for machine learning (ML)/deep learning (DL) performance optimization, second-generation NVIDIA NVLink high-speed GPU interconnect, and highly tuned HBM2 memory for higher-efficiency.

G3 instances use NVIDIA Tesla M60 GPUs and provide a high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA Tesla M60 GPUs support NVIDIA GRID Virtual Workstation features, and H.265 (HEVC) hardware encoding. Each M60 GPU in G3 instances supports 4 monitors with resolutions up to 4096x2160, and is licensed to use NVIDIA GRID Virtual Workstation for one Concurrent Connected User. Example applications of G3 instances include 3D visualizations, graphics-intensive remote workstation, 3D rendering, application streaming, video encoding, and other server-side graphics workloads.

Q: What are the benefits of NVIDIA Volta GV100 GPUs?

The new NVIDIA Tesla V100 accelerator incorporates the powerful new Volta GV100 GPU. GV100 not only builds upon the advances of its predecessor, the Pascal GP100 GPU, it significantly improves performance and scalability, and adds many new features that improve programmability. These advances will supercharge HPC, data center, supercomputer, and deep learning systems and applications.

Q: Who will benefit from P3 instances?

P3 instances with their high computational performance will benefit users in artificial intelligence (AI), machine learning (ML), deep learning (DL) and high performance computing (HPC) applications. Users include data scientists, data architects, data analysts, scientific researchers, ML engineers, IT managers and software developers. Key industries include transportation, energy/oil & gas, financial services (banking, insurance), healthcare, pharmaceutical, sciences, IT, retail, manufacturing, high-tech, transportation, government, and academia, among many others.

Q: What are some key use cases of P3 instances?

P3 instances use GPUs to accelerate numerous deep learning systems and applications including autonomous vehicle platforms, speech, image, and text recognition systems, intelligent video analytics, molecular simulations, drug discovery, disease diagnosis, weather forecasting, big data analytics, financial modeling, robotics, factory automation, real-time language translation, online search optimizations, and personalized user recommendations, to name just a few.

Q: Why should customers use GPU-powered Amazon P3 instances for AI/ML and HPC?

GPU-based compute instances provide greater throughput and performance because they are designed for massively parallel processing using thousands of specialized cores per GPU, versus CPUs offering sequential processing with a few cores. In addition, developers have built hundreds of GPU-optimized scientific HPC applications such as quantum chemistry, molecular dynamics, and meteorology, among many others. Research indicates that over 70% of the most popular HPC applications provide built-in support for GPUs.

Q: Will P3 instances support EC2 Classic networking and Amazon VPC?

P3 instances will support VPC only.

Q: How are G3 instances different from P2 instances?

G3 instances use NVIDIA Tesla M60 GPUs and provide a high-performance platform for graphics applications using DirectX or OpenGL. NVIDIA Tesla M60 GPUs support NVIDIA GRID Virtual Workstation features, and H.265 (HEVC) hardware encoding. Each M60 GPU in G3 instances supports 4 monitors with resolutions up to 4096x2160, and is licensed to use NVIDIA GRID Virtual Workstation for one Concurrent Connected User. Example applications of G3 instances include 3D visualizations, graphics-intensive remote workstation, 3D rendering, application streaming, video encoding, and other server-side graphics workloads.

P2 instances use NVIDIA Tesla K80 GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. P2 instances provide customers with high bandwidth 25 Gbps networking, powerful single and double precision floating-point capabilities, and error-correcting code (ECC) memory, making them ideal for deep learning, high performance databases, computational fluid dynamics, computational finance, seismic analysis, molecular modeling, genomics, rendering, and other server-side GPU compute workloads.

Q: How are P3 instances different from P2 instances?

P3 Instances are the next-generation of EC2 general-purpose GPU computing instances, powered by up to 8 of the latest-generation NVIDIA Volta GV100 GPUs. These new instances significantly improve performance and scalability and add many new features, including new Streaming Multiprocessor (SM) architecture, optimized for machine learning (ML)/deep learning (DL) performance, second-generation NVIDIA NVLink high-speed GPU interconnect, and highly tuned HBM2 memory for higher-efficiency.

P2 instances use NVIDIA Tesla K80 GPUs and are designed for general purpose GPU computing using the CUDA or OpenCL programming models. P2 instances provide customers with high bandwidth 25 Gbps networking, powerful single and double precision floating-point capabilities, and error-correcting code (ECC) memory.

Q: What APIs and programming models are supported by GPU Graphics and Compute instances?

P3 instances support CUDA 9 and OpenCL, P2 instances support CUDA 8 and OpenCL 1.2 and G3 instances support DirectX 12, OpenGL 4.5, CUDA 8, and OpenCL 1.2.

Q: Where do I get NVIDIA drivers for P3 and G3 instances?

There are two methods by which NVIDIA drivers may be obtained. There are listings on the AWS Marketplace that offer Amazon Linux AMIs and Windows Server AMIs with the NVIDIA drivers pre-installed. You may also launch 64-bit, HVM AMIs and install the drivers yourself. You must visit the NVIDIA driver website and search for the NVIDIA Tesla V100 for P3, NVIDIA Tesla K80 for P2, and NVIDIA Tesla M60 for G3 instances.

Q: Which AMIs can I use with P3, P2 and G3 instances?

You can currently use Windows Server, SUSE Enterprise Linux, Ubuntu, and Amazon Linux AMIs on P2 and G3 instances. P3 instances only support HVM AMIs.

Q: Does the use of G2 and G3 instances require third-party licenses?

Aside from the NVIDIA drivers and GRID SDK, the use of G2 and G3 instances does not necessarily require any third-party licenses. However, you are responsible for determining whether your content or technology used on G2 and G3 instances requires any additional licensing. For example, if you are streaming content you may need licenses for some or all of that content. If you are using third-party technology such as operating systems, audio and/or video encoders, and decoders from Microsoft, Thomson, Fraunhofer IIS, Sisvel S.p.A., MPEG-LA, and Coding Technologies, please consult these providers to determine if a license is required. For example, if you leverage the on-board h.264 video encoder on the NVIDIA GRID GPU you should reach out to MPEG-LA for guidance, and if you use mp3 technology you should contact Thomson for guidance.

Q: Why am I not getting NVIDIA GRID features on G3 instances using the driver downloaded from the NVIDIA website?

The NVIDIA Tesla M60 GPU used in G3 instances requires a special NVIDIA GRID driver to enable all advanced graphics features, and 4 monitors support with resolution up to 4096x2160. You need to use an AMI with NVIDIA GRID driver pre-installed, or download and install the NVIDIA GRID driver following the AWS documentation.

Q: Why am I unable to see the GPU when using Microsoft Remote Desktop?

When using Remote Desktop, GPUs using the WDDM driver model are replaced with a non-accelerated Remote Desktop display driver. In order to access your GPU hardware, you need to utilize a different remote access tool, such as VNC.

Q: What is Amazon EC2 F1?

Amazon EC2 F1 is a compute instance with programmable hardware you can use for application acceleration. The new F1 instance type provides a high performance, easy to access FPGA for developing and deploying custom hardware accelerations.

Q: What are FPGAs and why do I need them?

FPGAs are programmable integrated circuits that you can configure using software. By using FPGAs you can accelerate your applications up to 30x when compared with servers that use CPUs alone. And, FPGAs are reprogrammable, so you get the flexibility to update and optimize your hardware acceleration without having to redesign the hardware.

Q: How does F1 compare with traditional FPGA solutions?

F1 is an AWS instance with programmable hardware for application acceleration. With F1, you have access to FPGA hardware in a few simple clicks, reducing the time and cost of full-cycle FPGA development and scale deployment from months or years to days. While FPGA technology has been available for decades, adoption of application acceleration has struggled to be successful in both the development of accelerators and the business model of selling custom hardware for traditional enterprises, due to time and cost in development infrastructure, hardware design, and at-scale deployment. With this offering, customers avoid the undifferentiated heavy lifting associated with developing FPGAs in on-premises data centers.

Q: What is an Amazon FPGA Image (AFI)?

The design that you create to program your FPGA is called an Amazon FPGA Image (AFI). AWS provides a service to register, manage, copy, query, and delete AFIs. After an AFI is created, it can be loaded on a running F1 instance. You can load multiple AFIs to the same F1 instance, and can switch between AFIs in runtime without reboot. This lets you quickly test and run multiple hardware accelerations in rapid sequence. You can also offer to other customers on the AWS Marketplace a combination of your FPGA acceleration and an AMI with custom software or AFI drivers.

Q: How do I list my hardware acceleration on the AWS Marketplace?

You would develop your AFI and the software drivers/tools to use this AFI. You would then package these software tools/drivers into an Amazon Machine Image (AMI) in an encrypted format. AWS manages all AFIs in the encrypted format you provide to maintain the security of your code. To sell a product in the AWS Marketplace, you or your company must sign up to be an AWS Marketplace reseller, you would then submit your AMI ID and the AFI ID(s) intended to be packaged in a single product. AWS Marketplace will take care of cloning the AMI and AFI(s) to create a product, and associate a product code to these artifacts, such that any end-user subscribing to this product code would have access to this AMI and the AFI(s).

Q: What is available with F1 instances?

For developers, AWS is providing a Hardware Development Kit (HDK) to help accelerate development cycles, a FPGA Developer AMI for development in the cloud, an SDK for AMIs running the F1 instance, and a set of APIs to register, manage, copy, query, and delete AFIs. Both developers and customers have access to the AWS Marketplace where AFIs can be listed and purchased for use in application accelerations.

Q: Do I need to be an FPGA expert to use an F1 instance?

AWS customers subscribing to an F1-optimized AMI from AWS Marketplace do not need to know anything about FPGAs to take advantage of the accelerations provided by the F1 instance

and the AWS Marketplace. Simply subscribe to an F1-optimized AMI from the AWS Marketplace with an acceleration that matches the workload. The AMI contains all the software necessary for using the FPGA acceleration. Customers need only write software to the specific API for that accelerator and start using the accelerator.

Q: I'm a FPGA developer; how do I get started with F1 instances?

Developers can get started on the F1 instance by creating an AWS account and downloading the AWS Hardware Development Kit (HDK). The HDK includes documentation on F1, internal FPGA interfaces, and compiler scripts for generating AFI. Developers can start writing their FPGA code to the documented interfaces included in the HDK to create their acceleration function. Developers can launch AWS instances with the FPGA Developer AMI. This AMI includes the development tools needed to compile and simulate the FPGA code. The Developer AMI is best run on the latest C5, M5, or R4 instances. Developers should have experience in the programming languages used for creating FPGA code (i.e. Verilog or VHDL) and an understanding of the operation they wish to accelerate.

Q: I'm not an FPGA developer; how do I get started with F1 instances?

Customers can get started with F1 instances by selecting an accelerator from the AWS Marketplace, provided by AWS Marketplace sellers, and launching an F1 instance with that AMI. The AMI includes all of the software and APIs for that accelerator. AWS manages programming the FPGA with the AFI for that accelerator. Customers do not need any FPGA experience or knowledge to use these accelerators. They can work completely at the software API level for that accelerator.

Q: Does AWS provide a developer kit?

Yes. The Hardware Development Kit (HDK) includes simulation tools and simulation models for developers to simulate, debug, build, and register their acceleration code. The HDK includes code samples, compile scripts, debug interfaces, and many other tools you will need to develop the FPGA code for your F1 instances. You can use the HDK either in an AWS provided AMI, or in your on-premises development environment. These models and scripts are available publicly with an AWS account.

Q: Can I use the HDK in my on-premises development environment?

Yes. You can use the Hardware Development Kit HDK either in an AWS-provided AMI, or in your on-premises development environment.

Q: Can I add an FPGA to any EC2 instance type?

No. F1 instances comes in two instance sizes: f1.2xlarge, f1.4xlarge, and f1.16 xlarge.

Q: How do I use the Inferentia chip in Inf1 instances?

You can start your workflow by building and training your model in one of the popular ML frameworks such as TensorFlow, PyTorch, or MXNet using GPU instances such as P4, P3, or P3dn. Once the model is trained to your required accuracy, you can use the ML framework's API to invoke Neuron, a software development kit for Inferentia, to compile the model for execution on Inferentia chips, load it in to Inferentia's memory, and then execute inference calls. In order to get started quickly, you can use AWS Deep Learning AMIs that come pre-installed with ML frameworks and the Neuron SDK. For a fully managed experience you will be able to use Amazon SageMaker, which will enable you to seamlessly deploy your trained models on Inf1 instances.

Q: When would I use Inf1 vs. C6i or C5 vs. G4 instances for inference?

Customers running machine learning models that are sensitive to inference latency and throughput can use Inf1 instances for high-performance cost-effective inference. For those ML models that are less sensitive to inference latency and throughput, customers can use EC2 C6i or C5 instances and utilize the AVX-512/VNNI instruction set. For ML models that require access to NVIDIA's CUDA, CuDNN or TensorRT libraries, we recommend using G4 instances.

Model Characteristics and Libraries Used	**EC2 Inf1**	**EC2 C6i or C5**	**EC2 G4**
Models that benefit from low latency and high throughput at low cost	X		
Models not sensitive to latency and throughput		X	
Models requiring NVIDIA's developer libraries			X

Q: When should I choose Elastic Inference (EI) for inference vs Amazon EC2 Inf1 instances?

There are two cases where developers would choose EI over Inf1 instances: (1) if you need different CPU and memory sizes than what Inf1 offers, then you can use EI to attach acceleration to the EC2 instance with the right mix of CPU and memory for your application (2) if your performance requirements are significantly lower than what the smallest Inf1 instance provides, then using EI could be a more cost effective choice. For example, if you only need 5 TOPS, enough for processing up to 6 concurrent video streams, then using the smallest slice of EI with a C5.large instance could be up to 50% cheaper than using the smallest size of an Inf1 instance.

Q: What ML models types and operators are supported by EC2 Inf1 instances using the Inferentia chip?

Inferentia chips support the commonly used machine learning models such as single shot detector (SSD) and ResNet for image recognition/classification and Transformer and BERT for natural language processing and translation and many others. A list of supported operators can be found on GitHub.

Q: How do I take advantage of AWS Inferentia's NeuronCore Pipeline capability to lower latency?

Inf1 instances with multiple Inferentia chips, such as Inf1.6xlarge or Inf1.24xlarge, support a fast chip-to-chip interconnect. Using the Neuron Processing Pipeline capability, you can split your model and load it to local cache memory across multiple chips. The Neuron compiler uses ahead-of-time (AOT) compilation technique to analyze the input model and compile it to fit across the on-chip memory of single or multiple Inferentia chips. Doing so enables the Neuron Cores to have high-speed access to models and not require access to off-chip memory, keeping latency bounded while increasing the overall inference throughput.

Q: What is the difference between AWS Neuron and Amazon SageMaker Neo?

AWS Neuron is a specialized SDK for AWS Inferentia chips that optimizes the machine learning inference performance of Inferentia chips. It consists of a compiler, run-time, and profiling tools for AWS Inferentia and is required to run inference workloads on EC2 Inf1 instances. On the other hand, Amazon SageMaker Neo is a hardware agnostic service that consists of a compiler and run-time that enables developers to train machine learning models once, and run them on many different hardware platforms.

Q: How do I use the Trainium chips in Trn1 instances?

The Trainium software stack, AWS Neuron SDK, integrates with leading ML frameworks, such as PyTorch and TensorFlow, so you can get started with minimal code changes. To get started quickly, you can use AWS Deep Learning AMIs and AWS Deep Learning Containers, which come preconfigured with AWS Neuron. If you are using containerized applications, you can deploy AWS Neuron by using Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), or your preferred native container engine. AWS Neuron also supports Amazon SageMaker, which you can use to build, train, and deploy machine learning models.

Q: Where can I deploy deep learning models trained on Trn1?

You can deploy deep learning models trained on Trn1 instances on any other Amazon EC2 instance that supports deep learning use cases, including instances based on CPUs, GPUs, or other accelerators. You can also deploy models trained on Trn1 instances outside of AWS, such as on-premises data centers or in embedded devices at the edge. For example, you can train your models on Trn1 instances and deploy them on Inf1 instances, G5 instances, G4 instances, or compute devices at the edge.

Q: When would I use Trn1 instances over GPU-based instances for training ML models?

Trn1 instances are a good fit for your natural language processing (NLP), large language model (LLM), and computer vision (CV) model training use cases. Trn1 instances focus on accelerating model training to deliver high performance while also lowering your model training costs. If you have ML models that need third-party proprietary libraries or languages, for example NVIDIA CUDA, CUDA Deep Neural Network (cuDNN), or TensorRT libraries, we recommend using the NVIDIA GPU-based instances (P4, P3).

Burstable instances

Q: How are Burstable Performance Instances different?

Amazon EC2 allows you to choose between Fixed Performance Instances (e.g. C, M and R instance families) and Burstable Performance Instances (e.g. T2). Burstable Performance Instances provide a baseline level of CPU performance with the ability to burst above the baseline.

T2 instances' baseline performance and ability to burst are governed by CPU Credits. Each T2 instance receives CPU Credits continuously, the rate of which depends on the instance size. T2 instances accrue CPU Credits when they are idle, and consume CPU credits when they are active. A CPU Credit provides the performance of a full CPU core for one minute.

Model	vCPUs	CPU Credits / hour	Maximum CPU Credit Balance	Baseline CPU Performance
t2.nano	1	3	72	5% of a core
t2.micro	1	6	144	10% of a core
t2.small	1	12	288	20% of a core
t2.medium	2	24	576	40% of a core*
t2.large	2	36	864	60% of a core**
t2.xlarge	4	54	1,296	90% of a core***

t2.2xlarge	8	81	1,944	135% of a core****

** For the t2.medium, single threaded applications can use 40% of 1 core, or if needed, multithreaded applications can use 20% each of 2 cores.*

***For the t2.large, single threaded applications can use 60% of 1 core, or if needed, multithreaded applications can use 30% each of 2 cores.*

**** For the t2.xlarge, single threaded applications can use 90% of 1 core, or if needed, multithreaded applications can use 45% each of 2 cores or 22.5% of all 4 cores.*

***** For the t2.2xlarge, single threaded applications can use all of 1 core, or if needed, multithreaded applications can use 67.5% each of 2 cores or 16.875% of all 8 cores.*

Q. How do I choose the right Amazon Machine Image (AMI) for my T2 instances?

You will want to verify that the minimum memory requirements of your operating system and applications are within the memory allocated for each T2 instance size (e.g. 512 MiB for t2.nano). Operating systems with Graphical User Interfaces (GUI) that consume significant memory and CPU, for example Microsoft Windows, might need a t2.micro or larger instance size for many use cases. You can find AMIs suitable for the t2.nano instance types on AWS Marketplace. Windows customers who do not need the GUI can use the Microsoft Windows Server 2012 R2 Core AMI.

Q: When should I choose a Burstable Performance Instance, such as T2?

T2 instances provide a cost-effective platform for a broad range of general purpose production workloads. T2 Unlimited instances can sustain high CPU performance for as long as required. If your workloads consistently require CPU usage much higher than the baseline, consider a dedicated CPU instance family such as the M or C.

Q: How can I see the CPU Credit balance for each T2 instance?

You can see the CPU Credit balance for each T2 instance in EC2 per-Instance metrics in Amazon CloudWatch. T2 instances have four metrics, CPUCreditUsage, CPUCreditBalance, CPUSurplusCreditBalance and CPUSurplusCreditsCharged. CPUCreditUsage indicates the amount of CPU Credits used. CPUCreditBalance indicates the balance of CPU Credits. CPUSurplusCredit Balance indicates credits used for bursting in the absence of earned credits. CPUSurplusCreditsCharged indicates credits that are charged when average usage exceeds the baseline.

Q: What happens to CPU performance if my T2 instance is running low on credits (CPU Credit balance is near zero)?

If your T2 instance has a zero CPU Credit balance, performance will remain at baseline CPU performance. For example, the t2.micro provides baseline CPU performance of 10% of a physical CPU core. If your instance's CPU Credit balance is approaching zero, CPU performance will be lowered to baseline performance over a 15-minute interval.

Q: Does my T2 instance credit balance persist at stop / start?

No, a stopped instance does not retain its previously earned credit balance.

Q: Can T2 instances be purchased as Reserved Instances or Spot Instances?

T2 instances can be purchased as On-Demand Instances, Reserved Instances or Spot Instances.
Q: What are Amazon EC2 T4g instances?

Amazon EC2 T4g instances are the next-generation of general purpose burstable instances powered by Arm-based AWS Graviton2 processors. T4g instances deliver up to 40% better price performance over T3 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for T4g instances?

T4g instances deliver up to 40% better price performance over T3 instances for a wide variety of burstable general purpose workloads such as micro-services, low-latency interactive applications, small and medium databases, virtual desktops, development environments, code repositories, and business-critical applications. Customers deploying applications built on open source software across the T instance family will find the T4g instances an appealing option to realize the best price performance within the instance family. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.
Q: How can customers get access to the T4g free trial?

Until December 31, 2022, all AWS customers will be enrolled automatically in the T4g free trial as detailed in the AWS Free Tier. During the free-trial period, customers who run a t4g.small instance will automatically get 750 free hours per month deducted from their bill during each month. The 750 hours are calculated in aggregate across all Regions in which the t4g.small instances are used. Customers must pay for surplus CPU credits when they exceed the instances allocated credits during the 750 free hours of the T4g free trial program.

Q: Who is eligible for the T4g free trial?

All existing and new customers with an AWS account can take advantage of the T4g free trial. The T4g free trial is available for a limited time until December 31, 2022. The start and end time of the free trial are based on the Coordinated Universal Time (UTC). The T4g free trial will be available in addition to the existing AWS Free Tier on t2.micro/t3.micro. Customers who have exhausted their t2.micro (or t3.micro, depending on the Region) Free Tier usage can still benefit from the T4g free trial.

Q: What is the regional availability of T4g free trial?

The T4g free trial is currently available across these AWS Regions: US East (Ohio), US East (N. Virginia), US West (N. California), US West (Oregon), South America (Sao Paulo), Asia Pacific (Hong Kong), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), Europe (London), and Europe (Stockholm). It is currently not available in the China (Beijing) and China (Ningxia) Regions.

As part of the free trial, customers can run t4g.small instances across one or multiple Regions from a single cumulative bucket of 750 free hours per month until December 31, 2022. For example, a customer can run t4g.small in Oregon for 300 hours for a month and run another t4g.small in Tokyo for 450 hours during the same month. This would add up to 750 hours per month of the free-trial limit.

Q: Is there an additional charge for running specific AMIs under the T4g free trial?

Under the t4g.small free trial, there will be no Amazon Machine Image (AMI) charge for Amazon Linux 2, RHEL and SUSE Linux AMIs that are available through the EC2 console Quick Start for

the first 750 free hours per month. After 750 free hours per month, regular On-Demand prices, including AMI charge (if any), will apply. The applicable software fees for AWS Marketplace offers with AMI fulfillment options is not included in the free trial. Only the t4g.small infrastructure cost is included and covered under the free trial.

Q: How will the t4g.small free trial be reflected on my AWS bill?

The T4g free trial has a monthly billing cycle that starts on the first of every month and ends on the last day of that month. Under the T4g free-trial billing plan, customers using t4g.small will see a $0 line item on their bill under the On-Demand pricing plan for the first 750 aggregate hours of usage for every month during the free-trial period. Customers can start any time during the free-trial period and get 750 free hours for the remainder of that month. Any unused hours from the previous month will not be carried over. Customers can launch multiple t4g.small instances under the free trial. Customers will be notified automatically through email using AWS Budgets when their aggregate monthly usage reaches 85% of 750 free hours. When the aggregate instance usage exceeds 750 hours for the monthly billing cycle, customers will be charged based on regular On-Demand pricing for the exceeded hours for that month. For customers with a Compute Savings Plan or T4g Instance Savings Plan, Savings Plan (SV) discount will be applied to On-Demand pricing for hours exceeding the 750 free trial hours. If customers have purchased the T4g Reserved Instance (RI) plan, the RI plan applies first to any usage on an hourly basis. For any remaining usage after the RI plan has been applied, the free trial billing plan is in effect.

Q: If customers sign up for consolidated billing (or a single payer account), can they get the T4g free trial for each account that is tied to the payer account?

No, customers who use consolidated billing to consolidate payment across multiple accounts will have access to one free trial per Organization. Each payer account gets a total aggregate of 750 free hours a month.

Q: Will customers get charged for surplus CPU credits as a part of T4g free trial?

Customers must pay for surplus CPU credits when they exceed the instances allocated credits during the 750 free hours of the T4g free trial program. For details about how CPU credits work, see Key concepts and definitions for burstable performance instances in the Amazon EC2 User Guide for Linux Instances.

Q: At the end of the free trial, how will customers be billed for t4g.small instances?

Starting January 1, 2023, customers running on t4g.small instances will be automatically switched from the free trial plan to the On-Demand pricing plan (or Reserved Instance (RI)/Savings Plan (SV) plan, if purchased). Accumulated credits will be set to zero. Customers will receive an email notification seven days before the end of the free trial period stating that the free trial period will be ending in seven days. Starting January 1, 2023, if the RI plan is purchased, the RI plans will apply. Otherwise, customers will be charged regular On-Demand pricing for t4g.small instances. For customers who have the T4g Instance Savings Plan or a Compute Savings Plan, t4g.small instance billing will apply the Savings Plan discount on their On-Demand pricing.

Compute Optimized instances

Q: When should I use Compute Optimized instances?

Compute Optimized instances are designed for applications that benefit from high compute power. These applications include compute-intensive applications like high-performance web servers, high-performance computing (HPC), scientific modelling, distributed analytics and machine learning inference.

Q: What are Amazon EC2 C7g instances?

Amazon EC2 C7g instances, powered by the latest generation AWS Graviton3 processors, provide the best price performance in Amazon EC2 for compute-intensive workloads. C7g instances are ideal for high performance computing (HPC), batch processing, electronic design automation (EDA), gaming, video encoding, scientific modeling, distributed analytics, CPU-based machine learning (ML) inference, and ad-serving. They offer up to 25% better performance over the sixth generation AWS Graviton2-based C6g instances.
Q: What are Amazon EC2 C6g instances?

Amazon EC2 C6g instances are the next-generation of compute-optimized instances powered by Arm-based AWS Graviton2 Processors. C6g instances deliver up to 40% better price performance over C5 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for C6g instances?

C6g instances deliver significant price performance benefits for compute-intensive workloads such as high performance computing (HPC), batch processing, ad serving, video encoding, gaming, scientific modelling, distributed analytics, and CPU-based machine learning inference. Customers deploying applications built on open source software across the C instance family will find the C6g instances an appealing option to realize the best price performance within the instance family. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: What are the various storage options available on C6g instances?

C6g instances are EBS-optimized by default and offer up to 19,000 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. C6g instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. Additionally, options with local NVMe instance storage are also available through the C6gd instance types.

Q: Which network interface is supported on C6g instances?

C6g instances support ENA based Enhanced Networking. With ENA, C6g instances can deliver up to 25 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Will customers need to modify their applications and workloads to be able to run on the C6g instances?

The changes required are dependent on the application. Customers running applications built on open source software will find that the Arm ecosystem is well developed and already likely supports their applications. Most Linux distributions as well as containers (Docker, Kubernetes, Amazon ECS, Amazon EKS, Amazon ECR) support the Arm architecture. Customers will find Arm versions of commonly used software packages available for installation through the same mechanisms that they currently use. Applications that are based on interpreted languages (such as Java, Node, Python) not reliant on native CPU instruction sets should run with minimal to no changes. Applications developed using compiled languages (C, C++, GoLang) will need to be re-compiled to generate Arm binaries. The Arm architecture is well supported in these popular programming languages and modern code usually requires a simple ‘Make’ command.

Q: Will there be more compute choices offered with the C6 instance families?

Yes, we plan to offer Intel and AMD CPU powered instances in the future as part of the C6 instance families.
Q. Can I launch C4 instances as Amazon EBS-optimized instances?

Each C4 instance type is EBS-optimized by default. C4 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance. Since this feature is always enabled on C4 instances, launching a C4 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q. How can I use the processor state control feature available on the c4.8xlarge instance?

The c4.8xlarge instance type provides the ability for an operating system to control processor C-states and P-states. This feature is currently available only on Linux instances. You may want to change C-state or P-state settings to increase processor performance consistency, reduce latency, or tune your instance for a specific workload. By default, Amazon Linux provides the highest-performance configuration that is optimal for most customer workloads; however, if your application would benefit from lower latency at the cost of higher single- or dual-core frequencies, or from lower-frequency sustained performance as opposed to bursty Turbo Boost frequencies, then you should consider experimenting with the C-state or P-state configuration options that are available to these instances.

Q. Which instances are available within Compute Optimized instances category?

C6g instances: Amazon EC2 C6g instances are powered by Arm-based AWS Graviton2 processors. They deliver up to 40% better price performance over C5 instances and are ideal for running advanced compute-intensive workloads. This includes workloads such as high performance computing (HPC), batch processing, ad serving, video encoding, gaming, scientific modelling, distributed analytics, and CPU-based machine learning inference.

Hpc6a instances: Hpc6a instances are powered by 96 cores of 3rd generation AMD EPYC processors with an all-core turbo frequency of 3.6 GHz, and 384 GB RAM. Hpc6a instances offer 100 Gbps Elastic Fabric Adapter networking enabled by default for high throughput inter-node communications to help you run your HPC workloads at scale. Hpc6a instances deliver up to 65% better price performance over comparable, compute-optimized, x-86 based instances.

C6a instances: C6a instances are powered by 3rd generation AMD EPYC processors with an all-core turbo frequency of 3.6 GHz, offer up to 15% better price performance over C5a instances for a wide variety of workloads, and support always-on memory encryption using AMD Transparent Single Key Memory Encryption (TSME). C6a instances provide new instance sizes with up to 192 vCPUs and 384 GiB of memory, double that of the largest C5a instance. C6a also gives customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, more than twice that of C5a instances.

C6i instances: C6i instances are powered by 3rd generation Intel Xeon Scalable processors with an all-core turbo frequency of 3.5 GHz, offer up to 15% better price performance over C5 instances for a wide variety of workloads, and always-on memory encryption using Intel Total Memory encryption (TME). C6i instances provide a new instance size (c6i.32xlarge) with 128 vCPUs and 256 GiB of memory, 33% more than the largest C5 instance. They also provide up to 9% higher memory bandwidth per vCPU compared to C5 instances. C6i also give customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of C5 instances. C6i are also available with local NVMe-based SSD block-level storage (C6id instances) for applications that need high-speed, low-latency local storage. Compared to previous generation C5d instances, C6id instances offer up to 138% higher TB storage per vCPU and 56% lower cost per TB.

C5 instances: C5 instances are based on Intel Xeon Platinum processors, part of the Intel Xeon Scalable (codenamed Skylake-SP or Cascade Lake) processor family, are available in 9 sizes, and offer up to 96 vCPUs and 192 GiB memory. C5 instances deliver 25% improvement in price/performance compared to C4 instances. The C5d instances have local NVMe storage for workloads that require very low latency and storage access with high random read and write IOPS ability.

C5a instances: C5a instances deliver leading x86 price-performance for a broad set of compute-intensive workloads including batch processing, distributed analytics, data transformations, log analysis, and web applications. C5a instances feature 2nd Gen 3.3GHz AMD EPYC processors with up to 96 vCPUs and up to 192 GiB of memory. The C5ad instances have local NVMe storage for workloads that require very low latency and storage access with high random read and write IOPS ability.

C5n instances: C5n instances are ideal for applications requiring high network bandwidth and packet rate. The C5n instances are ideal for applications like HPC, data lakes, network appliances as well as applications that require inter-node communication and the Message Passing Interface (MPI). C5n offer a choice of Intel Xeon Platinum 3.0 GHz processors with up to 72 vCPUs and 192GiB of Memory.

C4 instances: C4 instances are based on Intel Xeon E5-2666 v3 (codenamed Haswell) processors. C4 instances are available in 5 sizes and offer up to 36 vCPUs and 60 GiB memory.

Q. Why should customers choose C6i instances over C5 instances?

C6i instances offer up to 15% better price performance over C5 instances, and always-on memory encryption using Intel Total Memory encryption (TME). C6i instances provide a new instance size (c6i.32xlarge) with 128 vCPUs and 256 GiB of memory, 33% more than the largest C5 instance. They also provide up to 9% higher memory bandwidth per vCPU compared to C5 instances. C6i also give customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of C5 instances.

Q. Why should customers choose C5 instances over C4 instances?

The generational improvement in CPU performance and lower price of C5 instances, which combined result in a 25% price/performance improvement relative to C4 instances, benefit a broad spectrum of workloads that currently run on C3 or C4 instances. For floating point intensive applications, Intel AVX-512 enables significant improvements in delivered TFLOPS by effectively extracting data level parallelism. Customers looking for absolute performance for graphics rendering and HPC workloads that can be accelerated with GPUs or FPGAs should also evaluate other instance families in the Amazon EC2 portfolio that include those resources to find the ideal instance for their workload.

Q. Which storage interface is supported on C5 instances?

C5 instances will support only NVMe EBS device model. EBS volumes attached to C5 instances will appear as NVMe devices. NVMe is a modern storage interface that provides latency reduction and results in increased disk I/O and throughput.

Q: Why does the total memory reported by the operating system not exactly match the advertised memory on instance types?

Portions of the EC2 instance memory are reserved and used by the virtual BIOS for video RAM, DMI, and ACPI. In addition, for instances that are powered by the AWS Nitro Hypervisor, a small percentage of the instance memory is reserved by the Amazon EC2 Nitro Hypervisor to manage virtualization.

Q: How are Hpc6a instances different from other EC2 instances?

Hpc6a instances offer optimal price performance for tightly coupled high performance computing (HPC) workloads. These instances deliver 100 Gbps Elastic Fabric Adapter network bandwidth optimized for traffic between instances in the same placement group, and have simultaneous multi threading disabled. These instances are available in a single Availability Zone to optimize for tightly-coupled workloads and have standard networking bandwidth.

Q: When should I choose between the Hpc6a and other network-optimized instances?

If you are running HPC applications that rely on inter-instance communications, you can use the Hpc6a instances to realize the best price performance in Amazon EC2. If you are running non-HPC workloads like network appliances and Data Lakes that require high availability, high network bandwidth to outside of the VPC, S3, or EBS for data volumes should continue to use their current high network instances.

Q: What is the regional availability of the Hpc6a instances?

The Hpc6a instances are available in US East (Ohio) and GovCloud (West). To optimize networking for tightly coupled workloads, you will be able to access Hpc6a instances in a single Availability Zone in each Region where available.

Q: Which AMIs are supported on Hpc6a instances?

Hpc6a instances support Amazon Linux 2, Amazon Linux, Ubuntu 18.04 or later, Red Hat Enterprise Linux 7.4 or later, SUSE Linux Enterprise Server 12 SP2 or later, CentOS 7 or later, and FreeBSD 11.1 or later. These instances also support Windows including Windows Server 2012, 2012 R2, 2016, and 2019.

Q: What pricing models does Hpc6a support?

Hpc6a instances will be available for purchase via 1 year and 3 year standard and convertible RIs, Savings Plans, and On-Demand Instances.

General Purpose instances

Q: What are Amazon EC2 M6g instances?

Amazon EC2 M6g instances are the next-generation of general-purpose instances powered by Arm-based AWS Graviton2 Processors. M6g instances deliver up to 40% better price/performance over M5 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are the specifications of the new AWS Graviton2 Processors?

The AWS Graviton2 processors deliver up to 7x performance, 4x the number of compute cores, 2x larger caches, 5x faster memory, and 50% faster per core encryption performance than first generation AWS Graviton processors. Each core of the AWS Graviton2 processor is a single-threaded vCPU. These processors also offer always-on fully encrypted DRAM memory, hardware acceleration for compression workloads, dedicated engines per vCPU that double the floating-point performance for workloads such as video encoding, and instructions for int8/fp16 CPU-based machine learning inference acceleration. The CPUs are built utilizing 64-bit Arm Neoverse cores and custom silicon designed by AWS on the advanced 7 nm manufacturing technology.

Q: Is memory encryption supported by AWS Graviton2 processors?

AWS Graviton2 processors support always-on 256-bit memory encryption to further enhance security. Encryption keys are securely generated within the host system, do not leave the host system, and are irrecoverably destroyed when the host is rebooted or powered down. Memory encryption does not support integration with AWS Key Management Service (KMS) and customers cannot bring their own keys.

Q: What are some of the ideal use cases for M6g instances?

M6g instances deliver significant performance and price performance benefits for a broad spectrum of general-purpose workloads such as application servers, gaming servers, microservices, mid-size databases, and caching fleets. Customers deploying applications built on open source software across the M instance family will find the M6g instances an appealing option to realize the best price-performance within the instance family. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: What are the various storage options available on M6g instances?

M6g instances are EBS-optimized by default and offer up to 19,000 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. M6g instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. Additionally, options with local NVMe instance storage are also available through the M6gd instance types.

Q: Which network interface is supported on M6g instances?

M6g instances support ENA based Enhanced Networking. With ENA, M6g instances can deliver up to 25 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Will customers need to modify their applications and workloads to be able to run on the M6g instances?

The changes required are dependent on the application. Customers running applications built on open source software will find that the Arm ecosystem is well developed and already likely supports their applications. Most Linux distributions as well as containers (Docker, Kubernetes, Amazon ECS, Amazon EKS, Amazon ECR) support the Arm architecture. Customers will find Arm versions of commonly used software packages available for installation through the same mechanisms that they currently use. Applications that are based on interpreted languages (such as Java, Node, Python) not reliant on native CPU instruction sets should run with minimal to no changes. Applications developed using compiled languages (C, C++, GoLang) will need to be re-compiled to generate Arm binaries. The Arm architecture is well supported in these popular programming languages and modern code usually requires a simple ‘Make’ command.

Q: What are Amazon EC2 A1 instances?

Amazon EC2 A1 instances are general purpose instances powered by the first-generation AWS Graviton Processors that are custom designed by AWS.

Q: What are the specifications of the first-generation AWS Graviton Processors?

AWS Graviton processors are custom designed by AWS utilizing Amazon’s extensive expertise in building platform solutions for cloud applications running at scale. These processors are based on the 64-bit Arm instruction set and feature Arm Neoverse cores as well as custom silicon designed by AWS. The cores operate at a frequency of 2.3 GHz.

Q: When should I use A1 instances?

A1 instances deliver significant cost savings for scale-out workloads that can fit within the available memory footprint. A1 instances are ideal for scale-out applications such as web servers, containerized microservices, and data/log processing. These instances will also appeal to developers, enthusiasts, and educators across the Arm developer community.

Q: Will customers have to modify applications and workloads to be able to run on the A1 instances?

The changes required are dependent on the application. Applications based on interpreted or run-time compiled languages (e.g. Python, Java, PHP, Node.js) should run without modifications. Other applications may need to be recompiled and those that don't rely on x86 instructions will generally build with minimal to no changes.

Q: Which operating systems/AMIs are supported on A1 Instances?

The following AMIs are supported on A1 instances: Amazon Linux 2, Ubuntu 16.04.4 or newer, Red Hat Enterprise Linux (RHEL) 7.6 or newer, SUSE Linux Enterprise Server 15 or newer. Additional AMI support for Fedora, Debian, NGINX Plus are also available through community AMIs and the AWS Marketplace. EBS backed HVM AMIs launched on A1 instances require NVMe and ENA drivers installed at instance launch.

Q: Are there specific AMI requirements to run on M6g and A1 instances?

You will need to use the “arm64” AMIs with the M6g and A1 instances. x86 AMIs are not compatible with M6g and A1 instances.

Q: When should customers use A1 instances versus the new M6g instances?

A1 instances continue to offer significant cost benefits for scale-out workloads that can run on multiple smaller cores and fit within the available memory footprint. The new M6g instances are a good fit for a broad spectrum of applications that require more compute, memory, networking resources and/or can benefit from scaling up across platform capabilities. M6g instances will deliver the best price-performance within the instance family for these applications. M6g supports up to 16xlarge instance size (A1 supports up to 4xlarge), 4GB of memory per vCPU (A1 supports 2GB memory per vCPU), and up to 25 Gbps of networking bandwidth (A1 supports up to 10 Gbps).

Q: What are the various storage options available to A1 customers?

A1 instances are EBS-optimized by default and offer up to 3,500 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. A1 instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. A1 instances will not support the blkfront interface.

Q: Which network interface is supported on A1 instances?

A1 instances support ENA based Enhanced Networking. With ENA, A1 instances can deliver up to 10 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Do A1 instances support the AWS Nitro System?

Yes, A1 instances are powered by the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: Why should customers choose EC2 M5 Instances over EC2 M4 Instances?

Compared with EC2 M4 Instances, the new EC2 M5 Instances deliver customers greater compute and storage performance, larger instance sizes for less cost, consistency and security. The biggest benefit of EC2 M5 Instances is based on its usage of the latest generation of Intel Xeon Scalable processors (Skylake-SP or Cascade Lake), which deliver up to 20% improvement in price/performance compared to M4. With AVX-512 support in M5 vs. the older AVX2 in M4, customers will gain 2x higher performance in workloads requiring floating point operations. M5 instances offer up to 25 Gbps of network bandwidth and up to 10 Gbps of dedicated bandwidth to Amazon EBS. M5 instances also feature significantly higher networking and Amazon EBS performance on smaller instance sizes with EBS burst capability.

Q: Why should customers choose M6i instances over M5 instances?

Amazon M6i instances are powered by 3rd generation Intel Xeon Scalable processors (code named Ice Lake) with an all-core turbo frequency of 3.5 GHz, offer up to 15% better compute price performance over M5 instances, and always-on memory encryption using Intel Total Memory Encryption (TME). Amazon EC2 M6i instances are the first to use a lower-case “i” to indicate they are Intel-powered instances. M6i instances provide a new instance size (m6i.32xlarge) with 128 vCPUs and 512 GiB of memory, 33% more than the largest M5 instance. They also provide up to 20% higher memory bandwidth per vCPU compared to M5 instances, allowing customers to efficiently perform real-time analysis for data-intensive AI/ML, gaming, and High Performance Computing (HPC) applications. M6i also give customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of M5 instances. M6i also allows customers to use Elastic Fabric Adapter on the 32xlarge size, enabling low latency and high scale inter-node communication. For optimal networking performance on these new instances, Elastic Network Adapter (ENA) driver update may be required.

Q: How does support for Intel AVX-512 benefit customers who use the EC2 M5 family or the M6i family?

Intel Advanced Vector Extensions 512 (AVX-512) is a set of new CPU instructions available on the latest Intel Xeon Scalable processor family, that can accelerate performance for workloads and usages such as scientific simulations, financial analytics, artificial intelligence, machine learning/deep learning, 3D modeling and analysis, image and video processing, cryptography and data compression, among others. Intel AVX-512 offers exceptional processing of encryption algorithms, helping to reduce the performance overhead for cryptography, which means customers who use the EC2 M5 family or M6i family can deploy more secure data and services into distributed environments without compromising performance.

Q: What are M5zn instances?

M5zn instances are a variant of the M5 general purpose instances that are powered by the fastest Intel Xeon Scalable processor in the cloud, with an all-core turbo frequency of up to 4.5 GHz, along with 100 Gbps networking and support for Amazon EFA. M5zn instances are an ideal fit for workloads such as gaming, financial applications, simulation modeling applications such as those used in the automotive, aerospace, energy, and telecommunication industries, and other High Performance Computing applications.

Q: How are M5zn instances different than z1d instances?

z1d instances are a memory-optimized instance, and feature a high frequency version of the Intel Xeon Scalable processors (up to 4.0 GHz), along with local NVMe storage. M5zn instances are a general purpose instance, and feature a high frequency version of the 2nd Generation Intel Xeon Scalable processors up to 4.5 GHz), along with up to 100 Gbps networking performance, and support for EFA. M5zn instances offer improved price performance compared to z1d.

High Memory instances

Q. What are EC2 High Memory instances?

Amazon EC2 High Memory instances offer 3, 6, 9, 12, 18, or 24 TiB of memory in a single instance. These instances are designed to run large in-memory databases, including production installations of SAP HANA, in the cloud.

EC2 High Memory instances with 3, 6, 9, and 12 TiB of memory are powered by an 8-socket platform with Intel® Xeon® Platinum 8176M (Skylake) processors. EC2 High Memory instances with 18 and 24 TiB of memory are the first Amazon EC2 instances powered by an 8-socket platform with 2nd Generation Intel® Xeon® Scalable (Cascade Lake) processors that are optimized for mission-critical enterprise workloads. EC2 High Memory instances deliver high networking throughput and low-latency with up to 100 Gbps of aggregate network bandwidth using Amazon Elastic Network Adapter (ENA)-based Enhanced Networking. EC2 High Memory instances are EBS-Optimized by default, and support encrypted and unencrypted EBS volumes.

Q. Are High Memory instances certified by SAP to run SAP HANA workloads?

High Memory instances are certified by SAP for running Business Suite on HANA, the next-generation Business Suite S/4HANA, Data Mart Solutions on HANA, Business Warehouse on HANA, and SAP BW/4HANA in production environments. For details, see SAP's Certified and Supported SAP HANA Hardware Directory.

Q. Are High Memory instances only available as bare metal?

High Memory instances are available as both bare metal and virtualized instances, giving customers the choice to have direct access to the underlying hardware resources, or to take advantage of the additional flexibility that virtualized instances offer including On-Demand and 1-year and 3-year Savings Plan purchase options.

Q. What are the storage options available with High Memory instances?

High Memory instances support Amazon EBS volumes for storage. High Memory instances are EBS-optimized by default, and offer up to 38 Gbps of storage bandwidth.

Q. Which storage interface is supported on High Memory instances?

High Memory instances access EBS volumes via PCI attached NVM Express (NVMe) interfaces. EBS volumes attached to High Memory instances appear as NVMe devices. NVMe is an efficient and scalable storage interface, which is commonly used for flash based SSDs and provides latency reduction and results in increased disk I/O and throughput. The EBS volumes are attached and detached by PCI hotplug.

Q. What network performance is supported on High Memory instances?

High Memory instances use the Elastic Network Adapter (ENA) for networking and enable Enhanced Networking by default. With ENA, High Memory instances can utilize up to 100 Gbps of network bandwidth.

Q. Can I run High Memory instances in my existing Amazon Virtual Private Cloud (VPC)?

You can run High Memory instances in your existing and new Amazon VPCs.

Q. What is the underlying hypervisor on High Memory instances?

High Memory instances use the lightweight Nitro Hypervisor that is based on core KVM technology.

Q. Do High Memory instances enable CPU power management state control?

Yes. You can configure C-states and P-states on High Memory instances. You can use C-states to enable higher turbo frequencies (as much as 4.0 GHz). You can also use P-states to lower performance variability by pinning all cores at P1 or higher P states, which is similar to disabling Turbo, and running consistently at the base CPU clock speed.

Q. What purchase options are available for High Memory instances?

EC2 High Memory bare metal instances (e.g. u-6tb1.metal) are only available as EC2 Dedicated Hosts on a 1-Yr and 3-Yr reservations. EC2 High Memory virtualized instances (e.g. u-6tb1.112xlarge) are available for purchase via 1-Yr and 3-Yr Savings Plan, On-Demand instances, and as Dedicated hosts.

Q. What is the lifecycle of a Dedicated Host?

Once a Dedicated Host is allocated within your account, it will be standing by for your use. You can then launch an instance with a tenancy of "host" using the RunInstances API, and can also stop/start/terminate the instance through the API. You can use the AWS Management Console to manage the Dedicated Host and the instance.

Q. Can I launch, stop/start, and terminate High Memory instances using AWS CLI/SDK?

You can launch, stop/start, and terminate instances using AWS CLI/SDK.

Q. Which AMIs are supported with High memory instances?

EBS-backed HVM AMIs with support for ENA networking can be used with High Memory instances. The latest Amazon Linux, Red Hat Enterprise Linux, SUSE Enterprise Linux Server, and Windows Server AMIs are supported. Operating system support for SAP HANA workloads on High Memory instances include: SUSE Linux Enterprise Server 12 SP3 for SAP, Red Hat Enterprise Linux 7.4 for SAP, Red Hat Enterprise Linux 7.5 for SAP, SUSE Linux Enterprise Server 12 SP4 for SAP, SUSE Linux Enterprise Server 15 for SAP, Red Had Enterprise Linux 7.6 for SAP. Refer to SAP's Certified and Supported SAP HANA Hardware Directory for latest detail on supported operating systems.

Q. Are there standard SAP HANA reference deployment frameworks available for the High Memory instance and the AWS Cloud?

You can use the AWS Quick Start reference SAP HANA deployments to rapidly deploy all the necessary SAP HANA building blocks on High Memory instances following SAP's recommendations for high performance and reliability. AWS Quick Starts are modular and customizable, so you can layer additional functionality on top or modify them for your own implementations.

Memory Optimized instances

Q. When should I use memory-optimized instances?

Memory-optimized instances offer large memory size for memory intensive applications including in-memory applications, in-memory databases, in-memory analytics solutions, High Performance Computing (HPC), scientific computing, and other memory-intensive applications.

Q: What are Amazon EC2 R6g instances?

Amazon EC2 R6g instances are the next-generation of memory-optimized instances powered by Arm-based AWS Graviton2 Processors. R6g instances deliver up to 40% better price performance over R5 instances. They are built on the AWS Nitro System, a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for R6g instances?

R6g instances deliver significant price performance benefits for memory-intensive workloads such as instances and are ideal for running memory-intensive workloads such as open-source databases, in-memory caches, and real time big data analytics. Customers deploying applications built on open source software across the R instance family will find the R6g instances an appealing option to realize the best price performance within the instance family. Arm developers can also build their applications directly on native Arm hardware as opposed to cross-compilation or emulation.

Q: What are the various storage options available on R6g instances?

R6g instances are EBS-optimized by default and offer up to 19,000 Mbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes. R6g instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. Additionally, options with local NVMe instance storage are also available through the R6gd instance types.

Q: Which network interface is supported on R6g instances?

R6g instances support ENA based Enhanced Networking. With ENA, R6g instances can deliver up to 25 Gbps of network bandwidth between instances when launched within a Placement Group.

Q: Will customers need to modify their applications and workloads to be able to run on the R6g instances?

The changes required are dependent on the application. Customers running applications built on open source software will find that the Arm ecosystem is well developed and already likely supports their applications. Most Linux distributions as well as containers (Docker, Kubernetes, Amazon ECS, Amazon EKS, Amazon ECR) support the Arm architecture. Customers will find Arm versions of commonly used software packages available for installation through the same mechanisms that they currently use. Applications that are based on interpreted languages (such as Java, Node, Python) not reliant on native CPU instruction sets should run with minimal to no changes. Applications developed using compiled languages (C, C++, GoLang) will need to be re-compiled to generate Arm binaries. The Arm architecture is well supported in these popular programming languages and modern code usually requires a simple ‘Make’ command.

Q: Why should customers choose R6i instances over R5 instances?

Amazon R6i instances are powered by 3rd generation Intel Xeon Scalable processors (code named Ice Lake) with an all-core turbo frequency of 3.5 GHz, offer up to 15% better compute price performance over R5 instances, and always-on memory encryption using Intel Total Memory Encryption (TME). Amazon EC2 R6i instances use a lower-case “i” to indicate they are Intel-powered instances. R6i instances provide a new instance size (r6i.32xlarge) with 128 vCPUs and 1,024 GiB of memory, 33% more than the largest R5 instance. They also provide up to 20% higher memory bandwidth per vCPU compared to R5 instances, allowing customers to efficiently perform real-time analysis for data-intensive AI/ML, gaming, and High Performance Computing (HPC) applications. R6i also give customers up to 50 Gbps of networking speed and 40 Gbps of bandwidth to the Amazon Elastic Block Store, twice that of R5 instances. R6i also

allows customers to use Elastic Fabric Adapter on the 32xlarge and metal sizes, enabling low latency and high scale inter-node communication. For optimal networking performance on these new instances, Elastic Network Adapter (ENA) driver update may be required.

Q: What are Amazon EC2 R5b instances?

R5b instances are EBS-optimized variants of memory-optimized R5 instances that deliver up to 3x better EBS performance compared to same sized R5 instances. R5b instances deliver up to 60 Gbps bandwidth and 260K IOPS of EBS performance, the fastest block storage performance on EC2. They are built on the AWS Nitro System, which is a combination of dedicated hardware and Nitro hypervisor.

Q: What are some of the ideal use cases for R5b instances?

R5b instances are ideal for large relational database workloads, including Microsoft SQL Server, SAP HANA, IBM DB2, and Oracle that run performance intensive applications such as commerce platforms, ERP systems, and health record systems. Customers looking to migrate large on-premises workloads with large storage performance requirements to AWS will find R5b instances to be a good fit.

Q: What are the various storage options available on R5b instances?

R5b instances are EBS-optimized by default and offer up to 60,000 Mbps of dedicated EBS bandwidth and 260K IOPS for both encrypted and unencrypted EBS volumes. R5b instances only support Non-Volatile Memory Express (NVMe) interface to access EBS storage volumes. R5b is supported by all volume types, with the exception of io2 volumes.

Q: When should I use R5b instances?

Customers running workloads such as large relational databases and data analytics that want to take advantage of the increased EBS storage network performance can use R5b instances to deliver higher performance and bandwidth. Customers can also lower costs by migrating their workloads to smaller size R5b instances or by consolidating workloads on fewer R5b instances.

Q. What are the storage options available with High Memory instances?

High Memory instances support Amazon EBS volumes for storage. High Memory instances are EBS-optimized by default, and offer up to 38Gbps of storage bandwidth to both encrypted and unencrypted EBS volumes.
Q: What are Amazon EC2 X2gd instances?

Amazon EC2 X2gd instances are the next generation of memory-optimized instances powered by AWS-designed Arm-based AWS Graviton2 processors. X2gd instances deliver up to 55% better price performance compared to x86-based X1 instances and offer the lowest cost per GiB of memory in Amazon EC2. They are the first in the X family of instances to be built on the AWS Nitro System, which is a combination of dedicated hardware and Nitro hypervisor.

Q: What workloads are suited for X2gd instances?

X2gd is ideal for customers with Arm-compatible memory bound scale-out workloads such as Redis and Memcached in-memory databases, that need low latency memory access and benefit from more memory per vCPU. X2gd is also well suited for relational databases such as PostgreSQL, MariaDB, MySQL, and RDS Aurora. Customers who run memory intensive workloads such as Apache Hadoop, real-time analytics, and real-time caching servers will benefit from 1:16 vCPU to memory ratio of X2gd. Single threaded workloads such as EDA backend verification jobs will benefit from physical core and more memory of X2gd instances,

allowing them to consolidate more workloads on to a single instance. X2gd instance also feature local NVMe SSD block storage to improve response times by acting as a caching layer.

Q: When should I use X2gd instances compared to the X1, X2i, or R-family of instances?

X2gd instances are suitable for Arm-compatible memory bound scale-out workloads such as in-memory databases, memory analytics applications, open-source relational database workloads, EDA workloads, and large caching servers. X2gd instances offer customers the lowest cost per gigabyte of memory within EC2, with sizes up to 1 TiB. X2iezn, X2idn, X2iedn, X1, and X1e instances use x86 processors and are suitable for memory-intensive enterprise-class, scale-up workloads such as Windows workloads, in-memory databases (e.g. SAP HANA), and relational databases (e.g. OracleDB). Customers can leverage the x86-based X-family of instances for larger memory sizes up to 4 TiB. R6g and R6gd instances are suitable for workloads such as web applications, databases, and search indexing queries that need more vCPUs during times of heavy data processing. Customers running memory bound workloads that need less than 1 TiB memory and have dependency on x86 instruction set such as Windows applications, and applications like Oracle or SAP can leverage the R5 and R6 family of instances.

Q: When should I use X2idn and X2iedn instances?

X2idn and X2iedn instances are powered by 3rd generation Intel Xeon Scalable processors with an all-core turbo frequency up to 3.5 GHz and deliver up to 50% higher compute price performance than comparable X1 instances. X2idn and X2iedn instances both include up to 3.8 TB of local NVMe SSD storage and up to 100 Gbps of networking bandwidth, while X2idn offers up to 2 TiB of memory and X2iedn offers up to 4 TiB of memory. X2idn and X2iedn instances are SAP-Certified and are a great fit for workloads such as small-to large-scale traditional and in-memory databases, and analytics.

Q: When should I use X2iezn instances?

X2iezn instances feature the fastest Intel Xeon Scalable processors in the cloud and are a great fit for workloads that need high single-threaded performance combined with a high memory-to-vCPU ratio and high speed networking. X2iezn instances have an all-core turbo frequency up to 4.5 GHz, feature a 32:1 ratio of memory to vCPU, and deliver up to 55% higher compute price performance compared to X1e instances. X2iezn instances are a great fit for electronic design automation (EDA) workloads like physical verification, static timing analysis, power signoff, and full chip gate-level simulation.

Q: Which operating systems/AMIs are supported on X2gd instances?

The following AMIs are supported: Amazon Linux 2, Ubuntu 18.04 or newer, Red Hat Enterprise Linux 8.2 or newer, and SUSE Enterprise Server 15 or newer. Customers will find additional AMIs such as Fedora, Debian, NetBSD, and CentOS available through community AMIs and the AWS Marketplace. For containerized applications, Amazon ECS and EKS optimized AMIs are available as well.

Q. When should I use X1 instances?

X1 instances are ideal for running in-memory databases like SAP HANA, big data processing engines like Apache Spark or Presto, and high performance computing (HPC) applications. X1 instances are certified by SAP to run production environments of the next-generation Business Suite S/4HANA, Business Suite on HANA (SoH), Business Warehouse on HANA (BW), and Data Mart Solutions on HANA on the AWS cloud.

Q. Do X1 and X1e instances enable CPU power management state control?

Yes. You can configure C-states and P-states on x1e.32xlarge, x1e.16xlarge, x1e.8xlarge, x1.32xlarge and x1.16xlarge instances. You can use C-states to enable higher turbo frequencies (as much as 3.1 GHz with one or two core turbo). You can also use P-states to lower performance variability by pinning all cores at P1 or higher P states, which is similar to disabling Turbo, and running consistently at the base CPU clock speed.

x1e.32xlarge will also support Windows Server 2012 R2 and 2012 RTM. x1e.xlarge, x1e.2xlarge, x1e.4xlarge, x1e.8xlarge, x1e.16xlarge and x1.32xlarge will also support Windows Server 2012 R2, 2012 RTM and 2008 R2 64bit (Windows Server 2008 SP2 and older versions will not be supported) and x1.16xlarge will support Windows Server 2012 R2, 2012 RTM, 2008 R2 64bit, 2008 SP2 64bit, and 2003 R2 64bit (Windows Server 32bit versions will not be supported).

Q. Are there standard SAP HANA reference deployment frameworks available for the High Memory instance and the AWS?

You can use AWS Launch Wizard for SAP or AWS Quick Start reference SAP HANA deployments to rapidly deploy all the necessary SAP HANA building blocks on High Memory instances following recommendations from AWS and SAP for high performance and reliability.

Previous Generation instances

Q: Why don't I see M1, C1, CC2 and HS1 instances on the pricing pages any more?

These have been moved to the Previous Generation Instance page.

Q: Are these Previous Generation instances still being supported?

Yes. Previous Generation instances are still fully supported.

Q: Can I still use/add more Previous Generation instances?

Yes. Previous Generation instances are still available as On-Demand, Reserved Instances, and Spot Instance, from our APIs, CLI and EC2 Management Console interface.

Q: Are my Previous Generation instances going to be deleted?

No. Your C1, C3, CC2, CR1, G2, HS1, M1, M2, M3, R3 and T1 instances are still fully functional and will not be deleted because of this change.

Q: Are Previous Generation instances being discontinued soon?

Currently, there are no plans to end of life Previous Generation instances. However, with any rapidly evolving technology the latest generation will typically provide the best performance for the price and we encourage our customers to take advantage of technological advancements.

Q: Will my Previous Generation instances I purchased as a Reserved Instance be affected or changed?

No. Your Reserved Instances will not change, and the Previous Generation instances are not going away.

Storage Optimized instances

Q. What is a Dense-storage Instance?

Dense-storage instances are designed for workloads that require high sequential read and write access to very large data sets, such as Hadoop distributed computing, massively parallel processing data warehousing, and log processing applications. The Dense-storage instances offer the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q. How do dense-storage instances compare to High I/O instances?

High I/O instances (Im4gn, Is4gen, I4i, I3, I3en) are targeted at workloads that demand low latency and high random I/O in addition to moderate storage density and provide the best price/IOPS across other EC2 instance types. Dense-storage instances (D3, D3en, D2) and HDD-storage instances (H1) are optimized for applications that require high sequential read/write access and low cost storage for very large data sets and provide the best price/GB-storage and price/disk-throughput across other EC2 instances.

Q. How much disk throughput can Dense-storage and HDD-storage instances deliver?

The largest current generation of Dense HDD-storage instances, d3en.12xlarge, can deliver up to 6.2 GiB/s read and 6.2 GiB/s write disk throughput with a 128k block size. Please see the product detail page for additional performance information. To ensure the best disk throughput performance from your D2, D3 and D3en instances on Linux, we recommend that you use the most recent version of the Amazon Linux AMI, or another Linux AMI with a kernel version of 3.8 or later that supports persistent grants - an extension to the Xen block ring protocol that significantly improves disk throughput and scalability.

Q. Do Dense-storage and HDD-storage instances provide any failover mechanisms or redundancy?

D2 and H1 instances provide notifications for hardware failures. Like all instance storage, Dense HDD-storage volumes persist only for the life of the instance. Hence, we recommend that you build a degree of redundancy (e.g. RAID 1/5/6) or use file systems (e.g. HDFS and MapR-FS) that support redundancy and fault tolerance. You can also back up data periodically to more data storage solutions such as Amazon Elastic Block Store (EBS) or Simple Storage Service (S3).

Q. How do dense HDD-storage instances differ from Amazon EBS?

Amazon EBS offers simple, elastic, reliable (replicated), and persistent block level storage for Amazon EC2 while abstracting the details of the underlying storage media in use. Amazon EC2 instance instances with local HDD or NVMe storage provide directly attached, high performance storage building blocks that can be used for a variety of storage applications. Dense-storage instances are specifically targeted at customers who want high sequential read/write access to large data sets on local storage, e.g. for Hadoop distributed computing and massively parallel processing data warehousing.

Q. Can I launch dense HDD-storage instances as Amazon EBS-optimized instances?

Each HDD-storage instance type (H1, D2, D3, and D3en) is EBS-optimized by default. Since this feature is always enabled, launching one of these instances explicitly as EBS-optimized will not affect the instance's behavior.

Q. Can I launch D2 instances as Amazon EBS-optimized instances?

Each D2 instance type is EBS-optimized by default. D2 instances 500 Mbps to 4,000 Mbps to EBS above and beyond the general-purpose network throughput provided to the instance.

Since this feature is always enabled on D2 instances, launching a D2 instance explicitly as EBS-optimized will not affect the instance's behavior.

Q. Are Dense-storage instances offered in EC2 Classic?

The current generation of Dense-storage instances (D2 instances) can be launched in both EC2-Classic and Amazon VPC. However, by launching a Dense-storage instance into a VPC, you can leverage a number of features that are available only on the Amazon VPC platform – such as enabling enhanced networking, assigning multiple private IP addresses to your instances, or changing your instances' security groups.

Q. What is a High I/O instance?

High I/O instances use NVMe based local instance storage to deliver very high, low latency, I/O capacity to applications, and are optimized for applications that require millions of IOPS. Like Cluster instances, High I/O instances can be clustered via cluster placement groups for low latency networking.

Q. Are all features of Amazon EC2 available for High I/O instances?

High I/O instances support all Amazon EC2 features. Im4gn, Is4gen, I4i, I3 and I3en instances offer NVMe only storage, while previous generation I2 instances allow legacy blkfront storage access.

Q. AWS has other database and Big Data offerings. When or why should I use High I/O instances?

High I/O instances are ideal for applications that require access to millions of low latency IOPS, and can leverage data stores and architectures that manage data redundancy and availability. Example applications are:

- NoSQL databases like Cassandra and MongoDB
- In-memory databases like Aerospike
- Elasticsearch and analytics workloads
- OLTP systems

Q. Do High I/O instances provide any failover mechanisms or redundancy?

Like other Amazon EC2 instance types, instance storage on Im4gn, Is4gen, I4i, I3 and I3en instances persists during the life of the instance. Customers are expected to build resilience into their applications. We recommend using databases and file systems that support redundancy and fault tolerance. Customers should back up data periodically to Amazon S3 for improved data durability.

Q. Do High I/O instances support TRIM?

The TRIM command allows the operating system to inform SSDs which blocks of data are no longer considered in use and can be wiped internally. In the absence of TRIM, future write operations to the involved blocks can slow down significantly. Im4gn, Is4gen, I4i, I3 and I3en instances support TRIM.

Q. How do D3 and D3en instances compare to D2 instances?

D3 and D3en instances offer improved specifications over D2 on the following compute, storage and network attributes:

- D3 and D3en instances offer up to 30% higher compute performances than equivalent D2 instances. Exact performance benefit will depend on the specific workload.
- D3 and D3en instances provide up to 45% and 100% higher disk throughput than D2 instances, respectively.
- D3 instances are available at a price that is 5% lower than D2 instances. D3en instances lower cost per TB of storage by up to 80% compared to D2 instances.
- D3 and D3en instances offer Intel Advanced Vector Extensions (AVX 512), which offer up to 2X the FLOPS per cycle compared to AVX 2 on D2.
- D3en instances offer a new instance size (12xl) with 48 vCPUs and 7 TB of storage per vCPU for 336 TB of total storage, but have half the memory per vCPU compared to D2 and 48 TB of total storage.
- D3 and D3en instances offer up to 25 Gbps and 75 Gbps of network bandwidth respectively on their largest sizes to meet customer needs for network performance for running big data workloads and file system clusters.

Q. Do D3 and D3en instances encrypt storage volumes and network traffic?

Yes; data written onto the storage volumes will be encrypted at rest using AES-256-XTS. Network traffic between D3 and D3en instances in the same VPC or a peered VPC are encrypted by default using a 256-bit key.

Storage

Amazon Elastic Block Store (EBS)

Q: What happens to my data when a system terminates?

The data stored on a local instance store will persist only as long as that instance is alive. However, data that is stored on an Amazon EBS volume will persist independently of the life of the instance. Therefore, we recommend that you use the local instance store for temporary data and, for data requiring a higher level of durability, we recommend using Amazon EBS volumes or backing up the data to Amazon S3. If you are using an Amazon EBS volume as a root partition, you will need to set the Delete On Terminate flag to "N" if you want your Amazon EBS volume to persist outside the life of the instance.

Q: What kind of performance can I expect from Amazon EBS volumes?

Amazon EBS provides four current generation volume types that are divided into two major categories: SSD-backed storage for transactional workloads and HDD-backed storage for throughput intensive workloads. These volume types differ in performance characteristics and price, allowing you to tailor your storage performance and cost to the needs of your applications.

Q: What are Throughput Optimized HDD (st1) and Cold HDD (sc1) volume types?

ST1 volumes are backed by hard disk drives (HDDs) and are ideal for frequently accessed, throughput intensive workloads with large datasets and large I/O sizes, such as MapReduce, Kafka, log processing, data warehouse, and ETL workloads. These volumes deliver performance in terms of throughput, measured in MB/s, and include the ability to burst up to 250 MB/s per TB, with a baseline throughput of 40 MB/s per TB and a maximum throughput of 500 MB/s per volume. ST1 is designed to deliver the expected throughput performance 99% of the time and has enough I/O credits to support a full-volume scan at the burst rate.

SC1 volumes are backed by hard disk drives (HDDs) and provide the lowest cost per GB of all EBS volume types. It is ideal for less frequently accessed workloads with large, cold datasets. Similar to st1, sc1 provides a burst model: these volumes can burst up to 80 MB/s per TB, with a baseline throughput of 12 MB/s per TB and a maximum throughput of 250 MB/s per volume. For infrequently accessed data, sc1 provides extremely inexpensive storage. SC1 is designed to deliver the expected throughput performance 99% of the time and has enough I/O credits to support a full-volume scan at the burst rate.

To maximize the performance of st1 and sc1, we recommend using EBS-optimized EC2 instances.

Q: Which volume type should I choose?

Amazon EBS includes two major categories of storage: SSD-backed storage for transactional workloads (performance depends primarily on IOPS) and HDD-backed storage for throughput workloads (performance depends primarily on throughput, measured in MB/s). SSD-backed volumes are designed for transactional, IOPS-intensive database workloads, boot volumes, and workloads that require high IOPS. SSD-backed volumes include Provisioned IOPS SSD (io1 and io2) and General Purpose SSD (gp2 and gp3). HDD-backed volumes are designed for throughput-intensive and big-data workloads, large I/O sizes, and sequential I/O patterns. HDD-backed volumes include Throughput Optimized HDD (st1) and Cold HDD (sc1). For more information on Amazon EBS see the EBS product details page.

Q: Do you support multiple instances accessing a single volume?

Yes, you can enable Multi-Attach on an EBS Provisioned IOPS io1 volume to allow a volume to be concurrently attached to up to sixteen Nitro-based EC2 instances within the same Availability Zone. For more information on Amazon EBS Multi-Attach, see the EBS product page.

Q: Will I be able to access my EBS snapshots using the regular Amazon S3 APIs?

No, EBS snapshots are only available through the Amazon EC2 APIs.

Q: Do volumes need to be un-mounted in order to take a snapshot? Does the snapshot need to complete before the volume can be used again?

No, snapshots can be done in real time while the volume is attached and in use. However, snapshots only capture data that has been written to your Amazon EBS volume, which might exclude any data that has been locally cached by your application or OS. In order to ensure consistent snapshots on volumes attached to an instance, we recommend cleanly detaching the volume, issuing the snapshot command, and then reattaching the volume. For Amazon EBS volumes that serve as root devices, we recommend shutting down the machine to take a clean snapshot.

Q: Are snapshots versioned? Can I read an older snapshot to do a point-in-time recovery?

Each snapshot is given a unique identifier, and customers can create volumes based on any of their existing snapshots.

Q: What charges apply when using Amazon EBS shared snapshots?

If you share a snapshot, you won't be charged when other users make a copy of your snapshot. If you make a copy of another user's shared volume, you will be charged normal EBS rates.

Q: Can users of my Amazon EBS shared snapshots change any of my data?

Users who have permission to create volumes based on your shared snapshots will first make a copy of the snapshot into their account. Users can modify their own copies of the data, but the data on your original snapshot and any other volumes created by other users from your original snapshot will remain unmodified.

Q: How can I discover Amazon EBS snapshots that have been shared with me?

You can find snapshots that have been shared with you by selecting “Private Snapshots” from the viewing dropdown in the Snapshots section of the AWS Management Console. This section will list both snapshots you own and snapshots that have been shared with you.

Q: How can I find what Amazon EBS snapshots are shared globally?

You can find snapshots that have been shared globally by selecting “Public Snapshots” from the viewing dropdown in the Snapshots section of the AWS Management Console.

Q: Do you offer encryption on Amazon EBS volumes and snapshots?

Yes. EBS offers seamless encryption of data volumes and snapshots. EBS encryption better enables you to meet security and encryption compliance requirements.

Q: How can I find a list of Amazon Public Data Sets?

All information on Public Data Sets is available in our Public Data Sets Resource Center. You can also obtain a listing of Public Data Sets within the AWS Management Console by choosing “Amazon Snapshots” from the viewing dropdown in the Snapshots section.

Amazon Elastic File System (EFS)

Q. How do I access a file system from an Amazon EC2 instance?

To access your file system, you mount the file system on an Amazon EC2 Linux-based instance using the standard Linux mount command and the file system’s DNS name. Once you’ve mounted, you can work with the files and directories in your file system just like you would with a local file system.

Amazon EFS uses the NFSv4.1 protocol.

Q. What Amazon EC2 instance types and AMIs work with Amazon EFS?

Amazon EFS is compatible with all Amazon EC2 instance types and is accessible from Linux-based AMIs. You can mix and match the instance types connected to a single file system.

Q. How do I load data into a file system?

You can load data into an Amazon EFS file system from your Amazon EC2 instances or from your on-premises datacenter servers.

Amazon EFS file systems can be mounted on an Amazon EC2 instance, so any data that is accessible to an Amazon EC2 instance can also be read and written to Amazon EFS. To load data that is not currently stored on the Amazon cloud, you can use the same methods you use to transfer files to Amazon EC2 today, such as Secure Copy (SCP).

Amazon EFS file systems can also be mounted on an on-premises server, so any data that is accessible to an on-premises server can be read and written to Amazon EFS using standard Linux tools.

Q. How do I access my file system from outside my VPC?

Amazon EC2 instances within your VPC can access your file system directly, and Amazon EC2 Classic instances outside your VPC can mount a file system via ClassicLink. On-premises servers can mount your file systems via an AWS Direct Connect connection to your VPC.

Q. How many Amazon EC2 instances can connect to a file system?

Amazon EFS supports one to thousands of Amazon EC2 instances connecting to a file system concurrently.

NVMe Instance storage

Q: Is data stored on Amazon EC2 NVMe instance storage encrypted?

Yes, all data is encrypted in an AWS Nitro hardware module prior to being written on the locally attached SSDs offered via NVMe instance storage.

Q: What encryption algorithm is used to encrypt Amazon EC2 NVMe instance storage?

Amazon EC2 NVMe instance storage is encrypted using an XTS-AES-256 block cipher.

Q: Are encryption keys unique to an instance or a particular device for NVMe instance storage?

Encryption keys are securely generated within the Nitro hardware module, and are unique to each NVMe instance storage device that is provided with an EC2 instance.

Q: What is the lifetime of encryption keys on NVMe instance storage?

All keys are irrecoverably destroyed on any de-allocation of the storage, including instance stop and instance terminate actions.

Q: Can I disable NVMe instance storage encryption?

No, NVMe instance storage encryption is always on, and cannot be disabled.

Q: Do the published IOPS performance numbers on I3 and I3en include data encryption?

Yes, the documented IOPS numbers for Im4gn, Is4gen, I4i, I3 and I3en NVMe instance storage include encryption.

Q: Does Amazon EC2 NVMe instance storage support AWS Key Management Service (KMS)?

No, disk encryption on NVMe instance storage does not support integration with AWS KMS system. Customers cannot bring in their own keys to use with NVMe instance storage.

Networking and security

Elastic Fabric Adapter (EFA)

Q: Why should I use EFA?

EFA brings the scalability, flexibility, and elasticity of cloud to tightly-coupled HPC applications. With EFA, tightly-coupled HPC applications have access to lower and more consistent latency and higher throughput than traditional TCP channels, enabling them to scale better. EFA

support can be enabled dynamically, on-demand on any supported EC2 instance without pre-reservation, giving you the flexibility to respond to changing business/workload priorities.

Q: What types of applications can benefit from using EFA?

High Performance Computing (HPC) applications distribute computational workloads across a cluster of instances for parallel processing. Examples of HPC applications include computational fluid dynamics (CFD), crash simulations, and weather simulations. HPC applications are generally written using the Message Passing Interface (MPI) and impose stringent requirements for inter-instance communication in terms of both latency and bandwidth. Applications using MPI and other HPC middleware that supports the libfabric communication stack can benefit from EFA.

Q: How does EFA communication work?

EFA devices provide all ENA devices' functionalities plus a new OS bypass hardware interface that allows user-space applications to communicate directly with the hardware-provided reliable transport functionality. Most applications will use existing middleware, such as the Message Passing Interface (MPI), to interface with EFA. AWS has worked with a number of middleware providers to ensure support for the OS bypass functionality of EFA. Please note that communication using the OS bypass functionality is limited to instances within a single subnet of a Virtual Private Cloud (VPC).

Q: Which instance types support EFA?

EFA is currently available on the m6a.48xlarge, m6i.32xlarge, m6i.metal, m6id.32xlarge, m6id.metal, m6idn.32xlarge, m6in.32xlarge, m5n.24xlarge, m5dn.24xlarge, m5n.metal, m5dn.metal, r6idn.32xlarge, r6in.32xlarge, r6a.48xlarge, r6a.metal, r6i.32xlarge, r6i.metal, r6id.32xlarge, r6id.metal, r5n.24xlarge, r5dn.24xlarge, r5n.metal, r5dn.metal, x2idn.32xlarge, x2iedn.32xlarge, c7g.16xlarge, c7gn.16xlarge, c6a.48xlarge, c6i.32xlarge, c6i.metal, c6id.32xlarge, c6id.metal,c6in.32xlarge, c5n.18xlarge, c5n.metal, p3dn.24xlarge, i3en.24xlarge, and i3en.metal instance sizes. Support for more instance types and sizes is being added in the coming months.

Q: What are the differences between an EFA ENI and an ENA ENI?

An ENA ENI provides traditional IP networking features necessary to support VPC networking. An EFA ENI provides all the functionality of an ENA ENI, plus hardware support for applications to communicate directly with the EFA ENI without involving the instance kernel (OS-bypass communication) using an extended programming interface. Due to the advanced capabilities of the EFA ENI, EFA ENIs can only be attached at launch or to stopped instances.

Q: What are the pre-requisites to enabling EFA on an instance?

EFA support can be enabled either at the launch of the instance or added to a stopped instance. EFA devices cannot be attached to a running instance.

Elastic IP

Q: Why am I limited to 5 Elastic IP addresses per region?

Public (IPV4) internet addresses are a scarce resource. There is only a limited amount of public IP space available, and Amazon EC2 is committed to helping use that space efficiently.

By default, all accounts are limited to 5 Elastic IP addresses per region. If you need more than 5 Elastic IP addresses, we ask that you apply for your limit to be raised. We will ask you to think through your use case and help us understand your need for additional addresses. You can apply for more Elastic IP addresses here. Any increases will be specific to the region they have been requested for.

Q: Why am I charged when my Elastic IP address is not associated with a running instance?

In order to help ensure our customers are efficiently using the Elastic IP addresses, we impose a small hourly charge for each address when it is not associated with a running instance.

Q: Do I need one Elastic IP address for every instance that I have running?

No. You do not need an Elastic IP address for all your instances. By default, every instance comes with a private IP address and an internet routable public IP address. The private IP address remains associated with the network interface when the instance is stopped and restarted, and is released when the instance is terminated. The public address is associated exclusively with the instance until it is stopped, terminated or replaced with an Elastic IP address. These IP addresses should be adequate for many applications where you do not need a long lived internet routable end point. Compute clusters, web crawling, and backend services are all examples of applications that typically do not require Elastic IP addresses.

Q: How long does it take to remap an Elastic IP address?

The remap process currently takes several minutes from when you instruct us to remap the Elastic IP until it fully propagates through our system.

Q: Can I configure the reverse DNS record for my Elastic IP address?

All Elastic IP addresses come with reverse DNS, in a standard template of the form ec2-1-2-3-4.region.compute.amazonaws.com. For customers requiring custom reverse DNS settings for internet-facing applications that use IP-based mutual authentication (such as sending email from EC2 instances), you can configure the reverse DNS record of your Elastic IP address by filling out this form. Alternatively, please contact AWS Customer Support if you want AWS to delegate the management of the reverse DNS for your Elastic IPs to your authoritative DNS name servers (such as Amazon Route 53), so that you can manage your own reverse DNS PTR records to support these use-cases. Note that a corresponding forward DNS record pointing to that Elastic IP address must exist before we can create the reverse DNS record.

Elastic Load Balancing

Q: What load balancing options does the Elastic Load Balancing service offer?

Elastic Load Balancing offers two types of load balancers that both feature high availability, automatic scaling, and robust security. These include the Classic Load Balancer that routes traffic based on either application or network level information, and the Application Load Balancer that routes traffic based on advanced application level information that includes the content of the request.

Q: When should I use the Classic Load Balancer and when should I use the Application Load Balancer?

The Classic Load Balancer is ideal for simple load balancing of traffic across multiple EC2 instances, while the Application Load Balancer is ideal for applications needing advanced routing capabilities, microservices, and container-based architectures.

Enhanced networking

Q: What networking capabilities are included in this feature?

We currently support enhanced networking capabilities using SR-IOV (Single Root I/O Virtualization). SR-IOV is a method of device virtualization that provides higher I/O performance and lower CPU utilization compared to traditional implementations. For supported Amazon EC2 instances, this feature provides higher packet per second (PPS) performance, lower inter-instance latencies, and very low network jitter.

Q: Why should I use Enhanced Networking?

If your applications benefit from high packet-per-second performance and/or low latency networking, Enhanced Networking will provide significantly improved performance, consistence of performance and scalability.

Q: How can I enable Enhanced Networking on supported instances?

In order to enable this feature, you must launch an HVM AMI with the appropriate drivers. The instances listed as current generation use ENA for enhanced networking. Amazon Linux AMI includes both of these drivers by default. For AMIs that do not contain these drivers, you will need to download and install the appropriate drivers based on the instance types you plan to use. You can use Linux or Windows instructions to enable Enhanced Networking in AMIs that do not include the SR-IOV driver by default. Enhanced Networking is only supported in Amazon VPC.

Q: Do I need to pay an additional fee to use Enhanced Networking?

No, there is no additional fee for Enhanced Networking. To take advantage of Enhanced Networking you need to launch the appropriate AMI on a supported instance type in a VPC.

Q: Why is Enhanced Networking only supported in Amazon VPC?

Amazon VPC allows us to deliver many advanced networking features to you that are not possible in EC2-Classic. Enhanced Networking is another example of a capability enabled by Amazon VPC.

Q: Which instance types support Enhanced Networking?

Depending on your instance type, enhanced networking can be enabled using one of the following mechanisms:

Intel 82599 Virtual Function (VF) interface - The Intel 82599 Virtual Function interface supports network speeds of up to 10 Gbps for supported instance types. C3, C4, D2, I2, M4 (excluding m4.16xlarge), and R3 instances use the Intel 82599 VF interface for enhanced networking.

Elastic Network Adapter (ENA) - The Elastic Network Adapter (ENA) supports network speeds of up to 100 Gbps for supported instance types. The instances listed as current generation use ENA for enhanced networking, with the exception of C4, D2, and M4 instances smaller than m4.16xlarge.

Q: What does it mean to have multiple network cards for an EC2 instance? Why is it needed?

Newer generation EC2 instances use Nitro network cards for VPC data-plane offloading. To

provide higher network bandwidth and improved packet rate performance, select EC2 instances can be configured to use multiple network cards for packet processing ultimately increasing overall system performance.

Q: What instance types support multiple Network cards?

Multiple network cards are supported on accelerated instances such as p4d.24xlarge and network optimized instances such as c6in.32xlarge. For a full list of instances supporting multiple network cards, refer to the Elastic network interfaces page.

Q: What is the default number of network interfaces a multiple card instance can launch with?

This depends on the instance type. Accelerated instances such as p4 scale up to 15 network interfaces per network card. High network instances such as the recently launched c6in instances support an aggregate 14 network interfaces split evenly (7 and 7) across the two network cards. Network interface scale per network card is available in the Network cards section of Elastic network interfaces page.

Security

Q: How do I prevent other people from viewing my systems?

You have complete control over the visibility of your systems. The Amazon EC2 security systems allow you to place your running instances into arbitrary groups of your choice. Using the web services interface, you can then specify which groups may communicate with which other groups, and also which IP subnets on the Internet may talk to which groups. This allows you to control access to your instances in our highly dynamic environment. Of course, you should also secure your instance as you would any other server.

Q: Can I get a history of all EC2 API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. To receive a history of all EC2 API calls (including VPC and EBS) made on your account, you simply turn on CloudTrail in the AWS Management Console. For more information, visit the CloudTrail home page.

Elastic Network Adapter (ENA) Express

Q. What is ENA Express?

ENA Express is an enhancement on the Elastic Network Adapter that brings the Scalable Reliable Datagram (SRD) protocol to traditional TCP and UDP networking. Transparent to the application, ENA Express improves single flow bandwidths and reduces tail latencies in throughput intensive workloads.

Q. How does ENA Express work?

When configured, ENA Express works between any two supported instances in an Availability Zone. ENA Express detects compatibility between your EC2 instances and establishes an SRD connection when both communicating instances have ENA Express enabled. Once a connection is established, your traffic takes advantage of SRD and its performance benefits.

Q. When should I use ENA Express?

ENA Express works best for applications requiring high, single flow throughput like distribute storage systems and live video encoding. These workloads require high single flow bandwidth and low tail latency.

Q. How do I enable ENA Express?

ENA Express can be enabled on a per -ENI basis. While attaching a network card to an instance or while running a modify command, ENA Express can be enabled. ENA Express must be enabled on both communicating ENIs to establish point-to-point communication with it. Additionally, if you are using Jumbo Frames, you must adjust your maximum MTU to 8900 to use ENA Express.

Q. What protocols are supported by ENA Express?

ENA Express supports TCP by default. UDP can be enabled through an API argument during the interface attachment (AttachNetworkInterface) or modification process (ModifyNetworkInterfaceAttribute).

Q. What instances are supported?

ENA Express is supported on C6gn.16xl. Support for more instance types and sizes will be added in the coming months.

Q. What is the difference between EFA and ENA Express?

EFA is a network adapter built for HPC and ML applications, and it also leverages the SRD protocol. Unlike EFA, ENA Express helps you run your application transparently on TCP and UDP. EFA requires a different network programming model, which uses the LibFabric interface to pass communication to the ENI. Additionally, ENA Express allows for intra-Availability Zone (AZ) communication, while EFA is currently limited to communication with the same subnet.

Q. What happens if I'm running ENA Express on one instance, and it is communicating with another instance that doesn't support ENA Express or hasn't enabled it on the ENI?

ENA Express will detect if ENA Express has been enabled on another instance. If that instance doesn't support or hasn't enabled ENA Express, your instance will fallback to normal ENA operation. You will not be able to achieve any of the SRD performance benefits in this case.

Q. What monitoring tools are available to track this?

ENA Express offers EthTool counter metrics to track packets that eligible for SRD transmission in addition to those actually sent and received with SRD. Additionally, EthTool will support an SRD resource utilization metric on a percent basis, providing insight into when you should consider scaling out your architecture. Finally, boolean metrics will provide on and off status for ENA Express and the UDP protocol.

Q. What operating systems are supported?

The SRD functionality will be supported with all operating systems, but please note that ENA Express metrics will be available on only the EthTool in the latest Amazon Linux AMI or by installing the ENA driver version 2.8.0.g or later from GitHub, with all operating systems supporting the metrics in the future.

Q. Where is ENA Express available?

ENA Express is available in all commercial Regions. It can be used to establish communication between any two enabled instances within the same AZ.

Q. EFA and ENA Express both use SRD, how do they differ at the transport layer?

EFA and ENA Express both use the Scalable Reliable Datagram (SRD) protocol, built by AWS. EFA is purpose built for tightly coupled workloads to have direct hardware-provide transport communication to the application layer. ENA Express is designed to use the SRD protocol transparently for traditional networking applications that use TCP and UDP protocols.

Q. Are there any additional costs to running ENA Express?

No, ENA Express is free to use.

Management

Amazon CloudWatch

Q: What is the minimum time interval granularity for the data that Amazon CloudWatch receives and aggregates?

Metrics are received and aggregated at 1 minute intervals.

Q: Which operating systems does Amazon CloudWatch support?

Amazon CloudWatch receives and provides metrics for all Amazon EC2 instances and should work with any operating system currently supported by the Amazon EC2 service.

Q: Will I lose the metrics data if I disable monitoring for an Amazon EC2 instance?

You can retrieve metrics data for any Amazon EC2 instance up to 2 weeks from the time you started to monitor it. After 2 weeks, metrics data for an Amazon EC2 instance will not be available if monitoring was disabled for that Amazon EC2 instance. If you want to archive metrics beyond 2 weeks you can do so by calling mon-get-stats command from the command line and storing the results in Amazon S3 or Amazon SimpleDB.

Q: Can I access the metrics data for a terminated Amazon EC2 instance or a deleted Elastic Load Balancer?

Yes. Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 2 weeks.

Q: Does the Amazon CloudWatch monitoring charge change depending on which type of Amazon EC2 instance I monitor?

No, the Amazon CloudWatch monitoring charge does not vary by Amazon EC2 instance type.

Q: Why does the graphing of the same time window look different when I view in 5 minute and 1 minute periods?

If you view the same time window in a 5 minute period versus a 1 minute period, you may see that data points are displayed in different places on the graph. For the period you specify in your graph, Amazon CloudWatch will find all the available data points and calculates a single, aggregate point to represent the entire period. In the case of a 5 minute period, the single data point is placed at the beginning of the 5 minute time window. In the case of a 1 minute period,

the single data point is placed at the 1 minute mark. We recommend using a 1 minute period for troubleshooting and other activities that require the most precise graphing of time periods.

Amazon EC2 Auto Scaling

Q: Can I automatically scale Amazon EC2 Auto Scaling Groups?

Yes. Amazon EC2 Auto Scaling is a fully managed service designed to launch or terminate Amazon EC2 instances automatically to help ensure you have the correct number of Amazon EC2 instances available to handle the load for your application. EC2 Auto Scaling helps you maintain application availability through fleet management for EC2 instances, which detects and replaces unhealthy instances, and by scaling your Amazon EC2 capacity up or down automatically according to conditions you define. You can use EC2 Auto Scaling to automatically increase the number of Amazon EC2 instances during demand spikes to maintain performance and decrease capacity during lulls to reduce costs.

Allocation strategies in EC2 Auto Scaling determine how Spot Instances in your fleet are fulfilled from Spot Instance pools. The capacity-optimized allocation strategy attempts to provision Spot Instances from the most available Spot Instance pools by analyzing capacity metrics. This strategy is a good choice for workloads that have a higher cost of interruption such as big data and analytics, image and media rendering, machine learning, and high performance computing. The lowest-price allocation strategy launches Spot Instances strictly based on diversification across 'N' lowest priced pools.

Hibernate

Q: Why should I hibernate an instance?

You can hibernate an instance to get your instance and applications up and running quickly, if they take a long time to bootstrap (e.g. load memory caches). You can start instances, bring them to a desired state and hibernate them. These "pre-warmed" instances can then be resumed to reduce the time it takes for an instance to return to service. Hibernation retains memory state across Stop/Start cycles.

Q: What happens when I hibernate my instance?

When you hibernate an instance, data from your EBS root volume and any attached EBS data volumes is persisted. Additionally, contents from the instance's memory (RAM) are persisted to EBS root volume. When the instance is restarted, it returns to its previous state and reloads the RAM contents.

Q: What is the difference between hibernate and stop?

In the case of hibernate, your instance gets hibernated and the RAM data persisted. In the case of Stop, your instance gets shut down and RAM is cleared.

In both the cases, data from your EBS root volume and any attached EBS data volumes is persisted. Your private IP address remains the same, as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow. Stop and hibernate are available for Amazon EBS backed instances only. Local instance storage is not persisted.

Q: How much does it cost to hibernate an instance?

Hibernating instances are charged at standard EBS rates for storage. As with a stopped instance, you do not incur instance usage fees while an instance is hibernating.

Q: How can I hibernate an instance?

Hibernation needs to be enabled when you launch the instance. Once enabled, you can use the StopInstances API with an additional 'Hibernate' parameter to trigger hibernation. You can also do this through the console by selecting your instance, then clicking Actions> Instance State > Stop - Hibernate. For more information on using hibernation, refer to the user guide.

Q: How can I resume a hibernating instance?

You can resume by calling the StartInstances API as you would for a regular stopped instance. You can also do this through the console by selecting your instance, then clicking Actions > Instance State > Start.

Q: Can I enable hibernation on an existing instance?

No, you cannot enable hibernation on an existing instance (running or stopped). This needs to be enabled during instance launch.

Q: How can I tell that an instance is hibernated?

You can tell that an instance is hibernated by looking at the state reason. It should be 'Client.UserInitiatedHibernate'. This is visible on the console under "Instances - Details" view or in the DescribeInstances API response as the "reason" field.

Q: What is the state of an instance when it is hibernating?

Hibernated instances are in 'Stopped' state.

Q: What data is saved when I hibernate an instance?

EBS volume storage (boot volume and attached data volumes) and memory (RAM) are saved. Your private IP address remains the same (for VPC), as does your elastic IP address (if applicable). The network layer behavior will be similar to that of EC2 Stop-Start workflow.

Q: Where is my data stored when I hibernate an instance?

As with the Stop feature, root device and attached device data are stored on the corresponding EBS volumes. Memory (RAM) contents are stored on the EBS root volume.

Q: Is my memory (RAM) data encrypted when it is moved to EBS?

Yes, RAM data is always encrypted when it is moved to the EBS root volume. Encryption on the EBS root volume is enforced at instance launch time. This is to ensure protection for any sensitive content that is in memory at the time of hibernation.

Q: How long can I keep my instance hibernated?

We do not support keeping an instance hibernated for more than 60 days. You need to resume the instance and go through Stop and Start (without hibernation) if you wish to keep the instance around for a longer duration. We are constantly working to keep our platform up-to-date with upgrades and security patches, some of which can conflict with the old hibernated instances. We will notify you for critical updates that require you to resume the hibernated instance to perform a shutdown or a reboot.

Q: What are the prerequisites to hibernate an instance?

To use hibernation, the root volume must be an encrypted EBS volume. The instance needs to be configured to receive the ACPID signal for hibernation (or use the Amazon published AMIs that are configured for hibernation). Additionally, your instance should have sufficient space available on your EBS root volume to write data from memory.

Q: Which instances and operating systems support hibernation?

For instances running Amazon Linux, Amazon Linux 2, Ubuntu, and Windows, Hibernation is supported across C3, C4, C5, C5d, I3, M3, M4, M5, M5a, M5ad, M5d, R3, R4, R5, R5a, R5ad, R5d, T2, T3, and T3a instances.

For instances running CentOS, Fedora, and Red Hat Enterprise Linux, Hibernation is supported across C5, C5d, M5, M5a, M5ad, M5d, R5, R5a, R5ad, R5d, T3, and T3a instances.

For Windows, Hibernation is supported for instances up to 16 GB of RAM. For other operating systems, Hibernation is supported for instances with less than 150 GB of RAM. To review the list of supported OS versions and instance types, refer to the user guide.

Q: Should I use specific Amazon Machine Image (AMIs) if I want to hibernate my instance?

You can use any AMI that is configured to support hibernation. You can use AWS published AMIs that support hibernation by default. Alternatively, you can create a custom image from an instance after following the hibernation pre-requisite checklist and configuring your instance appropriately.

Q: What if my EBS root volume is not large enough to store memory state (RAM) for hibernation?

To enable hibernation, space is allocated on the root volume to store the instance memory (RAM). Make sure that the root volume is large enough to store the RAM contents and accommodate your expected usage, e.g. OS, applications. If the EBS root volume does not have enough space, hibernation will fail and the instance will get shut down instead.

VM Import/Export

Q. What is VM Import/Export?

VM Import/Export enables customers to import Virtual Machine (VM) images in order to create Amazon EC2 instances. Customers can also export previously imported EC2 instances to create VMs. Customers can use VM Import/Export to leverage their previous investments in building VMs by migrating their VMs to Amazon EC2.
Q. What operating systems are supported?

VM Import/Export currently supports Windows and Linux VMs, including Windows Server 2003, Windows Server 2003 R2, Windows Server 2008, Windows Server 2012 R1, Red Hat Enterprise Linux (RHEL) 5.1-6.5 (using Cloud Access), Centos 5.1-6.5, Ubuntu 12.04, 12.10, 13.04, 13.10, and Debian 6.0.0-6.0.8, 7.0.0-7.2.0. For more details on VM Import, including supported file formats, architectures, and operating system configurations, please see the VM Import/Export section of the Amazon EC2 User Guide.

Q. What virtual machine file formats are supported?

You can import VMware ESX VMDK images, Citrix Xen VHD images, Microsoft Hyper-V VHD images and RAW images as Amazon EC2 instances. You can export EC2 instances to VMware ESX VMDK, VMware ESX OVA, Microsoft Hyper-V VHD or Citrix Xen VHD images.

Q. What is VMDK?

VMDK is a file format that specifies a virtual machine hard disk encapsulated within a single file. It is typically used by virtual IT infrastructures such as those sold by VMware, Inc.

Q. How do I prepare a VMDK file for import using the VMware vSphere client?

The VMDK file can be prepared by calling File-Export-Export to OVF template in VMware vSphere Client. The resulting VMDK file is compressed to reduce the image size and is compatible with VM Import/Export. No special preparation is required if you are using the Amazon EC2 VM Import Connector vApp for VMware vCenter.

Q. What is VHD?

VHD (Virtual Hard Disk) is a file format that specifies a virtual machine hard disk encapsulated within a single file. The VHD image format is used by virtualization platforms such as Microsoft Hyper-V and Citrix Xen.

Q. How do I prepare a VHD file for import from Citrix Xen?

Open Citrix XenCenter and select the virtual machine you want to export. Under the Tools menu, choose "Virtual Appliance Tools" and select "Export Appliance" to initiate the export task. When the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q. How do I prepare a VHD file for import from Microsoft Hyper-V?

Open the Hyper-V Manager and select the virtual machine you want to export. In the Actions pane for the virtual machine, select "Export" to initiate the export task. Once the export completes, you can locate the VHD image file in the destination directory you specified in the export dialog.

Q. Are there any other requirements when importing a VM into Amazon EC2?

The virtual machine must be in a stopped state before generating the VMDK or VHD image. The VM cannot be in a paused or suspended state. We suggest that you export the virtual machine with only the boot volume attached. You can import additional disks using the ImportVolume command and attach them to the virtual machine using AttachVolume. Additionally, encrypted disks (e.g. Bit Locker) and encrypted image files are not supported. You are also responsible for ensuring that you have all necessary rights and licenses to import into AWS and run any software included in your VM image.

Q. Does the virtual machine need to be configured in any particular manner to enable import to Amazon EC2?

Ensure Remote Desktop (RDP) or Secure Shell (SSH) is enabled for remote access and verify that your host firewall (Windows firewall, iptables, or similar), if configured, allows access to RDP or SSH. Otherwise, you will not be able to access your instance after the import is complete. Please also ensure that Windows VMs are configured to use strong passwords for all users including the administrator and that Linux VMs are configured with a public key for SSH access.

Q. How do I import a virtual machine to an Amazon EC2 instance?

You can import your VM images using the Amazon EC2 API tools:

- Import the VMDK, VHD or RAW file via the ec2-import-instance API. The import instance task captures the parameters necessary to properly configure the Amazon EC2 instance properties (instance size, Availability Zone, and security groups) and uploads the disk image into Amazon S3.
- If ec2-import-instance is interrupted or terminates without completing the upload, use ec2-resume-import to resume the upload. The import task will resume where it left off.
- Use the ec2-describe-conversion-tasks command to monitor the import progress and obtain the resulting Amazon EC2 instance ID.
- Once your import task is completed, you can boot the Amazon EC2 instance by specifying its instance ID to the ec2-run-instances API.
- Finally, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3 as it is no longer needed.

Alternatively, if you use the VMware vSphere virtualization platform, you can import your virtual machine to Amazon EC2 using a graphical user interface provided through AWS Management Portal for vCenter. Please refer to the Getting Started Guide in AWS Management Portal for vCenter. AWS Management Portal for vCenter includes integrated support for VM Import. Once the portal is installed within vCenter, you can right-click on a VM and select "Migrate to EC2" to create an EC2 instance from the VM. The portal will handle exporting the VM from vCenter, uploading it to S3, and converting it into an EC2 instance for you, with no additional work required. You can also track the progress of your VM migrations within the portal.

Q. How do I export an Amazon EC2 instance back to my on-premise virtualization environment?

You can export your Amazon EC2 instance using the Amazon EC2 CLI tools:

- Export the instance using the ec2-create-instance-export-task command. The export command captures the parameters necessary (instance ID, S3 bucket to hold the exported image, name of the exported image, VMDK, OVA or VHD format) to properly export the instance to your chosen format. The exported file is saved in an S3 bucket that you previously created.
- Use ec2-describe-export-tasks to monitor the export progress.
- Use ec2-cancel-export-task to cancel an export task prior to completion.

Q. Are there any other requirements when exporting an EC2 instance using VM Import/Export?

You can export running or stopped EC2 instances that you previously imported using VM Import/Export. If the instance is running, it will be momentarily stopped to snapshot the boot volume. EBS data volumes cannot be exported. EC2 instances with more than one network interface cannot be exported.

Q. Can I export Amazon EC2 instances that have one or more EBS data volumes attached?

Yes, but VM Import/Export will only export the boot volume of the EC2 instance.

Q. What does it cost to import a virtual machine?

You will be charged standard Amazon S3 data transfer and storage fees for uploading and storing your VM image file. Once your VM is imported, standard Amazon EC2 instance hour and EBS service fees apply. If you no longer wish to store your VM image file in S3 after the import process completes, use the ec2-delete-disk-image command line tool to delete your disk image from Amazon S3.

Q. What does it cost to export a virtual machine?

You will be charged standard Amazon S3 storage fees for storing your exported VM image file. You will also be charged standard S3 data transfer charges when you download the exported VM file to your on-premise virtualization environment. Finally, you will be charged standard EBS charges for storing a temporary snapshot of your EC2 instance. To minimize storage charges, delete the VM image file in S3 after downloading it to your virtualization environment.

Q. When I import a VM of Windows Server 2003 or 2008, who is responsible for supplying the operating system license?

When you launch an imported VM using Microsoft Windows Server 2003 or 2008, you will be charged standard instance hour rates for Amazon EC2 running the appropriate Windows Server version, which includes the right to utilize that operating system within Amazon EC2. You are responsible for ensuring that all other installed software is properly licensed.

So then, what happens to my on-premise Microsoft Windows license key when I import a VM of Windows Server 2003 or 2008? Since your on-premise Microsoft Windows license key that was associated with that VM is not used when running your imported VM as an EC2 instance, you can reuse it for another VM within your on-premise environment.

Q. Can I continue to use the AWS-provided Microsoft Windows license key after exporting an EC2 instance back to my on-premise virtualization environment?

No. After an EC2 instance has been exported, the license key utilized in the EC2 instance is no longer available. You will need to reactivate and specify a new license key for the exported VM after it is launched in your on-premise virtualization platform.

Q. When I import a VM with Red Hat Enterprise Linux (RHEL), who is responsible for supplying the operating system license?

When you import Red Hat Enterprise Linux (RHEL) VM images, you can use license portability for your RHEL instances. With license portability, you are responsible for maintaining the RHEL licenses for imported instances, which you can do using Cloud Access subscriptions for Red Hat Enterprise Linux. Please contact Red Hat to learn more about Cloud Access and to verify your eligibility.

Q. How long does it take to import a virtual machine?

The length of time to import a virtual machine depends on the size of the disk image and your network connection speed. As an example, a 10 GB Windows Server 2008 SP2 VMDK image takes approximately 2 hours to import when it's transferred over a 10 Mbps network connection. If you have a slower network connection or a large disk to upload, your import may take significantly longer.

Q. In which Amazon EC2 regions can I use VM Import/Export?

Visit the Region Table page to see product service availability by region.

Q. How many simultaneous import or export tasks can I have?

Each account can have up to five active import tasks and five export tasks per region.

Q. Can I run imported virtual machines in Amazon Virtual Private Cloud (VPC)?

Yes, you can launch imported virtual machines within Amazon VPC.

Q. Can I use the AWS Management Console with VM Import/Export?

No. VM Import/Export commands are available via EC2 CLI and API. You can also use the AWS Management Portal for vCenter to import VMs into Amazon EC2. Once imported, the resulting instances are available for use via the AWS Management Console.

Platform

Amazon Time Sync Service

Q. How do I use this service?

The service provides an NTP endpoint at a link-local IP address (169.254.169.123) accessible from any instance running in a VPC. Instructions for configuring NTP clients are available for Linux and Windows.

Q. What are the key benefits of using this service?

A consistent and accurate reference time source is crucial for many applications and services. The Amazon Time Sync Service provides a time reference that can be securely accessed from an instance without requiring VPC configuration changes and updates. It is built on Amazon's proven network infrastructure and uses redundant reference time sources to ensure high accuracy and availability.

Q. Which instance types are supported for this service?

All instances running in a VPC can access the service.

Availability zones

Q: How isolated are Availability Zones from one another?

Each Availability Zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Common points of failures like generators and cooling equipment are not shared across Availability Zones. Additionally, they are physically separate, such that even extremely uncommon disasters such as fires, tornados or flooding would only affect a single Availability Zone.

Q: How can I make sure that I am in the same Availability Zone as another developer?

We do not currently support the ability to coordinate launches into the same Availability Zone across AWS developer accounts. One Availability Zone name (for example, us-east-1a) in two AWS customer accounts may relate to different physical Availability Zones.

Q: If I transfer data between Availability Zones using public IP addresses, will I be charged twice for Regional Data Transfer (once because it's across zones, and a second time because I'm using public IP addresses)?

No. Regional Data Transfer rates apply if at least one of the following is true, but you are only charged once for a given instance even if both are true:

- The other instance is in a different Availability Zone, regardless of which type of address is used.
- Public or Elastic IP addresses are used, regardless of which Availability Zone the other instance is in.

Cluster instances

Q. What is a Cluster Compute Instance?

Cluster Compute Instances combine high compute resources with high performance networking for High Performance Compute (HPC) applications and other demanding network-bound applications. Cluster Compute Instances provide similar functionality to other Amazon EC2 instances but have been specifically engineered to provide high performance networking.

Amazon EC2 cluster placement group functionality allows users to group Cluster Compute Instances in clusters – allowing applications to get the low-latency network performance necessary for tightly-coupled node-to-node communication typical of many HPC applications. Cluster Compute Instances also provide significantly increased network throughput both within the Amazon EC2 environment and to the Internet. As a result, these instances are also well suited for customer applications that need to perform network-intensive operations.

Q. What kind of network performance can I expect when I launch instances in a cluster placement group?

The bandwidth an EC2 instance can utilize in a cluster placement group depends on the instance type and its networking performance specification. Inter-instance traffic within the same region can utilize 5 Gbps for single-flow and up to 25 Gbps for multi-flow traffic. When launched in a placement group, select EC2 instances can utilize up to 10 Gbps for single-flow traffic.

Q. What is a Cluster GPU Instance?

Cluster GPU Instances provide general-purpose graphics processing units (GPUs) with proportionally high CPU and increased network performance for applications benefiting from highly parallelized processing that can be accelerated by GPUs using the CUDA and OpenCL programming models. Common applications include modeling and simulation, rendering and media processing.

Cluster GPU Instances give customers with HPC workloads an option beyond Cluster Compute Instances to further customize their high performance clusters in the cloud for applications that can benefit from the parallel computing power of GPUs.

Cluster GPU Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters – allowing applications to get the low-latency, high bandwidth network performance required for tightly-coupled node-to-node communication typical of many HPC applications.

Q. What is a High Memory Cluster Instance?

High Memory Cluster Instances provide customers with large amounts of memory and CPU capabilities per instance in addition to high network capabilities. These instance types are ideal for memory intensive workloads including in-memory analytics systems, graph analysis and many science and engineering applications.

High Memory Cluster Instances use the same cluster placement group functionality as Cluster Compute Instances for grouping instances into clusters – allowing applications to get the low-latency, high bandwidth network performance required for tightly-coupled node-to-node communication typical of many HPC and other network intensive applications.

Q. Does use of Cluster Compute and Cluster GPU Instances differ from other Amazon EC2 instance types?

Cluster Compute and Cluster GPU Instances use differs from other Amazon EC2 instance types in two ways.

First, Cluster Compute and Cluster GPU Instances use Hardware Virtual Machine (HVM) based virtualization and run only Amazon Machine Images (AMIs) based on HVM virtualization. Paravirtual Machine (PVM) based AMIs used with other Amazon EC2 instance types cannot be used with Cluster Compute or Cluster GPU Instances.

Second, in order to fully benefit from the available low latency, full bisection bandwidth between instances, Cluster Compute and Cluster GPU Instances must be launched into a cluster placement group through the Amazon EC2 API or AWS Management Console.

Q. What is a cluster placement group?

A cluster placement group is a logical entity that enables creating a cluster of instances by launching instances as part of a group. The cluster of instances then provides low latency connectivity between instances in the group. Cluster placement groups are created through the Amazon EC2 API or AWS Management Console.

Q. Are all features of Amazon EC2 available for Cluster Compute and Cluster GPU Instances?

Currently, Amazon DevPay is not available for Cluster Compute or Cluster GPU Instances.

Q. Is there a limit on the number of Cluster Compute or Cluster GPU Instances I can use and/or the size of cluster I can create by launching Cluster Compute Instances or Cluster GPU into a cluster placement group?

There is no limit specific for Cluster Compute Instances. For Cluster GPU Instances, you can launch 2 Instances on your own. If you need more capacity, please complete the Amazon EC2 instance request form (selecting the appropriate primary instance type).

Q. Are there any ways to optimize the likelihood that I receive the full number of instances I request for my cluster via a cluster placement group?

We recommend that you launch the minimum number of instances required to participate in a cluster in a single launch. For very large clusters, you should launch multiple placement groups, e.g. two placement groups of 128 instances, and combine them to create a larger, 256 instance cluster.

Q. Can Cluster GPU and Cluster Compute Instances be launched into a single cluster placement group?

While it may be possible to launch different cluster instance types into a single placement group, at this time we only support homogenous placement groups.

Q. If an instance in a cluster placement group is stopped then started again, will it maintain its presence in the cluster placement group?

Yes. A stopped instance will be started as part of the cluster placement group it was in when it stopped. If capacity is not available for it to start within its cluster placement group, the start will fail.

Hardware information

Q: What CPU options are available on EC2 instances?

EC2 instances offer a variety of CPU options to help customers balance performance and cost requirements. Depending on the instance type, EC2 offers a choice in CPU including AWS Graviton/Graviton2 processors (Arm), AMD processors (x86), and Intel processors (x86).

Q: What kind of hardware will my application stack run on?

Visit Amazon EC2 Instance Type for a list of EC2 instances available by region.

Q: How does EC2 perform maintenance?

AWS regularly performs routine hardware, software, power, and network maintenance with minimal disruption across all EC2 instance types. This is achieved by a combination of technologies and methods across the entire AWS Global infrastructure, such as live update and live migration as well as redundant and concurrently maintainable systems. Non-intrusive maintenance technologies such as live update and live migration do not require instances to be stopped or rebooted. Customers are not required to take any action prior to, during or after live migration or live update. These technologies help improve application uptime and reduce your operational effort. Amazon EC2 uses live update to deploy software to servers quickly with minimal impact to customer instances. Live update ensures that customers' workloads run on servers with software that is up-to-date with security patches, new instance features and performance improvements. Amazon EC2 uses live migration when running instances need to be moved from one server to another for hardware maintenance or to optimize placement of instances or to dynamically manage CPU resources. Amazon EC2 has been expanding the scope and coverage of non-intrusive maintenance technologies over the years so that scheduled maintenance events are a fallback option rather than the primary means of enabling routine maintenance.

Q: How do I select the right instance type?

Amazon EC2 instances are grouped into 5 families: General Purpose, Compute Optimized, Memory Optimized, Storage Optimized and Accelerated Computing instances. General Purpose Instances have memory to CPU ratios suitable for most general purpose applications and come with fixed performance or burstable performance; Compute Optimized instances have proportionally more CPU resources than memory (RAM) and are well suited for scale out compute-intensive applications and High Performance Computing (HPC) workloads; Memory Optimized Instances offer larger memory sizes for memory-intensive applications, including database and memory caching applications; Accelerated Computing instances use hardware accelerators, or co-processors, to perform functions such as floating point number calculations, graphics processing, or data pattern matching, more efficiently than is possible in software running on CPUs; Storage Optimized Instances provide low latency, I/O capacity using SSD-based local instance storage for I/O-intensive applications, as well as dense HDD-storage instances, which provide local high storage density and sequential I/O performance for data warehousing, Hadoop and other data-intensive applications. When choosing instance types, you should consider the characteristics of your application with regards to resource utilization (i.e. CPU, Memory, Storage) and select the optimal instance family and instance size.

Q: What is an "EC2 Compute Unit" and why did you introduce it?

Transitioning to a utility computing model fundamentally changes how developers have been trained to think about CPU resources. Instead of purchasing or leasing a particular processor to use for several months or years, you are renting capacity by the hour. Because Amazon EC2 is built on commodity hardware, over time there may be several different types of physical hardware underlying EC2 instances. Our goal is to provide a consistent amount of CPU capacity no matter what the actual underlying hardware.

Amazon EC2 uses a variety of measures to provide each instance with a consistent and predictable amount of CPU capacity. In order to make it easy for developers to compare CPU capacity between different instance types, we have defined an Amazon EC2 Compute Unit. The amount of CPU that is allocated to a particular instance is expressed in terms of these EC2 Compute Units. We use several benchmarks and tests to manage the consistency and predictability of the performance from an EC2 Compute Unit. The EC2 Compute Unit (ECU) provides the relative measure of the integer processing power of an Amazon EC2 instance. Over time, we may add or substitute measures that go into the definition of an EC2 Compute Unit, if we find metrics that will give you a clearer picture of compute capacity.

Q: How does EC2 ensure consistent performance of instance types over time?

AWS conducts yearly performance benchmarking of Linux and Windows compute performance on EC2 instance types. Benchmarking results, a test suite that customers can use to conduct independent testing, and guidance on expected performance variance is available under NDA for M,C,R, T and z1d instances; please contact your sales representative to request them.

Micro instances

Q. How much compute power do Micro instances provide?

Micro instances provide a small amount of consistent CPU resources and allow you to burst CPU capacity up to 2 ECUs when additional cycles are available. They are well suited for lower throughput applications and web sites that consume significant compute cycles periodically but very little CPU at other times for background processes, daemons, etc.

Q. How does a Micro instance compare in compute power to a Standard Small instance?

At steady state, Micro instances receive a fraction of the compute resources that Small instances do. Therefore, if your application has compute-intensive or steady state needs we recommend using a Small instance (or larger, depending on your needs). However, Micro instances can periodically burst up to 2 ECUs (for short periods of time). This is double the number of ECUs available from a Standard Small instance. Therefore, if you have a relatively low throughput application or web site with an occasional need to consume significant compute cycles, we recommend using Micro instances.

Q. How can I tell if an application needs more CPU resources than a Micro instance is providing?

The CloudWatch metric for CPU utilization will report 100% utilization if the instance bursts so much that it exceeds its available CPU resources during that CloudWatch monitored minute. CloudWatch reporting 100% CPU utilization is your signal that you should consider scaling – manually or via Auto Scaling – up to a larger instance type or scale out to multiple Micro instances.

Q. Are all features of Amazon EC2 available for Micro instances?

Currently Amazon DevPay is not available for Micro instances.

Nitro Hypervisor

Q. What is the Nitro Hypervisor?

The launch of C5 instances introduced a new hypervisor for Amazon EC2, the Nitro Hypervisor. As a component of the Nitro system, the Nitro Hypervisor primarily provides CPU and memory

isolation for EC2 instances. VPC networking and EBS storage resources are implemented by dedicated hardware components, Nitro Cards that are part of all current generation EC2 instance families. The Nitro Hypervisor is built on core Linux Kernel-based Virtual Machine (KVM) technology, but does not include general-purpose operating system components.

Q. How does the Nitro Hypervisor benefit customers?

The Nitro Hypervisor provides consistent performance and increased compute and memory resources for EC2 virtualized instances by removing host system software components. It allows AWS to offer larger instance sizes (like c5.18xlarge) that provide practically all of the resources from the server to customers. Previously, C3 and C4 instances each eliminated software components by moving VPC and EBS functionality to hardware designed and built by AWS. This hardware enables the Nitro Hypervisor to be very small and uninvolved in data processing tasks for networking and storage.

Q. Will all EC2 instances use the Nitro Hypervisor?

Eventually all new instance types will use the Nitro Hypervisor, but in the near term, some new instance types will use Xen depending on the requirements of the platform.

Q. Will AWS continue to invest in its Xen-based hypervisor?

Yes. As AWS expands its global cloud infrastructure, EC2's use of its Xen-based hypervisor will also continue to grow. Xen will remain a core component of EC2 instances for the foreseeable future. AWS is a founding member of the Xen Project since its establishment as a Linux Foundation Collaborative Project and remains an active participant on its Advisory Board. As AWS expands its global cloud infrastructure, EC2's Xen-based hypervisor also continues to grow. Therefore EC2's investment in Xen continues to grow, not shrink.

Q. How many EBS volumes and Elastic Network Interfaces (ENIs) can be attached to instances running on the Nitro Hypervisor?

Instances running on the Nitro Hypervisor support a maximum of 27 additional PCI devices for EBS volumes and VPC ENIs. Each EBS volume or VPC ENI uses a PCI device. For example, if you attach 3 additional network interfaces to an instance that uses the Nitro Hypervisor, you can attach up to 24 EBS volumes to that instance.

Q. Will the Nitro Hypervisor change the APIs used to interact with EC2 instances?

No, all the public facing APIs for interacting with EC2 instances that run using the Nitro Hypervisor will remain the same. For example, the "hypervisor" field of the DescribeInstances response will continue to report "xen" for all EC2 instances, even those running under the Nitro Hypervisor. This field may be removed in a future revision of the EC2 API.

Q. Which AMIs are supported on instances that use the Nitro Hypervisor?

EBS backed HVM AMIs with support for ENA networking and booting from NVMe storage can be used with instances that run under the Nitro Hypervisor. The latest Amazon Linux AMI and Windows AMIs provided by Amazon are supported, as are the latest AMI of Ubuntu, Debian, Red Hat Enterprise Linux, SUSE Enterprise Linux, CentOS, and FreeBSD.

Q. Will I notice any difference between instances using Xen hypervisor and those using the Nitro Hypervisor?

Yes. For example, instances running under the Nitro Hypervisor boot from EBS volumes using an NVMe interface. Instances running under Xen boot from an emulated IDE hard drive, and switch to the Xen paravirtualized block device drivers.

Operating systems can identify when they are running under a hypervisor. Some software assumes that EC2 instances will run under the Xen hypervisor and rely on this detection. Operating systems will detect they are running under KVM when an instance uses the Nitro Hypervisor, so the process to identify EC2 instances should be used to identify EC2 instances that run under both hypervisors.

All the features of EC2 such as Instance Metadata Service work the same way on instances running under both Xen and the Nitro Hypervisor. The majority of applications will function the same way under both Xen and the Nitro Hypervisor as long as the operating system has the needed support for ENA networking and NVMe storage.

Q. How are instance reboot and termination EC2 API requests implemented by the Nitro Hypervisor?

The Nitro Hypervisor signals the operating system running in the instance that it should shut down cleanly by industry standard ACPI methods. For Linux instances, this requires that acpid be installed and functioning correctly. If acpid is not functioning in the instance, termination events will be delayed by multiple minutes and will then execute as a hard reset or power off.

Q. How do EBS volumes behave when accessed by NVMe interfaces?

There are some important differences in how operating system NVMe drivers behave compared to Xen paravirtual (PV) block drivers.

First, the NVMe device names used by Linux based operating systems will be different than the parameters for EBS volume attachment requests and block device mapping entries such as /dev/xvda and /dev/xvdf. NVMe devices are enumerated by the operating system as /dev/nvme0n1, /dev/nvme1n1, and so on. The NVMe device names are not persistent mappings to volumes, therefore other methods like file system UUIDs or labels should be used when configuring the automatic mounting of file systems or other startup activities. When EBS volumes are accessed via the NVMe interface, the EBS volume ID is available via the controller serial number and the device name specified in EC2 API requests is provided by an NVMe vendor extension to the Identify Controller command. This enables backward compatible symbolic links to be created by a utility script. For more information see the EC2 documentation on device naming and NVMe based EBS volumes.

Second, by default the NVMe drivers included in most operating systems implement an I/O timeout. If an I/O does not complete in an implementation specific amount of time, usually tens of seconds, the driver will attempt to cancel the I/O, retry it, or return an error to the component that issued the I/O. The Xen PV block device interface does not time out I/O, which can result in processes that cannot be terminated if it is waiting for I/O. The Linux NVMe driver behavior can be modified by specifying a higher value for the nvme.io timeout kernel module parameter.

Third, the NVMe interface can transfer much larger amounts of data per I/O, and in some cases may be able to support more outstanding I/O requests, compared to the Xen PV block interface. This can cause higher I/O latency if very large I/Os or a large number of I/O requests are issued to volumes designed to support throughput workloads like EBS Throughput Optimized HDD (st1) and Cold HDD (sc1) volumes. This I/O latency is normal for throughput optimized volumes in these scenarios, but may cause I/O timeouts in NVMe drivers. The I/O timeout can be adjusted in the Linux driver by specifying a larger value for the nvme_core.io_timeout kernel module parameter.

Optimize CPUs

Q: What is Optimize CPUs?

Optimize CPUs gives you greater control of your EC2 instances on two fronts. First, you can specify a custom number of vCPUs when launching new instances to save on vCPU-based licensing costs. Second, you can disable Intel Hyper-Threading Technology (Intel HT Technology) for workloads that perform well with single-threaded CPUs, such as certain high-performance computing (HPC) applications.

Q: Why should I use Optimize CPUs feature?

You should use Optimize CPUs if:

- You are running EC2 workloads that are not compute bound and are incurring vCPU-based licensing costs. By launching instances with a custom number of vCPUs you may be able to optimize your licensing spend.
- You are running workloads that will benefit from disabling hyper-threading on EC2 instances.

Q: How will the CPU optimized instances be priced?

CPU optimized instances will be priced the same as equivalent full-sized instances.

Q: How will my application performance change when using Optimize CPUs on EC2?

Your application performance change with Optimize CPUs will be largely dependent on the workloads you are running on EC2. We encourage you to benchmark your application performance with Optimize CPUs to arrive at the right number of vCPUs and optimal hyper-threading behavior for your application.

Q: Can I use Optimize CPUs on EC2 Bare Metal instance types (such as i3.metal)?

No. You can use Optimize CPUs with only virtualized EC2 instances.

Workloads

Amazon EC2 running IBM

Q. How am I billed for my use of Amazon EC2 running IBM?

You pay only for what you use and there is no minimum fee. Pricing is per instance-hour consumed for each instance type. Partial instance-hours consumed are billed as full hours. Data transfer for Amazon EC2 running IBM is billed and tiered separately from Amazon EC2. There is no Data Transfer charge between two Amazon Web Services within the same region (i.e. between Amazon EC2 US West and another AWS service in the US West). Data transferred between AWS services in different regions will be charged as Internet Data Transfer on both sides of the transfer.

Q. Can I use Amazon DevPay with Amazon EC2 running IBM?

No, you cannot use DevPay to bundle products on top of Amazon EC2 running IBM at this time.

Amazon EC2 running Microsoft Windows and other third-party software

Q. Can I use my existing Windows Server license with EC2?

Yes you can. After you've imported your own Windows Server machine images using the ImportImage tool, you can launch instances from these machine images on EC2 Dedicated Hosts and effectively manage instances and report usage. Microsoft typically requires that you track usage of your licenses against physical resources such as sockets and cores and Dedicated Hosts helps you to do this. Visit the Dedicated Hosts detail page for more information on how to use your own Windows Server licenses on Amazon EC2 Dedicated Hosts.

Q. What software licenses can I bring to the Windows environment?

Specific software license terms vary from vendor to vendor. Therefore, we recommend that you check the licensing terms of your software vendor to determine if your existing licenses are authorized for use in Amazon EC2.

macOS workloads

Q: What are Amazon EC2 Mac instances?

Amazon EC2 Mac instances allow customers to run on-demand macOS workloads in the cloud for the first time, extending the flexibility, scalability, and cost benefits of AWS to all Apple developers. With EC2 Mac instances, developers creating apps for iPhone, iPad, Mac, Apple Watch, Apple TV, and Safari can provision and access macOS environments within minutes, dynamically scale capacity as needed, and benefit from AWS's pay-as-you-go pricing.

Q: What workloads should you run on EC2 Mac instances?

Amazon EC2 Mac instances are designed to build, test, sign, and publish applications for Apple platforms such as iOS, iPadOS, watchOS, tvOS, macOS, and Safari. Customers such as Pinterest, Intuit, FlipBoard, Twitch, and Goldman Sachs have seen up to 75% better build performance, up to 80% lower build failure rates, and up to 5x the number of parallel builds compared to running macOS on premises.

Q: What are EC2 x86 Mac instances?

x86-based EC2 Mac instances are built on Apple Mac mini computers featuring Intel Core i7 processors and are powered by the AWS Nitro System. They offer customers a choice of macOS Mojave (10.14), macOS Catalina (10.15), macOS Big Sur (11), and macOS Monterey (12) as Amazon Machine Images (AMIs). x86-based EC2 Instances are available in 12 Regions: US East (Ohio, N. Virginia), US West (Oregon), Europe (Amsterdam, Frankfurt, Ireland, London), and Asia Pacific (Mumbai, Seoul, Singapore, Sydney, Tokyo). Learn more and get started with x86-based EC2 Mac instances here.

Q: What are EC2 M1 Mac instances?

EC2 M1 Mac instances are built on Apple M1 Mac mini computers and are powered by the AWS Nitro System. They deliver up to 60 percent better price performance over x86-based EC2 Mac instances for iOS and macOS application build workloads. EC2 M1 Mac instances enable ARM64 macOS environments for the first time in AWS, and support macOS Big Sur (11) and macOS Monterey (12) as Amazon Machine Images (AMIs). EC2 M1 Mac instances are available in 4 Regions: US East (N. Virginia), US West (Oregon), Europe (Ireland), and Asia Pacific (Singapore). Learn more and get started with EC2 M1 Mac instances here.
Q: What pricing models are available for EC2 Mac instances?

Amazon EC2 Mac instances are available as Dedicated Hosts through both On-Demand and Savings Plans pricing models. The Dedicated Host is the unit of billing for EC2 Mac instances. Billing is per second, with a 24-hour minimum allocation period for the Dedicated Host to comply with the Apple macOS Software License Agreement. At the end of the 24-hour minimum allocation period, the host can be released at any time with no further commitment. Both Compute and Instance Savings Plans are available for EC2 Mac instances and offer up to 44 percent off On-Demand pricing. Visit the Dedicated Host pricing page for more information. (Note: Please select “Dedicated Host” tenancy and “Linux” operating system to view details.) You can also access EC2 Mac instances pricing on the AWS Pricing Calculator for Dedicated Hosts.

Q: How do you release a Dedicated Host?

The minimum allocation period for an EC2 Mac instance Dedicated Host is 24 hours. After the allocation period has exceeded 24 hours, first stop or terminate the instance running on the host, then release the host using the aws ec2 release-hosts CLI command or the AWS Management Console.

Q: Can you share EC2 Mac Dedicated Hosts with other AWS accounts in your organization?

Yes. You can share EC2 Mac Dedicated Hosts with AWS accounts inside your AWS organization, an organizational unit inside your AWS organization, or your entire AWS organization via AWS Resource Access Manager. For more information, please refer to the AWS Resource Access Manager documentation.

Q: How many EC2 Mac instances can you run on an EC2 Mac Dedicated Host?

EC2 Mac instances leverage the full power of the underlying Mac mini hardware. You can run 1 EC2 Mac instance on each EC2 Mac Dedicated Host.
Q: Can you update the EFI NVRAM variables on an EC2 Mac instance?

Yes, you can update certain EFI NVRAM variables on an EC2 Mac instance that will persist across reboots. However, EFI NVRAM variables will be reset if the instance is stopped or terminated. Please see the EC2 Mac instances documentation for more information.

Q: Can you use FileVault to encrypt the Amazon Elastic Block Store (Amazon EBS) boot volume on EC2 Mac instances?

FileVault requires a login before booting into macOS and before remote access can be enabled. If FileVault is enabled, you will lose access to your data on the boot volume at instance reboot, stop, or terminate. We strongly recommend you do not enable FileVault. Instead, we recommend using Amazon EBS encryption for both boot and data EBS volumes on EC2 Mac instances.

Q: Can you access to the microphone input or audio output on an EC2 Mac instance?

There is no access to the microphone input on an EC2 Mac instance. The built-in Apple Remote Desktop VNC server does not support audio output. Third party remote desktop software, such as Teradici CAS, supports remote audio on macOS.
Q: What macOS-based Amazon Machine Images (AMIs) are available for EC2 Mac instances?

EC2 Mac instances use physical Mac mini hardware to run macOS. Apple hardware only supports the macOS version shipped with the hardware (or later). x86-based EC2 Mac instances use the 2018 Intel Core i7 Mac mini, which means macOS Mojave (10.14.x) is as 'far back' as you can go, since the 2018 Mac mini shipped with Mojave. EC2 M1 Mac instances use

2020 M1 Mac mini, which shipped with macOS Big Sur (11.x). To see which latest versions of macOS are available as EC2 Mac AMIs, please visit the documentation.

Q: How can you run older versions of macOS on EC2 Mac instances?

EC2 Mac instances are bare metal instances and do not use the Nitro hypervisor. You can install and run a type-2 virtualization layer on x86-based EC2 Mac instances to get access to macOS High Sierra, Sierra, or older macOS versions. On EC2 M1 Mac instances, as macOS Big Sur is the first macOS version to support Apple Silicon, older macOS versions will not run even under virtualization.

Q: How can I run beta or preview versions of macOS on EC2 Mac instances?

Beta and preview versions of macOS are currently not supported on EC2 Mac instances.

Q: How can you use EC2 user data with EC2 Mac instances?

As with EC2 Linux and Windows instances, you can pass custom user data to EC2 Mac instances. Instead of using cloud-init, EC2 Mac instances use an open-source launch daemon: ec2-macos-init. You can pass this data into the EC2 Launch Wizard as plain-text, as a file, or as base64-encoded-text.

Q: How do you install Xcode on an EC2 M1 Mac instance?

AWS provides base macOS AMIs without any prior Xcode IDE installation. You can install Xcode (and accept the EULA) just like you would on any other macOS system. You can install the latest Xcode IDE from the App Store, or earlier Xcode versions from the Apple Developer website. Once you have Xcode installed, we recommend creating a snapshot of your AMI for future use.

Q: What is the release cadence of macOS AMIs?

We make new macOS AMIs available on a best effort basis. You can subscribe to SNS notifications for updates. We are targeting 30-60 days after a macOS minor version update and 90-120 days after a macOS major version update to release official macOS AMIs.

Q: What agents and packages are included in EC2 macOS AMIs?

The following agents and packages are included by default in EC2 macOS AMIs:

- ENA Driver for macOS
- AWS CLI
- EC2-macos-init
- Amazon CloudWatch Agent
- Chrony
- Homebrew
- AWS Systems Manger Agent

Q: Can you update the agents and packages included in macOS AMIs?

There is a public Github repository of the Homebrew tap for all agents and packages added to the base macOS image. You can use Homebrew to install the latest versions of agents and packages on macOS instances.

Q: Can you apply OS and software updates to your Mac instances directly from Apple Update Servers?

Automatic macOS software updates are disabled on EC2 Mac instances. We recommend using our officially vended macOS AMIs to launch the version of macOS you need. On x86-based EC2 Mac instances, you can update the version of macOS via the Software Update preferences pane, or via the software update CLI command. We do not support macOS updates on EC2 M1 Mac instances at this time. On both EC2 Mac instances, you can install and update applications and any other user-space software.
Q: How do you connect to an EC2 Mac instance over SSH?

After launching your instance and receiving an instance id, you can use the following command to poll the instance and determine when it is ready for SSH access. Connecting over SSH to EC2 Mac instances follows the same process as connecting to other EC2 instances, such as those running Linux or Windows. To support connecting to your instance using SSH, launch the instance using a key pair and a security group that allows SSH access. Provide the .pem file for the key pair when you connect to the instance.

Q: How do you connect to an EC2 Mac instance over VNC?

macOS has built-in Screen Sharing functionality that is disabled by default, but can be enabled and used to connect to a Graphical (Desktop) session of your EC2 Mac instance..

Q: How do you connect to an EC2 Mac instance using AWS Systems Manager Session Manager?

You can connect to your EC2 Mac instances with AWS Systems Manager Session Manager (SSM). Session Manager is a fully managed AWS Systems Manager feature that provides secure and auditable instance management. It removes the need to keep open inbound ports, maintain bastion hosts, or manage SSH keys. The SSM Agent is pre-installed by default on all EC2 macOS AMIs. For more information, please see this blog.

Q: How many Amazon Elastic Block Store (Amazon EBS) volumes and Elastic Network Interfaces (ENIs) are supported by EC2 Mac instances?

x86-based EC2 Mac instances support 16 EBS volumes and 8 ENI attachments, and EC2 M1 Mac instances support up to 10 EBS volumes and 8 ENI attachments.

Q: Do EC2 Mac instances support EBS?

EC2 Mac instances are EBS optimized by default and offer up to 8 Gbps of dedicated EBS bandwidth to both encrypted and unencrypted EBS volumes.

Q: Do EC2 Mac instances support booting from local storage?

EC2 Mac instances can only boot from EBS-backed macOS AMIs. The internal SSD of the Mac mini is present in Disk Utility, but is not bootable.

Q: Do EC2 Mac instances support Amazon FSx?

Yes. EC2 Mac instances support FSx using the SMB protocol. You will need to enroll the EC2 Mac instance into a supported directory service (such as Active Directory or the AWS Directory Service) to enable FSx on EC2 Mac instances.

Q: Do EC2 Mac instances support Amazon Elastic File System (Amazon EFS)?

Yes, EC2 Mac instances support EFS over the NFSv4 protocol. For more information on EFS, visit the product page.

Nitro System Support for Previous Generation

Q. What is Nitro System Support for Older Generation instances?

The AWS Nitro System now will provide its modern hardware and software components for previous generation EC2 instances to extend the length of service beyond the typical lifetime of the underlying hardware. With the Nitro System support customers can continue running their workloads and applications on the instance families they were built on.

Q. Which previous generation instances will receive Nitro System support and within what time frame?

Starting 2022 the following previous generation instances will receive Nitro System support: C1, C3, I2, M1, M2, M3, R3, and T1. Customers of these instances will receive a maintenance notification of migration to Nitro System.

Q. What actions do I need to take to migrate my existing previous generation instances?

Customers do not need to take any action for migrating active previous generation instances running on older generation hardware. For instances that are on older generation hardware, each customer account ID mapped to instance(s) will receive an email notification 2 weeks prior to the scheduled maintenance.

Similar to our typical maintenance events, customers will have the option to reschedule their maintenance as many times as needed within 2 additional weeks from the original scheduled maintenance time.

Q. What will happen to my instance during this maintenance event?

We will work in conjunction with the customer as a part of our standard AWS maintenance process. Several AWS teams have already migrated and are running previous generation instances on Nitro hardware. During maintenance, the instance will be rebooted which can take up to 30 minutes depending upon the instance size and attributes. For example: Instances with local disk take longer to migrate than instances without local disk. After the reboot, your instance retains its IP address, DNS name, and any data on local instance-store volumes.

Q. Do I need to rebuild/recertify workloads to run on previous generation instances migrated to AWS Nitro System?

No, customers don't need to rebuild/recertify workloads on previous generation instances migrated to AWS Nitro System.

Q. Will there be any changes to my instance specifications once migrated to AWS Nitro System?

There will be no change to instance specification of previous generation instances when instances are migrated to AWS Nitro System.

Q. Will all features and AMIs on my previous generation instances be supported as a part of this migration?

Yes, all existing features and AMIs supported on previous generation instances will be supported as we migrate these instances to AWS Nitro System. However, please note that

classic networking that has been announced for retirement and will not be supported on previous generation instances running on the Nitro System. We will migrate previous generation of instances running classic networking only after customers have moved to VPC.

Q. Will there be changes to pricing and billing when previous generation instances are migrated to AWS Nitro System?

There will be no change to billing and pricing. We will continue to support the same pricing models we support today for the previous generation instances (on-demand, 1yr/3yr Reserved Instance, Savings Plan, Spot).

Elastic Load Balancing

Distribute network traffic to improve application scalability.

How it works

Elastic Load Balancing (ELB) automatically distributes incoming application traffic across multiple targets and virtual appliances in one or more Availability Zones (AZs).

Application Load Balancer

Gateway Load Balancer

Network Load Balancer

Cross-zone load balancing

The nodes for your load balancer distribute requests from clients to registered targets. When cross-zone load balancing is enabled, each load balancer node distributes traffic across the registered targets in all enabled Availability Zones. When cross-zone load balancing is disabled, each load balancer node distributes traffic only across the registered targets in its Availability Zone.

The following diagrams demonstrate the effect of cross-zone load balancing with round robin as the default routing algorithm. There are two enabled Availability Zones, with two targets in Availability Zone A and eight targets in Availability Zone B. Clients send requests, and Amazon Route 53 responds to each request with the

IP address of one of the load balancer nodes. Based on the round robin routing algorithm, traffic is distributed such that each load balancer node receives 50% of the traffic from the clients. Each load balancer node distributes its share of the traffic across the registered targets in its scope.

If cross-zone load balancing is enabled, each of the 10 targets receives 10% of the traffic. This is because each load balancer node can route its 50% of the client traffic to all 10 targets.

If cross-zone load balancing is disabled:

- Each of the two targets in Availability Zone A receives 25% of the traffic.
- Each of the eight targets in Availability Zone B receives 6.25% of the traffic.

This is because each load balancer node can route its 50% of the client traffic only to targets in its Availability Zone.

Use cases

Modernize applications with serverless and containers
Scale modern applications to meet demand without complex configurations or API gateways.

Improve hybrid cloud network scalability
Load balance across AWS and on-premises resources using a single load balancer.

Retain your existing network appliances
Deploy network appliances from your preferred vendor while taking advantage of the scale and flexibility of the cloud.

What is Elastic Load Balancing?

Elastic Load Balancing automatically distributes your incoming traffic across multiple targets, such as EC2 instances, containers, and IP addresses, in one or more Availability Zones. It monitors the health of its registered targets, and routes traffic only to the healthy targets. Elastic Load Balancing scales your load balancer capacity automatically in response to changes in incoming traffic.

Load balancer benefits

A load balancer distributes workloads across multiple compute resources, such as virtual servers. Using a load balancer increases the availability and fault tolerance of your applications.

You can add and remove compute resources from your load balancer as your needs change, without disrupting the overall flow of requests to your applications.

You can configure health checks, which monitor the health of the compute resources, so that the load balancer sends requests only to the healthy ones. You can also offload the work of encryption and decryption to your load balancer so that your compute resources can focus on their main work.

Features of Elastic Load Balancing

Elastic Load Balancing supports the following load balancers: Application Load Balancers, Network Load Balancers, Gateway Load Balancers, and Classic Load Balancers. You can select the type of load balancer that best suits your needs. For more information, see Product comparisons.

Accessing Elastic Load Balancing

You can create, access, and manage your load balancers using any of the following interfaces:

- **AWS Management Console**— Provides a web interface that you can use to access Elastic Load Balancing.
- **AWS Command Line Interface (AWS CLI)** — Provides commands for a broad set of AWS services, including Elastic Load Balancing. The AWS CLI is supported on Windows, macOS, and Linux. For more information, see AWS Command Line Interface.
- **AWS SDKs** — Provide language-specific APIs and take care of many of the connection details, such as calculating signatures, handling request retries, and error handling. For more information, see AWS SDKs.
- **Query API**— Provides low-level API actions that you call using HTTPS requests. Using the Query API is the most direct way to access Elastic Load Balancing. However, the Query API requires that your application handle low-level details such as generating the hash to sign the request, and error handling. For more information, see the following:

- Application Load Balancers and Network Load Balancers — API version 2015-12-01
- Classic Load Balancers — API version 2012-06-01

Related services

Elastic Load Balancing works with the following services to improve the availability and scalability of your applications.

- **Amazon EC2** — Virtual servers that run your applications in the cloud. You can configure your load balancer to route traffic to your EC2 instances. For more information, see the Amazon EC2 User Guide for Linux Instances or the Amazon EC2 User Guide for Windows Instances.
- **Amazon EC2 Auto Scaling** — Ensures that you are running your desired number of instances, even if an instance fails. Amazon EC2 Auto Scaling also enables you to automatically increase or decrease the number of instances as the demand on your instances changes. If you enable Auto Scaling with Elastic Load Balancing, instances that are launched by Auto Scaling are automatically registered with the load balancer. Likewise, instances that are terminated by Auto Scaling are automatically de-registered from the load balancer. For more information, see the Amazon EC2 Auto Scaling User Guide.
- **AWS Certificate Manager** — When you create an HTTPS listener, you can specify certificates provided by ACM. The load balancer uses certificates to terminate connections and decrypt requests from clients.
- **Amazon CloudWatch** — Enables you to monitor your load balancer and to take action as needed. For more information, see the Amazon CloudWatch User Guide.
- **Amazon ECS** — Enables you to run, stop, and manage Docker containers on a cluster of EC2 instances. You can configure your load balancer to route traffic to your containers. For more information, see the Amazon Elastic Container Service Developer Guide.
- **AWS Global Accelerator** — Improves the availability and performance of your application. Use an accelerator to distribute traffic across multiple load balancers in one or more AWS Regions. For more information, see the AWS Global Accelerator Developer Guide.
- **Route 53** — Provides a reliable and cost-effective way to route visitors to websites by translating domain names into the numeric IP addresses that computers use to connect to each other. For example, it would

translate www.example.com into the numeric IP address 192.0.2.1. AWS assigns URLs to your resources, such as load balancers. However, you might want a URL that is easy for users to remember. For example, you can map your domain name to a load balancer. For more information, see the Amazon Route 53 Developer Guide.

- **AWS WAF** — You can use AWS WAF with your Application Load Balancer to allow or block requests based on the rules in a web access control list (web ACL). For more information, see the AWS WAF Developer Guide.

Elastic Load Balancing FAQs

General

Q: How do I decide which load balancer to select for my application?

A: Elastic Load Balancing (ELB) supports four types of load balancers. You can select the appropriate load balancer based on your application needs. If you need to load balance HTTP requests, we recommend you use the Application Load Balancer (ALB). For network/transport protocols (layer4 – TCP, UDP) load balancing, and for extreme performance/low latency applications we recommend using Network Load Balancer. If your application is built within the Amazon Elastic Compute Cloud (Amazon EC2) Classic network, you should use Classic Load Balancer. If you need to deploy and run third-party virtual appliances, you can use Gateway Load Balancer.

Q: Can I privately access Elastic Load Balancing APIs from my Amazon Virtual Private Cloud (VPC) without using public IPs?

A: Yes, you can privately access Elastic Load Balancing APIs from your Amazon Virtual Private Cloud (VPC) by creating VPC endpoints. With VPC endpoints, the routing between the VPC and Elastic Load Balancing APIs is handled by the AWS network without the need for an Internet gateway, network address translation (NAT) gateway, or virtual private network (VPN) connection. The latest generation of VPC Endpoints used by Elastic Load Balancing are powered by AWS PrivateLink, an AWS technology enabling the private connectivity between AWS services using Elastic Network Interfaces (ENI) with private IPs in your VPCs. To learn more about AWS PrivateLink, visit the AWS PrivateLink documentation.

Q: Is there an SLA for load balancers?

A: Yes, Elastic Load Balancing guarantees a monthly availability of at least 99.99% for your load balancers (Classic, Application or Network). To learn more about the SLA and know if you are qualified for a credit, visit here.

Application Load Balancer

Q: Which operating systems does an Application Load Balancer support?

A: An Application Load Balancer supports targets with any operating system currently supported by the Amazon EC2 service.

Q: Which protocols does an Application Load Balancer support?

A: An Application Load Balancer supports load balancing of applications using HTTP and HTTPS (Secure HTTP) protocols.

Q: Is HTTP/2 Supported on an Application Load Balancer?

A: Yes. HTTP/2 support is enabled natively on an Application Load Balancer. Clients supporting HTTP/2 can connect to an Application Load Balancer over TLS.

Q: How can I use static IP or PrivateLink on my Application Load Balancer?

A: You can forward traffic from your Network Load Balancer, which provides support for PrivateLink and a static IP address per Availability Zone, to your Application Load Balancer. Create an Application Load Balancer-type target group, register your Application Load Balancer to it, and configure your Network Load Balancer to forward traffic to the Application Load Balancer-type target group.

Q: What TCP ports can I use to load balance?

A: You can perform load balancing for the following TCP ports: 1-65535

Q: Is WebSockets supported on an Application Load Balancer?

A: Yes. WebSockets and Secure WebSockets support is available natively and ready for use on an Application Load Balancer.

Q: Is Request tracing supported on an Application Load Balancer?

A: Yes. Request tracing is enabled by default on your Application Load Balancer.

Q: Does a Classic Load Balancer have the same features and benefits as an Application Load Balancer?

A: While there is some overlap, there is no feature parity between the two types of load balancers. Application Load Balancers are the foundation of our application layer load-balancing platform for the future.

Q: Can I configure my Amazon EC2 instances to accept traffic only from my Application Load Balancers?

A: Yes.

Q: Can I configure a security group for the front end of an Application Load Balancer?

A: Yes.

Q: Can I use the existing APIs that I use with my Classic Load Balancer with an Application Load Balancer?

A: No. Application Load Balancers require a new set of application programming interfaces (APIs).

Q: How do I manage both Application and Classic Load Balancers simultaneously?

A: The ELB Console will allow you to manage Application and Classic Load Balancers from the same interface. If you are using the command-line interface (CLI) or a software development kit (SDK), you will use a different ‘service’ for Application Load Balancers. For example, in the CLI

you will describe your Classic Load Balancers using `aws elb describe-load-balancers` and your Application Load Balancers using `aws elbv2 describe-load-balancers`.

Q: Can I convert my Classic Load Balancer to an Application Load Balancer (and vice-versa)?

A: No, you cannot convert one load balancer type into another.

Q: Can I migrate to Application Load Balancer from Classic Load Balancer?

A: Yes. You can migrate to Application Load Balancer from Classic Load Balancer using one of the options listed in this document.

Q: Can I use an Application Load Balancer as a Layer-4 load balancer?

A: No. If you need Layer-4 features, you should use Network Load Balancer.

Q: Can I use a single Application Load Balancer for handling HTTP and HTTPS requests?

A: Yes, you can add listeners for HTTP port 80 and HTTPS port 443 to a single Application Load Balancer.

Q: Can I get a history of Application Load Balancing API calls made on my account for security analysis and operational troubleshooting purposes?

A: Yes. To receive a history of Application Load Balancing API calls made on your account, use AWS CloudTrail.

Q: Does an Application Load Balancer support HTTPS termination?

A: Yes, you can terminate HTTPS connection on the Application Load Balancer. You must install a Secure Sockets Layer (SSL) certificate on your load balancer. The load balancer uses this certificate to terminate the connection and then decrypt requests from clients before sending them to targets.

Q: What are the steps to get a SSL certificate?

A: You can either use AWS Certificate Manager to provision an SSL/TLS certificate or you can obtain the certificate from other sources by creating the certificate request, getting the certificate request signed by a CA, and then uploading the certificate either using AWS Certification Manager or the AWS Identity and Access Management (IAM) service.

Q: How does an Application Load Balancer integrate with AWS Certificate Manager (ACM)?

A: An Application Load Balancer is integrated with AWS Certificate Management (ACM). Integration with ACM simplifies binding a certificate to the load balancer, thereby streamlining the entire SSL offload process. Purchasing, uploading, and renewing SSL/TLS certificates is a complex, manual, and time-consuming process. With ACM integration with Application Load Balancer, this whole process has been shortened to simply requesting a trusted SSL/TLS certificate and selecting the ACM certificate to provision it with the load balancer.

Q: Is back-end server authentication supported with an Application Load Balancer?

A: No, only encryption is supported to the back-ends with an Application Load Balancer.

Q: How can I enable Server Name Indication (SNI) for my Application Load Balancer?

A: SNI is automatically enabled when you associate more than one TLS certificate with the same secure listener on a load balancer. Similarly, SNI mode for a secure listener is automatically disabled when you have only one certificate associated to a secure listener.

Q: Can I associate multiple certificates for the same domain to a secure listener?

A: Yes, you can associate multiple certificates for the same domain to a secure listener. For example, you can associate:

- ECDSA and RSA certificates
- Certificates with different key sizes (e.g. 2K and 4K) for SSL/TLS certificates
- Single-Domain, Multi-Domain (SAN) and Wildcard certificates

Q: Is IPv6 supported with an Application Load Balancer?

A: Yes, IPv6 is supported with an Application Load Balancer.

Q: How do you set up rules on an Application Load Balancer?

A: You can configure rules for each of the listeners on the load balancer. The rules include conditions and corresponding actions if the conditions are satisfied. The supported conditions are Host header, path, HTTP headers, methods, query parameters, and source IP classless inter-domain routing (CIDR). The supported actions are redirect, fixed response, authenticate, and forward. Once you have set this up, the load balancer will use the rules to determine how a particular HTTP request should be routed. You can use multiple conditions and actions in a rule, and in each condition can specify a match on multiple values.

Q: Are there limits on the resources for an Application Load Balancer?

A: Your AWS account has these limits for an Application Load Balancer.

Q: How can I protect my web applications behind a load balancer from web attacks?

A: You can integrate your Application Load Balancer with AWS Web Application Firewall (WAF), a web application firewall that helps protect web applications from attacks by allowing you to configure rules based on IP addresses, HTTP headers, and custom uniform resource identifier (URI) strings. Using these rules, AWS WAF can block, allow, or monitor (count) web requests for your web application. Please see AWS WAF developer guide for more information.

Q: Can I load balance to any arbitrary IP address?

A: You can use any IP address from the load balancer's VPC CIDR for targets within load balancer's VPC, and any IP address from RFC 1918 ranges (10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16) or RFC 6598 range (100.64.0.0/10) for targets located outside the load balancer's VPC (for example, targets in Peered VPC, Amazon EC2 Classic, and on-premises locations reachable over AWS Direct Connect or VPN connection).

Q: How can I load balance applications distributed across a VPC and on-premises location?

A: There are various ways to achieve hybrid load balancing. If an application runs on targets distributed between a VPC and an on-premises location, you can add them to the same target group using their IP addresses. To migrate to AWS without impacting your application, gradually add VPC targets to the target group and remove on-premises targets from the target group.

If you have two different applications such that the targets for one application are in a VPC and the targets for other applications are in on-premises location, you can put the VPC targets in one target group and the on-premises targets in another target group and use content based routing to route traffic to each target group. You can also use separate load balancers for VPC and on-premises targets and use DNS weighting to achieve weighted load balancing between VPC and on-premises targets.

Q: How can I load balance to EC2-Classic instances?

A: You cannot load balance to EC2-Classic Instances when registering their Instance IDs as targets. However if you link these EC2-Classic instances to the load balancer's VPC using ClassicLink and use the private IPs of these EC2-Classic instances as targets, then you can load balance to the EC2-Classic instances. If you are using EC2 Classic instances today with a Classic Load Balancer, you can easily migrate to an Application Load Balancer.

Q: How do I enable cross-zone load balancing in Application Load Balancer?

A: Cross-zone load balancing is already enabled by default in Application Load Balancer.

Q: When should I authenticate users using the Application Load Balancer's integration with Amazon Cognito vs. the Application Load Balancers' native support for OpenID Connect (IODC) identity providers (IdPs)?

A: You should use authentication through Amazon Cognito if:

- You want to provide flexibility to your users to authenticate via social network identities (Google, Facebook, and Amazon) or enterprise identities (SAML) or via your own user directories provided by Amazon Cognito's User Pool.
- You are managing multiple identity providers including OpenID Connect and want to create a single authentication rule in Application Load Balancer (ALB) that can use Amazon Cognito to federate your multiple identity providers.
- You need to actively manage user profiles with one or more social or OpenID Connect identity providers from one central place. For example, you can put users in groups and add custom attributes to represent user status and control access for paid users.

Alternatively, if you have invested in developing custom IdP solutions and simply want to authenticate with a single OpenID Connect-compatible identity provider, you may prefer using Application Load Balancer's native OIDC solution.

Q: What type of redirects does Application Load Balancer support?

A: The following three types of redirects are supported.

Types of redirects	Examples
HTTP to HTTP	http://hostA to http://hostB
HTTP to HTTPS	http://hostA to https://hostB https://hostA:portA/pathA to https://hostB:portB/pathB
HTTPS to HTTPS	https://hostA to https://hostB

Q: What content types does ALB support for the message body of fixed-response action?

A: The following content types are supported: text/plain, text/css, text/html, application/javascript, application/json.

Q: How does AWS Lambda invocation via Application Load Balancer work?

A: HTTP(S) requests received by a load balancer are processed by the content-based routing rules. If the request content matches the rule—with an action to forward it to a target group through a Lambda function as a target—then the corresponding Lambda function is invoked. The content of the request (including headers and body) is passed on to the Lambda function in JavaScript object notation (JSON) format. The response from the Lambda function should be in JSON format. The response from the Lambda function is transformed into an HTTP response and sent to the client. The load balancer invokes your Lambda function using the AWS Lambda Invoke API, and requires that you provide invoke permissions for your Lambda function to the Elastic Load Balancing service.

Q: Does Lambda invocation via Application Load Balancer support requests over both HTTP and HTTPS protocol?

A: Yes. Application Load Balancer supports Lambda invocation for requests over both HTTP and HTTPS protocol.

Q: In which AWS Regions can I use Lambda functions as targets with the Application Load Balancer?

A: You can use Lambda as a target with the Application Load Balancer in US East (N. Virginia), US East (Ohio), US West (Northern California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), EU (Frankfurt), EU (Ireland), EU (London), EU (Paris), South America (São Paulo), and GovCloud (US-West) AWS Regions.

Q: Is the Application Load Balancer available in AWS Local Zones?

A: Yes, Application Load Balancer is available in the Local Zone in Los Angeles. Within the Los Angeles Local Zone, Application Load Balancer will operate in a single subnet and scale automatically to meet varying levels of application load without manual intervention.
Application Load Balancer Pricing FAQs

Q: How does Application Load Balancer pricing work?

A: You are charged for each hour or partial hour that an Application Load Balancer is running and the number of Load Balancer Capacity Units (LCU) used per hour.

Q: What is a Load Balancer Capacity Unit (LCU)?

A: An LCU is a new metric for determining how you pay for an Application Load Balancer. An LCU defines the maximum resource consumed in any one of the dimensions (new connections, active connections, bandwidth and rule evaluations) the Application Load Balancer processes your traffic.

Q: Will I be billed on Classic Load Balancers by LCU?

A: No, Classic Load Balancers will continue to be billed for bandwidth and hourly usage.

Q: How do I know the number of LCUs an Application Load Balancer is using?

A: We expose the usage of all four dimensions that constitute an LCU via Amazon CloudWatch.

Q: Will I be billed on all the dimensions in an LCU?

A: No. The number of LCUs per hour will be determined based on maximum resource consumed amongst the four dimensions that constitutes a LCU.

Q: Will I be billed on partial LCUs?

A: Yes.

Q: Is a free tier offered on an Application Load Balancer for new AWS accounts?

A: Yes. For new AWS accounts, a free tier for an Application Load Balancer offers 750 hours and 15 LCUs. This free tier offer is only available to new AWS customers, and is available for 12 months following your AWS sign-up date.

Q: Can I use a combination of Application Load Balancer and Classic Load Balancer as part of my free tier?

A: Yes. You can use both Classic and Application Load Balancers for 15 GB and 15 LCUs respectively. The 750 load balancer hours are shared between both Classic and Application Load Balancers.

Q: What are rule evaluations?

A: Rule evaluations are defined as the product of number of rules processed and the request rate averaged over an hour.

Q: How does the LCU billing work with different certificate types and key sizes?

A: Certificate key size affects only the number of new connections per second in the LCU computation for billing. The following table lists the value of this dimension for different key sizes for RSA and ECDSA certificates.

RSA certificates				
Key Size	<=2K	<=4K	<=8K	>8K
New connections/sec	25	5	1	0.25
ECDSA Certificates				
Key Size	<=256	<=384	<=521	>521
New connections/sec	25	5	1	0.25

Q: Am I charged for regional AWS data transfer when enabling cross-zone load balancing in Application Load Balancer?

A: No. Since cross-zone load balancing is always on with Application Load Balancer, you are not charged for this type of regional data transfer.

Q: Is user authentication in Application Load Balancer charged separately?

A: No. There is no separate charge for enabling the authentication functionality in Application Load Balancer. When using Amazon Cognito with Application Load Balancer, Amazon Cognito pricing will apply.

Q: How do you charge for Application Load Balancer usage with AWS Lambda targets?

A: You are charged as usual for each hour or partial hour that an Application Load Balancer is running and the number of Load Balancer Capacity Units (LCU) used per hour. For Lambda targets, each LCU offers 0.4 GB processed bytes per hour, 25 new connections per second, 3,000 active connections per minute, and 1,000 rule evaluations per second. For the processed bytes dimension, each LCU provides 0.4 GB per hour for Lambda targets versus 1 GB per hour for all other target types like Amazon EC2 instances, containers, and IP addresses. Please note that usual AWS Lambda charges apply to Lambda invocations by Application Load Balancer.

Q: How can I differentiate the bytes processed by Lambda targets versus bytes processed by other targets (Amazon EC2, containers, and on-premises servers)?

A: Applications Load Balancers emit two new CloudWatch metrics. LambdaTargetProcessedBytes metric indicates the bytes processed by Lambda targets, and the StandardProcessedBytes metric indicates bytes processed by all other target types.

Network Load Balancer

Q: Can I create a TCP or UDP (Layer 4) listener for my Network Load Balancer?

A: Yes. Network Load Balancers support both TCP, UDP, and TCP+UDP (Layer 4) listeners, as well as TLS listeners.

Q: What are the key features available with the Network Load Balancer?

A: Network Load Balancer provides both TCP and UDP (Layer 4) load balancing. It is architected to handle millions of requests per second and sudden volatile traffic patterns, and provides extremely low latencies. In addition, Network Load Balancer also supports TLS termination, preserves the source IP of the clients, and provides stable IP support and zonal isolation. It also supports long-running connections that are useful for WebSocket type applications.

Q: Can Network Load Balancer process both TCP and UDP protocol traffic on the same port?

A: Yes. To achieve this, you can use a TCP+UDP listener. For example, for a DNS service using both TCP and UDP, you can create a TCP+UDP listener on port 53, and the load balancer will process traffic for both UDP and TCP requests on that port. You must associate a TCP+UDP listener with a TCP+UDP target group.

Q: How does Network Load Balancer compare to what I get with the TCP listener on a Classic Load Balancer?

A: Network Load Balancer preserves the source IP of the client, which is not preserved in the Classic Load Balancer. Customers can use proxy protocol with Classic Load Balancer to get the source IP. Network Load Balancer automatically provides a static IP per Availability Zone (AZ) to the load balancer and also enables assigning an Elastic IP to the load balancer per AZ. This is not supported with Classic Load Balancer.

Q: Can I migrate to Network Load Balancer from Classic Load Balancer?

A: Yes. You can migrate to Network Load Balancer from Classic Load Balancer using one of the options listed in this document.

Q: Are there limits on the resources for my Network Load Balancer?

A: Yes, please refer to Network Load Balancer limits documentation for more information.

Q: Can I use the AWS Management Console to set up my Network Load Balancer?

A: Yes, you can use the AWS Management Console, AWS CLI, or the API to set up a Network Load Balancer.

Q: Can I use the existing API for Classic Load Balancers for my Network Load Balancers?

A: No. To create a Classic Load Balancer, use the 2012-06-01 API. To create a Network Load Balancer or an Application Load Balancer, use the 2015-12-01 API.

Q: Can I create my Network Load Balancer in a single Availability Zone?

A: Yes, you can create your Network Load Balancer in a single AZ by providing a single subnet when you create the load balancer.

Q: Does Network Load Balancer support DNS regional and zonal fail-over?

A: Yes, you can use Amazon Route 53 health checking and DNS failover features to enhance the availability of the applications running behind Network Load Balancers. Using Route 53 DNS failover, you can run applications in multiple AWS Availability zones and designate alternate load balancers for failover across regions.

In the event that you have your Network Load Balancer configured for multi-AZ, if there are no healthy Amazon EC2 instances registered with the load balancer for that AZ, or if the load balancer nodes in a given zone are unhealthy, then Route 53 will fail away to alternate load balancer nodes in other healthy AZs.

Q: Can I have a Network Load Balancer with a mix of ELB-provided IPs and Elastic IPs or assigned private IPs?

A: No. A Network Load Balancer's addresses must be completely controlled by you, or completely controlled by ELB. This is to ensure that when using Elastic IPs with a Network Load Balancer, all addresses known to your clients do not change.

Q: Can I assign more than one EIP to my Network Load Balancer in each subnet?

A: No. For each associated subnet a Network Load Balancer is in, the Network Load Balancer can only support a single public/internet facing IP address.

Q: If I remove/delete a Network Load Balancer what will happen to the Elastic IP addresses that were associated with it?

A: The Elastic IP Addresses that were associated with your load balancer will return to your allocated pool and be available for future use.

Q: Does Network Load Balancer support internal load balancers?

A: Network Load Balancer can be set up as an internet-facing load balancer or an internal load balancer, similar to what is possible with Application Load Balancer and Classic Load Balancer.

Q: Can the internal Network Load balancer support more than one private IP in each subnet?

A: No. For each associated subnet that a load balancer is in, the Network Load Balancer can only support a single private IP.

Q: Can I set up Websockets with my Network Load Balancer?

A: Yes, configure TCP listeners that route the traffic to the targets that implement WebSockets protocol (https://tools.ietf.org/html/rfc6455). Because WebSockets is a layer 7 protocol and

Network Load Balancer is operating at layer 4, no special handling exists in Network Load Balancer for WebSockets or other higher level protocols.

Q: Can I load balance to any arbitrary IP address?

A: Yes. You can use any IP address from the load balancer's VPC CIDR for targets within load balancer's VPC and any IP address from RFC 1918 ranges (10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16) or RFC 6598 range (100.64.0.0/10) for targets located outside the load balancer's VPC (EC2-Classic and on-premises locations reachable over AWS Direct Connect). Load balancing to IP address target type is supported for TCP listeners only, and is currently not supported for UDP listeners.

Q: Can I use Network Load Balancer to setup AWS PrivateLink?

A: Yes, Network Load Balancers with TCP and TLS Listeners can be used to setup AWS PrivateLink. You cannot set up PrivateLink with UDP listeners on Network Load Balancers.

Q: What is a UDP flow?

A: While user datagram protocol (UDP) is connectionless, the load balancer maintains UDP flow state based on 5-tuple hash, ensuring that packets sent in the same context are consistently forwarded to the same target. The flow is considered active as long as traffic is flowing and until the idle timeout is reached. Once the timeout threshold is reached, the load balancer will forget the affinity, and the incoming UDP packet will be considered a new flow and load-balanced to a new target.

Q: What is the idle timeout supported by Network Load Balancer?

A: Network Load Balancer idle timeout for TCP connections is 350 seconds. The idle timeout for UDP flows is 120 seconds.

Q: What is the benefit of targeting containers behind a load balancer with IP addresses instead of instance IDs?

A: Each container on an instance can now have its own security group, and does not need to share security rules with other containers. You can attach security groups to an ENI, and each ENI on an instance can have a different security group. You can map a container to the IP address of a particular ENI to associate security group(s) per container. Load balancing using IP addresses also allows multiple containers running on an instance use the same port (say port 80). The ability to use the same port across containers allows containers on an instance to communicate with each other through well-known ports instead of random ports.

Q: How can I load balance applications distributed across a VPC and on-premises location?

A: There are various ways to achieve hybrid load balancing. If an application runs on targets distributed between a VPC and an on-premises location, you can add them to the same target group using their IP addresses. To migrate to AWS without impacting your application, gradually add VPC targets to the target group and remove on-premises targets from the target group. You can also use separate load balancers for VPC and on-premises targets and use DNS weighting to achieve weighted load balancing between VPC and on-premises targets.

Q: How can I load balance to EC2-Classic instances?

A: You cannot load balance to EC2-Classic Instances when registering their Instance IDs as targets. However if you link these EC2-Classic instances to the load balancer's VPC using ClassicLink and use the private IPs of these EC2-Classic instances as targets, then you can

load balance to the EC2-Classic instances. If you are using EC2 Classic instances today with a Classic Load Balancer, you can easily migrate to a Network Load Balancer.

Q: How do I enable cross-zone load balancing in Network Load Balancer?

A: You can enable cross-zone loading balancing only after creating your Network Load Balancer. You achieve this by editing the load balancing attributes section and then selecting the cross-zone load balancing support checkbox.

Q: Am I charged for regional AWS data-transfer when I enable cross-zone load balancing in Network Load Balancer?

A: Yes, you will be charged for regional data transfer between Availability Zones with Network Load Balancer when cross-zone load balancing is enabled..

Q: Is there any impact of cross-zone load balancing on Network Load Balancer limits?

A: Yes. Network Load Balancer currently supports 200 targets per Availability Zone. For example, if you are in two AZs, you can have up to 400 targets registered with Network Load Balancer. If cross-zone load balancing is on, then the maximum targets reduce from 200 per AZ to 200 per load balancer. So, in the example above: When cross-zone load balancing is on, even though your load balancer is in two AZs, you are limited to 200 targets that can be registered to the load balancer.

Q: Does Network Load Balancer support TLS termination?

A: Yes, you can terminate TLS connections on the Network Load Balancer. You must install an SSL certificate on your load balancer. The load balancer uses this certificate to terminate the connection and then decrypt requests from clients before sending them to targets.

Q: Is source IP is preserved when terminating TLS on Network Load Balancer?

A: Source IP continues to be preserved even if you terminate TLS on the Network Load Balancer.

Q: What are the steps to get a SSL certificate?

A: You can either use AWS Certificate Manager to provision an SSL/TLS certificate, or you can obtain the certificate from other sources by creating the certificate request, getting the certificate request signed by a certificate authority (CA), and then uploading the certificate either using AWS Certification Manager (ACM) or the AWS Identity and Access Management (IAM) service.

Q: How can I enable Server Name Indication (SNI) for my Network Load Balancer?

A: SNI is automatically enabled when you associate more than one TLS certificate with the same secure listener on a load balancer. Similarly, SNI mode for a secure listener is automatically disabled when you have only one certificate associated to a secure listener.

Q: How does the Network Load Balancer integrate with AWS Certificate Manager (ACM) or Identity Access Manager (IAM)?

A: Network Load Balancer is integrated with AWS Certificate Management (ACM). Integration with ACM makes it very simple to bind a certificate to the load balancer thereby making the entire SSL offload process very easy. Purchasing, uploading, and renewing SSL/TLS certificates is a time-consuming manual and complex process. With ACM integration with Network Load Balancer, this whole process has been shortened to simply requesting a trusted SSL/TLS

certificate and selecting the ACM certificate to provision it with the load balancer. Once you create a Network Load balancer, you can now configure a TLS listener followed by an option to select a certificate from either ACM or Identity Access Manager (IAM). This experience is similar to what you have in Application Load Balancer or Classic Load Balancer.

Q: Is back-end server authentication supported with Network Load Balancer?

A: No, only encryption is supported to the back-ends with Network Load Balancer.

Q: What are the certificate types supported by Network Load Balancer?

A: Network Load Balancer only supports RSA certificates with 2K key size. We currently do not support RSA certificate key sizes greater than 2K or ECDSA certificates on the Network Load Balancer.

Q: In which AWS Regions is TLS Termination on Network Load Balancer supported?

A: You can use TLS Termination on Network Load Balancer in US East (N. Virginia), US East (Ohio), US West (Northern California), US West (Oregon), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), EU (Frankfurt), EU (Ireland), EU (London), EU (Paris), South America (São Paulo), and GovCloud (US-West) AWS Regions.

Network Load Balancer Pricing FAQs
Q: How does Network Load Balancer pricing work?

A: You are charged for each hour or partial hour that a Network Load Balancer is running and the number of Load Balancer Capacity Units (LCU) used by Network Load Balancer per hour.

Q: What is a Load Balancer Capacity Unit (LCU)?

A: An LCU is a new metric for determining how you pay for a Network Load Balancer. An LCU defines the maximum resource consumed in any one of the dimensions (new connections/flows, active connections/flows, and bandwidth) the Network Load Balancer processes your traffic.

Q: What are the LCU metrics for TCP traffic on Network Load Balancer?

A: The LCU metrics for the TCP traffic are as follows:

- 800 new TCP connections per second.
- 100,000 active TCP connections (sampled per minute).
- 1 GB per hour for Amazon EC2 instances, containers, and IP addresses as targets.

Q: What are the LCU metrics for UDP traffic on Network Load Balancer?

A: The LCU metrics for the UDP traffic are as follows:

- 400 new flows per second.
- 50,000 active UDP flows (sampled per minute).
- 1 GB per hour for Amazon EC2 instances, containers, and IP addresses as targets.

Q: What are the LCU metrics for TLS traffic on Network Load Balancer?

A: The LCU metrics for the TLS traffic are as follows:

- 50 new TLS connections per second.
- 3,000 active TLS connections (sampled per minute).
- 1 GB per hour for Amazon EC2 instances, containers, and IP addresses as targets.

Q: Will I be billed on all the dimensions (Processed Bytes, New Flows and Active Flows)?

A: No, for each protocol you are charged only on one of the three dimensions (the highest for the hour).

Q: Is new connections/flows per sec same as requests/sec?

A: No. Multiple requests can be sent in a single connection.

Q: Will I be billed on Classic Load Balancers by LCU?

A: No. Classic Load Balancers will continue to be billed for bandwidth and hourly charge.

Q: How do I know the number of LCUs a Network Load Balancer is using?

A: We will expose the usage of all three dimensions that constitutes a LCU via Amazon CloudWatch.

Q: Will I be billed on all the dimensions in an LCU?

A: No. The number of LCUs per hour will be determined based on maximum resource consumed amongst the three dimensions that constitutes a LCU.

Q: Will I be billed on partial LCUs?

A: Yes.

Q: Is a free tier offered on a Network Load Balancer for new AWS accounts?

A: Yes. For new AWS accounts, a free tier for a Network Load Balancer offers 750 hours and 15 LCUs. This free tier offer is only available to new AWS customers, and is available for 12 months following your AWS sign-up date.

Q: Can I use a combination of Network Load Balancer, Application Load Balancer and Classic Load Balancer as part of my free tier?

A: Yes. You can use Application and Network each for 15 LCUs and Classic for 15 GB respectively. The 750 load balancer hours are shared between Application, Network, and Classic Load Balancers.

Gateway Load Balancer

Getting started

Q: When should I use Gateway Load Balancer, as opposed to Network Load Balancer or Application Load Balancer?

A: You should use Gateway Load Balancer when deploying inline virtual appliances where network traffic is not destined for the Gateway Load Balancer itself. Gateway Load Balancer transparently passes all Layer 3 traffic through third-party virtual appliances, and is invisible to

the source and destination of the traffic. For more details on how these load balancers compare, see the features comparison page.

Q: Where is Gateway Load Balancer available?

A: Gateway Load Balancer is available in the following regions: AWS GovCloud (US-East), AWS GovCloud (US-West), US East (N. Virginia) - except in zone us1-az3, US East (Ohio), US West (Oregon), US West (N. California), Canada (Central), South America (Sao Paulo), EU (Ireland), EU (Frankfurt), EU (Stockholm), EU (London), EU (Paris), EU (Milan), Africa (Cape Town), Middle East (Bahrain), Asia Pacific (Sydney), Asia Pacific (Tokyo), Asia Pacific (Hong Kong), Asia Pacific (Singapore), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Osaka), China (Beijing), China (Ningxia).

Q: Is Gateway Load Balancer deployed per Region or per Availability Zone (AZ)?

A: Gateway Load Balancer runs within one AZ.

Q: What are the key features available with the Gateway Load Balancer?

A: Gateway Load Balancer provides both Layer 3 gateway and Layer 4 load balancing capabilities. It is a transparent bump-in-the-wire device that does not change any part of the packet. It is architected to handle millions of requests/second, volatile traffic patterns, and introduces extremely low latency. See Gateway Load Balancer features in this table.

Q: Does Gateway Load Balancer perform TLS termination?

A: Gateway Load Balancer does not perform TLS termination and does not maintain any application state. These functions are performed by the third-party virtual appliances it directs traffic to, and receives traffic from.

Q: Does Gateway Load Balancer maintain application state?

A: Gateway Load Balancer does not maintain application state, but it maintains flow stickiness to a specific appliance using 5-tuple (for TCP/UDP flows) or 3-tuple (for non-TCP/UDP flows).

Q: How does Gateway Load Balancer define a flow?

A: By default, Gateway Load Balancer defines a flow as a combination of a 5-tuple that comprises of Source IP, Destination IP, Protocol, Source Port, and Destination Port. Using the default 5-tuple hash, Gateway Load Balancer makes sure that both directions of a flow (i.e., source to destination, and destination to source) are consistently forwarded to the same target. The flow is considered active as long as traffic is flowing and until the idle timeout is reached. Once the timeout threshold is reached, the load balancer will forget the affinity, and incoming traffic packet will be considered as a new flow and may be load-balanced to a new target.

Note that the default 5-tuple hash can affect TCP based applications that use separate streams or port numbers for control and data, such as FTP, Microsoft RDP, Windows RPC and SSL VPN. Control and data flows of such applications can land on different target appliances and can cause traffic disruption. If you want to support such protocols, you can enable GWLB flow stickiness using 3-tuple (source IP, destination IP, transport protocol) or 2-tuple (source IP, destination IP). Please see flow stickiness documentation for how to change flow stickiness type.

Q: What is the idle timeout supported by Gateway Load Balancer?

A: Gateway Load Balancer idle timeout for TCP connections is 350 seconds. The idle timeout for non-TCP flows is 120 seconds. These timeouts are fixed and cannot be changed.

Q: Can appliances fragment the packets?

A: No. Target appliances of Gateway Load Balancer (GWLB) should not fragment packets they have received when sending them back to the GWLB. Fragments created by the appliance are dropped at Gateway Load Balancer because layer 4 header is not present in the IP fragments. To prevent fragmentation from happening on the appliance, we recommend enabling jumbo frame on your appliance or setting your appliance's network interface to use the maximum desired MTU, thus achieving transparent forwarding behavior by keeping the original packet contents as is.

Q: How does Gateway Load Balancer handle the failure of one virtual appliance instance in a single Availability Zone?

A: When a single virtual appliance instance fails, Gateway Load Balancer removes it from the routing list and reroutes traffic to a healthy appliance instance.

Q: How does Gateway Load Balancer handle the failure of all virtual appliances within a single AZ?

A: If all virtual appliances within an Availability Zone fail, Gateway Load Balancer will drop the network traffic. We recommend deploying Gateway Load Balancers in multiple AZs for greater availability. If all appliances fail in one AZ, scripts can be used to either add new appliances, or direct traffic to a Gateway Load Balancer in a different AZ.

Q: Can I configure an appliance to be a target for more than one Gateway Load Balancer?

A: Yes, multiple Gateway Load Balancers can point to same set of virtual appliances.

Q: What type of listener can I create for my Gateway Load Balancer?

A: Gateway Load Balancer is a transparent bump-in-the-wire device and listens to all types of IP traffic (including TCP, UDP, ICMP, GRE, ESP and others). Hence only IP listener is created on a Gateway Load Balancer.

Q: Are there limits on the resources for my Gateway Load Balancer?

A: Yes, please refer to Gateway Load Balancer limits documentation for more information.

Q: Can I use the AWS Management Console to set up my Gateway Load Balancer?

A: Yes, you can use the AWS Management Console, AWS CLI, or the API to set up a Gateway Load Balancer.

Q: Can I create my Gateway Load Balancer in a single Availability Zone?

A: Yes, you can create your Gateway Load Balancer in a single availability zone by providing a single subnet when you create the load balancer. However, we recommend using multiple availability zones for improved availability. You cannot add or remove availability zones for a Gateway Load Balancer after you create it.

Q: How do I enable cross-zone load balancing in Gateway Load Balancer?

A: By default, cross-zone load balancing is disabled. You can enable cross-zone loading balancing only after creating your Gateway Load Balancer. You achieve this by editing the load balancing attributes section and then by selecting the cross-zone load balancing support checkbox.

Q: Am I charged for AWS data-transfer when I enable cross-zone load balancing in Gateway Load Balancer?

A: Yes, you will be charged for data transfer between Availability Zones with Gateway Load Balancer when cross-zone load balancing is enabled. Check the charges in the data-transfer section at Amazon EC2 On-Demand Pricing page.

Q: Is there any impact of cross-zone load balancing on Gateway Load Balancer limits?

A: Yes. Gateway Load Balancer currently supports 300 targets per Availability Zone. For example, if you created Gateway Load Balancer in 3 Availability-Zones, you can have up to 900 targets registered. If cross-zone load balancing is on, then the maximum number of targets reduces from 300 per Availability Zone to 300 per Gateway Load Balancer.

Gateway Load Balancer Endpoints

Q: Why do I need a Gateway Load Balancer Endpoint?

In order to be valuable, virtual appliances need to introduce as little additional latency as possible, and traffic flowing to and from the virtual appliance must follow a secure connection. Gateway Load Balancer Endpoints create the secured, low-latency connections necessary to meet these requirements.

Q: How do Gateway Load Balancer Endpoints help with centralization?

Using a Gateway Load Balancer Endpoint, appliances can reside in different AWS accounts and VPCs. This allows appliances to be centralized in one location for easier management and reduced operational overhead.

Q: How do Gateway Load Balancer Endpoints work?

Gateway Load Balancer Endpoints are a new type of VPC endpoint that uses PrivateLink technology. As network traffic flows from a source (an Internet Gateway, a VPC, etc.) to the Gateway Load Balancer, and back, a Gateway Load Balancer Endpoint ensures private connectivity between the two. All traffic flows over the AWS network and data is never exposed to the internet, increasing both security and performance.

Q: How are PrivateLink Interface endpoints different than Gateway Load Balancer Endpoints?

A PrivateLink Interface endpoint is paired with a Network Load Balancer (NLB) in order to distribute TCP and UDP traffic that is destined for the web applications. In contrast, Gateway Load Balancer Endpoints are used with Gateway Load Balancers to connect the source and destination of traffic. Traffic flows from the Gateway Load Balancer Endpoint to the Gateway Load Balancer, through the virtual appliances, and back to the destination over secured PrivateLink connections.

Classic Load Balancer

Q: Which operating systems does the Classic Load Balancer support?

A: The Classic Load Balancer supports Amazon EC2 instances with any operating system currently supported by the Amazon EC2 service.

Q: Which protocols does the Classic Load Balancer support?

A: The Classic Load Balancer supports load balancing of applications using HTTP, HTTPS (Secure HTTP), SSL (Secure TCP) and TCP protocols.

Q: What TCP ports can I load balance?

A: You can perform load balancing for the following TCP ports:

- [EC2-VPC] 1-65535
- [EC2-Classic] 25, 80, 443, 465, 587, 1024-65535

Q: Does the Classic Load Balancer support IPv6 traffic?

A: Yes. Each Classic Load Balancer has an associated IPv4, IPv6, and dualstack (both IPv4 and IPv6) DNS name. IPv6 is not supported in VPC. You can use an Application Load Balancer for native IPv6 support in VPC.

Q: Can I configure my Amazon EC2 instances to only accept traffic from Classic Load Balancers?

A: Yes.

Q: Can I configure a security group for the front-end of Classic Load Balancers?

A: If you are using Amazon Virtual Private Cloud, you can configure security groups for the front end of your Classic Load Balancers.

Q: Can I use a single Classic Load Balancer for handling HTTP and HTTPS requests?

A: Yes, you can map HTTP port 80 and HTTPS port 443 to a single Classic Load Balancer.

Q: How many connections will my load balanced Amazon EC2 instances need to accept from each Classic Load Balancer?

A: Classic Load Balancers do not cap the number of connections that they can attempt to establish with your load balanced Amazon EC2 instances. You can expect this number to scale with the number of concurrent HTTP, HTTPS, or SSL requests or the number of concurrent TCP connections that the Classic load balancers receive.

Q: Can I load balance Amazon EC2 instances launched using a Paid AMI?

A: You can load balance Amazon EC2 instances launched using a paid AMI from AWS Marketplace. However, Classic Load Balancers do not support instances launched using a paid AMI from Amazon DevPay site.

Q: Can I use Classic Load Balancers in Amazon Virtual Private Cloud?

A: Yes. See the Elastic Load Balancing web page.

Q: Can I get a history of Classic Load Balancer API calls made on my account for security analysis and operational troubleshooting purposes?

A: Yes. To receive a history of Classic Load Balancer API calls made on your account, simply turn on CloudTrail in the AWS Management Console.

Q: Do Classic Load Balancers support SSL termination?

A: Yes, you can terminate SSL on Classic Load Balancers. You must install an SSL certificate on each load balancer. The load balancers use this certificate to terminate the connection and then decrypt requests from clients before sending them to the back-end instances.

Q: What are the steps to get a SSL certificate?

A: You can either use AWS Certificate Manager to provision a SSL/TLS certificate or you can obtain the certificate from other sources by creating the certificate request, getting the certificate request signed by a CA, and then uploading the certificate using the AWS Identity and Access Management (IAM) service.

Q: How do Classic Load Balancers integrate with AWS Certificate Manager (ACM)?

A: Classic Load Balancers are now integrated with AWS Certificate Management (ACM). Integration with ACM makes it very simple to bind a certificate to each load balancer thereby making the entire SSL offload process very easy. Typically purchasing, uploading, and renewing SSL/TLS certificates is a time-consuming manual and complex process. With ACM integrated with Classic Load Balancers, this whole process has been shortened to simply requesting a trusted SSL/TLS certificate and selecting the ACM certificate to provision it with each load balancer.

Q: How do I enable cross-zone load balancing in Classic Load Balancer?

A: You can enable cross-zone load balancing using the console, the AWS CLI, or an AWS SDK. See Cross-Zone Load Balancing documentation for more details.

Q: Am I charged for regional AWS data-transfer when I enable cross-zone load balancing in Classic Load Balancer?

A: No, you are not charged for regional data transfer between Availability Zones when you enable cross-zone load balancing for your Classic Load Balancer.

Amazon CloudWatch

Observe and monitor AWS resources and applications in the cloud and on premises.

How it works

Amazon CloudWatch collects and visualizes real-time logs, metrics, and event data in automated dashboards to streamline your infrastructure and application maintenance.

Use cases

Monitor application performance
Visualize performance data, create alarms, and correlate data to understand and resolve the root cause of performance issues in your AWS resources.

Perform root cause analysis
Analyze metrics, logs, logs analytics, and user requests to speed up debugging and reduce overall mean time to resolution.

Optimize resources proactively
Automate resource planning and lower costs by setting actions to occur when thresholds are met based on your specifications or machine learning models.

Test website impacts
Find out exactly when your website is impacted and for how long by viewing screenshots, logs, and web requests at any point in time.

Amazon CloudWatch is basically a metrics repository. An AWS service—such as Amazon EC2—puts metrics into the repository, and you retrieve statistics based on those metrics. If you put your own custom metrics into the repository, you can retrieve statistics on these metrics as well.

What is Amazon CloudWatch?

Amazon CloudWatch monitors your Amazon Web Services (AWS) resources and the applications you run on AWS in real time. You can use CloudWatch to collect and track metrics, which are variables you can measure for your resources and applications.

The CloudWatch home page automatically displays metrics about every AWS service you use. You can additionally create custom dashboards to display metrics about your custom applications, and display custom collections of metrics that you choose.

You can create alarms that watch metrics and send notifications or automatically make changes to the resources you are monitoring when a threshold is breached. For example, you can monitor the CPU usage and disk reads and writes of your

Amazon EC2 instances and then use that data to determine whether you should launch additional instances to handle increased load. You can also use this data to stop under-used instances to save money.

With CloudWatch, you gain system-wide visibility into resource utilization, application performance, and operational health.

Accessing CloudWatch

You can access CloudWatch using any of the following methods:

- **Amazon CloudWatch console** – https://console.aws.amazon.com/cloudwatch/
- **AWS CLI** – For more information, see Getting Set Up with the AWS Command Line Interface in the *AWS Command Line Interface User Guide*.
- **CloudWatch API** – For more information, see the Amazon CloudWatch API Reference.
- **AWS SDKs** – For more information, see Tools for Amazon Web Services.

Related AWS services

The following services are used along with Amazon CloudWatch:

- **Amazon Simple Notification Service (Amazon SNS)** coordinates and manages the delivery or sending of messages to subscribing endpoints or clients. You use Amazon SNS with CloudWatch to send messages when an alarm threshold has been reached. For more information, see Setting up Amazon SNS notifications.
- **Amazon EC2 Auto Scaling** enables you to automatically launch or terminate Amazon EC2 instances based on user-defined policies, health status checks, and schedules. You can use a CloudWatch alarm with Amazon EC2 Auto Scaling to scale your EC2 instances based on demand. For more information, see Dynamic Scaling in the *Amazon EC2 Auto Scaling User Guide*.
- **AWS CloudTrail** enables you to monitor the calls made to the Amazon CloudWatch API for your account, including calls made by the AWS Management Console, AWS CLI, and other services. When CloudTrail logging is turned on, CloudWatch writes log files to the Amazon S3 bucket that you specified when you configured CloudTrail. For more information, see Logging Amazon CloudWatch API calls with AWS CloudTrail.

- **AWS Identity and Access Management (IAM)** is a web service that helps you securely control access to AWS resources for your users. Use IAM to control who can use your AWS resources (authentication) and what resources they can use in which ways (authorization). For more information, see Identity and access management for Amazon CloudWatch.

Amazon CloudWatch FAQs

General

Q: What is Amazon CloudWatch?

Amazon CloudWatch is a monitoring service for AWS cloud resources and the applications you run on AWS. You can use Amazon CloudWatch to collect and track metrics, collect and monitor log files, and set alarms. Amazon CloudWatch can monitor AWS resources such as Amazon EC2 instances, Amazon DynamoDB tables, and Amazon RDS DB instances, as well as custom metrics generated by your applications and services, and any log files your applications generate. You can use Amazon CloudWatch to gain system-wide visibility into resource utilization, application performance, and operational health. You can use these insights to react and keep your application running smoothly.

To get started with monitoring, you can use Automatic Dashboards with built-in AWS best practices, explore account and resource-based view of metrics and alarms, and easily drill down to understand the root cause of performance issues.

Q: What can I use to access CloudWatch?

Amazon CloudWatch can be accessed via API, command-line interface, AWS SDKs, and the AWS Management Console.

Q: Which operating systems does Amazon CloudWatch support?

Amazon CloudWatch receives and provides metrics for all Amazon EC2 instances and should work with any operating system currently supported by the Amazon EC2 service.

Q: What access management policies can I implement for CloudWatch?

Amazon CloudWatch integrates with AWS Identity and Access Management (IAM) so that you can specify which CloudWatch actions a user in your AWS Account can perform. For example, you could create an IAM policy that gives only certain users in your organization permission to use GetMetricStatistics. They could then use the action to retrieve data about your cloud resources.

You can't use IAM to control access to CloudWatch data for specific resources. For example, you can't give a user access to CloudWatch data for only a specific set of instances or a specific LoadBalancer. Permissions granted using IAM cover all the cloud resources you use with CloudWatch. In addition, you can't use IAM roles with the Amazon CloudWatch command line tools.

Q: What is Amazon CloudWatch Logs?

Amazon CloudWatch Logs lets you monitor and troubleshoot your systems and applications using your existing system, application and custom log files.

With CloudWatch Logs, you can monitor your logs, in near real time, for specific phrases, values or patterns. For example, you could set an alarm on the number of errors that occur in your system logs or view graphs of latency of web requests from your application logs. You can then view the original log data to see the source of the problem. Log data can be stored and accessed indefinitely in highly durable, low-cost storage so you don't have to worry about filling up hard drives.

Q: What kinds of things can I do with CloudWatch Logs?

CloudWatch Logs is capable of monitoring and storing your logs to help you better understand and operate your systems and applications. You can use CloudWatch Logs in a number of ways.

Real-time application and system monitoring: You can use CloudWatch Logs to monitor applications and systems using log data. For example, CloudWatch Logs can track the number of errors that occur in your application logs and send you a notification whenever the rate of errors exceeds a threshold you specify. CloudWatch Logs uses your log data for monitoring, so no code changes are required.

Long-term log retention: You can use CloudWatch Logs to store your log data indefinitely in highly durable and cost effective storage without worrying about hard drives running out of space. The CloudWatch Logs Agent makes it easy to quickly move both rotated and non-rotated log files off of a host and into the log service. You can then access the raw log event data when you need it.

Q: What platforms does the CloudWatch Logs Agent support?

The CloudWatch Logs Agent is supported on Amazon Linux, Ubuntu, CentOS, Red Hat Enterprise Linux, and Windows. This agent will support the ability to monitor individual log files on the host.
Q: Does the CloudWatch Logs Agent support IAM roles?

Yes. The CloudWatch Logs Agent is integrated with Identity and Access Management (IAM) and includes support for both access keys and IAM roles.

Q: What is Amazon CloudWatch Logs Insights?

Amazon CloudWatch Logs Insights is an interactive, pay-as-you-go, and integrated log analytics capability for CloudWatch Logs. It helps developers, operators, and systems engineers understand, improve, and debug their applications, by allowing them to search and visualize their logs. Logs Insights is fully integrated with CloudWatch, enabling you to manage, explore, and analyze your logs. You can also leverage CloudWatch Metrics, Alarms and Dashboards with Logs to get full operational visibility into your applications. This empowers you to understand your applications, make improvements, and find and fix problems quickly, so that you can continue to innovate rapidly. You can write queries with aggregations, filters, and regular expressions to derive actionable insights from your logs. You can also visualize timeseries data, drill down into individual log events, and export your query results to CloudWatch Dashboards.

Q: How can I get started with CloudWatch Logs Insights?

You can immediately start using Logs Insights to run queries on all your logs being sent to CloudWatch Logs. There is no setup required and no infrastructure to manage. You can access Logs Insights from the AWS Management Console or programmatically through your applications by using the AWS SDK.

Q: What is Amazon CloudWatch Anomaly Detection?

Amazon CloudWatch Anomaly Detection applies machine-learning algorithms to continuously analyze single time series of systems and applications, determine a normal baseline, and surface anomalies with minimal user intervention. It allows you to create alarms that auto-adjust thresholds based on natural metric patterns, such as time of day, day of week, seasonality, or changing trends. You can also visualize metrics with anomaly detection bands on dashboards, monitoring, isolating, and troubleshooting unexpected changes in your metrics.

Q: How can I get started with Amazon CloudWatch Anomaly Detection?

It is easy to get started with Anomaly Detection. In the CloudWatch console, go to Alarms in the navigation pane to create an alarm, or start with Metrics to overlay the metric's expected values onto the graph as a band. You can also enable Anomaly Detection using the AWS CLI, AWS SDKs, or AWS CloudFormation templates.

Q: What is Amazon CloudWatch Contributor Insights?

Amazon CloudWatch now includes Contributor Insights, which analyzes time-series data to provide a view of the top contributors influencing system performance. Once set up, Contributor Insights runs continuously without needing additional user intervention. This helps developers and operators more quickly isolate, diagnose, and remediate issues during an operational event.

Q: How can I get started with CloudWatch Contributor Insights?

In the CloudWatch console, go to Contributor Insights in the navigation pane to create a Contributor Insights rule. You can also enable Contributor Insights using the AWS CLI, AWS SDKs, or AWS CloudFormation templates. Contributor Insights is available in all commercial AWS Regions. To learn more, please visit the documentation on CloudWatch Contributor Insights.

Q: What is Amazon CloudWatch ServiceLens?

Amazon CloudWatch ServiceLens is a new feature that enables you to visualize and analyze the health, performance, and availability of your applications in a single place. CloudWatch ServiceLens ties together CloudWatch metrics and logs as well as traces from AWS X-Ray to give you a complete view of your applications and their dependencies. This enables you to quickly pinpoint performance bottlenecks, isolate root causes of application issues, and determine users impacted. CloudWatch ServiceLens enables you to gain visibility into your applications in three main areas: Infrastructure monitoring (using metrics and logs to understand the resources supporting your applications), transaction monitoring (using traces to understand dependencies between your resources), and end user monitoring (using canaries to monitor your endpoints and notify you when your end user experience has degraded).

Q: How can I get started with CloudWatch ServiceLens?

If you already use AWS X-Ray, you can access CloudWatch ServiceLens on the CloudWatch console by default. If you do not yet use AWS X-Ray, you can get started by enabling AWS X-Ray on your applications using the X-Ray SDK. Amazon CloudWatch ServiceLens is available in all public AWS Regions where AWS-X-Ray is available. To learn more, visit the documentation on Amazon CloudWatch ServiceLens.

Q: What is Amazon CloudWatch Synthetics?

Amazon CloudWatch Synthetics allows you to monitor application endpoints more easily. It runs tests on your endpoints every minute, 24x7, and alerts you as soon as your application endpoints don't behave as expected. These tests can be customized to check for availability, latency, transactions, broken or dead links, step by step task completions, page load errors, load latencies for UI assets, complex wizard flows, or checkout flows in your applications. You can also use CloudWatch Synthetics to isolate alarming application endpoints and map them back to underlying infrastructure issues to reduce mean time to resolution.

Q: How can I get started with CloudWatch Synthetics?

It's easy to get started with CloudWatch Synthetics. You can write your first passing canary in minutes. To learn more, visit the documentation on Amazon CloudWatch Synthetics.

AWS resource & custom metrics monitoring

Q: What can I measure with Amazon CloudWatch Metrics?

Amazon CloudWatch allows you to monitor AWS cloud resources and the applications you run on AWS. Metrics are provided automatically for a number of AWS products and services, including Amazon EC2 instances, EBS volumes, Elastic Load Balancers, Auto Scaling groups, EMR job flows, RDS DB instances, DynamoDB tables, ElastiCache clusters, RedShift clusters, OpsWorks stacks, Route 53 health checks, SNS topics, SQS queues, SWF workflows, and Storage Gateways. You can also monitor custom metrics generated by your own applications and services.

Q: What is the retention period of all metrics?

You can publish and store custom metrics down to one-second resolution. Extended retention of metrics was launched on November 1, 2016, and enabled storage of all metrics for customers from the previous 14 days to 15 months. CloudWatch retains metric data as follows:

- Data points with a period of less than 60 seconds are available for 3 hours. These data points are high-resolution custom metrics.
- Data points with a period of 60 seconds (1 minute) are available for 15 days
- Data points with a period of 300 seconds (5 minute) are available for 63 days
- Data points with a period of 3600 seconds (1 hour) are available for 455 days (15 months)

Data points that are initially published with a shorter period are aggregated together for long-term storage. For example, if you collect data using a period of 1 minute, the data remains available for 15 days with 1-minute resolution. After 15 days this data is still available, but is aggregated and is retrievable only with a resolution of 5 minutes. After 63 days, the data is further aggregated and is available with a resolution of 1 hour. If you need availability of metrics longer than these periods, you can use the GetMetricStatistics API to retrieve the datapoints for offline or different storage.

The feature is currently available in US East (N. Virginia), US West (Oregon), US West (N. California), EU (Ireland), EU (Frankfurt), S. America (São Paulo), Asia Pacific (Singapore), Asia Pacific (Tokyo), Asia Pacific (Seoul), Asia Pacific (Mumbai), Asia Pacific (Sydney), EU (London), Canada (Central), US East (Ohio), and China (Beijing).

Q: What is the minimum resolution for the data that Amazon CloudWatch receives and aggregates?

The minimum resolution supported by CloudWatch is one-second data points, which is a high-resolution metric, or you can store metrics at one-minute granularity. Sometimes metrics are received by Cloudwatch at varying intervals, such as three-minute or five-minute intervals. If you do not specify that a metric is high resolution, by setting the StorageResolution field in the PutMetricData API request, then by default CloudWatch will aggregate and store the metrics at one-minute resolution.

Depending on the age of data requested, metrics will be available at the resolutions defined in the retention schedules above. For example, if you request for one-minute data for a day from 10 days ago, you will receive the 1440 data points. However, if you request for one-minute data from five months back, the UI will automatically change the granularity to one-hour and the GetMetricStatistics API will not return any output.

Q: Can I delete any metrics?

CloudWatch does not support metric deletion. Metrics expire based on the retention schedules described above.

Q: Will I lose the metrics data if I disable monitoring for an Amazon EC2 instance?

No. You can always retrieve metrics data for any Amazon EC2 instance based on the retention schedules described above. However, the CloudWatch console limits the search of metrics to two weeks after a metric is last ingested to ensure that the most up-to-date instances are shown in your namespace.

Q: Can I access the metrics data for a terminated Amazon EC2 instance or a deleted Elastic Load Balancer?

Yes. Amazon CloudWatch stores metrics for terminated Amazon EC2 instances or deleted Elastic Load Balancers for 15 months.

Q: Why does the graphing of the same time window look different when I view the metrics in five-minute and one-minute periods?

If you view the same time window in a 5 minute period versus a 1 minute period, you may see that data points are displayed in different places on the graph. For the period you specify in your graph, Amazon CloudWatch will find all the available data points and calculates a single, aggregate point to represent the entire period. In the case of a 5 minute period, the single data point is placed at the beginning of the 5 minute time window. In the case of a 1 minute period, the single data point is placed at the 1 minute mark. We recommend using a one minute period for troubleshooting and other activities that require the most precise graphing of time periods.

Q: What is a Custom Metric?

You can use Amazon CloudWatch to monitor data produced by your own applications, scripts, and services. A custom metric is any metric you provide to Amazon CloudWatch. For example, you can use custom metrics as a way to monitor the time to load a web page, request error rates, number of processes or threads on your instance, or amount of work performed by your application. You can get started with custom metrics by using the PutMetricData API, our sample monitoring scripts for Windows and Linux, CloudWatch collectd plugin, as well as a number of applications and tools offered by AWS partners.

Q: What resolution can I get from a Custom Metric?

A custom metric can be one of the following:

- Standard resolution, with data having one-minute granularity
- High resolution, with data at a granularity of one second

By default, metrics are stored at one-minute resolution in CloudWatch. You can define a metric as high-resolution by setting the StorageResolution parameter to one in the PutMetricData API request. If you do not set the optional StorageResolution parameter, then CloudWatch will default to storing the metrics at one-minute resolution.

When you publish a high-resolution metric, CloudWatch stores it with a resolution of one second, and you can read and retrieve it with a period of one second, five seconds, 10 seconds, 30 seconds, or any multiple of 60 seconds.

Custom metrics follow the same retention schedule listed above.

Q: What metrics are available at high resolution?

Currently, only custom metrics that you publish to CloudWatch are available at high resolution. High-resolution custom metrics are stored in CloudWatch at one-second resolution. High resolution is defined by the StorageResolution parameter in the PutMetricData API request, with a value of one, and is not a required field. If you do not specify a value for the optional StorageResolution field, then CloudWatch will store the custom metric at one-minute resolution by default.

Q: Are high-resolution custom metrics priced differently than regular custom metrics?

No, high-resolution custom metrics are priced in the same manner as standard one-minute custom metrics.

Q: When would I use a Custom Metric over having my program emit a log to CloudWatch Logs?

You can monitor your own data using custom metrics, CloudWatch Logs, or both. You may want to use custom metrics if your data is not already produced in log format, for example operating system processes or performance measurements. Or, you may want to write your own application or script, or one provided by an AWS partner. If you want to store and save individual measurements along with additional detail, you may want to use CloudWatch Logs.

Q: What statistics can I view and graph in CloudWatch?

You can retrieve, graph, and set alarms on the following statistical values for Amazon CloudWatch metrics: Average, Sum, Minimum, Maximum, and Sample Count. Statistics can be computed for any time periods between 60 seconds and one day. For high-resolution custom metrics, statistics can be computed for time periods between one second and three hours.

Q: What is CloudWatch Application Insights for .NET and SQL Server?

Amazon CloudWatch Application Insights for .NET and SQL Server is a capability that you can use to easily monitor your .NET and SQL Server applications. It helps identify and set up key metrics and logs across your application resources and technology stack, i.e. database, web (IIS) and application servers, OS, load balancers, queues, etc. It constantly monitors these telemetry data to detect and correlate anomalies and errors, to notify you of any problems in your application. To aid in troubleshooting, it creates automatic dashboards to visualize problems it detects which includes correlated metric anomalies and log errors, along with additional insights to point you to their potential root cause.

Q: What are the benefits of using CloudWatch Application Insights for .NET and SQL Server?

- Automatically recognize application metrics and logs: It scans your application resources, provides a list of recommended metrics and logs to monitor, and sets them up automatically, making it easier to set up monitoring for your applications.
- Intelligent problem detection: It uses built-in rules and machine learning algorithms to dynamically monitor and analyze symptoms of a problem across your application stack and detect application problems. It helps you reduce the overhead of dealing with individual metric spikes, or events, or log exceptions, and instead get notified on real problems, along with contextual information these problems.
- Faster troubleshooting: It assesses the detected problems to give you insights on them, such as the possible root cause of the detected problem and list of metrics and logs impacted because of the problem. You can provide feedback on generated insights to make the problem detection engine specific to your use case.

Q: How do I get started with monitoring using CloudWatch Application Insights for .NET and SQL Server?

On-board application: Specify the application you want to monitor by choosing the AWS Resource Group associated with it.

Identify application components: It analyzes your application resources to identify application components (standalone resources, or groups of related resources such as auto scaling groups and load balancer groups). You can also customize components by grouping resources for better insights and easy onboarding.

Enable monitoring: For your application components, you can specify the technology tier i.e. IIS front-end, .NET worker tier, etc. Based on your selection it provides a recommended set of metrics and logs that can be customized based on your needs. Once you save these “monitors”, Application Insights for .NET and SQL Server sets up CloudWatch to collect these on your behalf.

Once onboarded, Application Insights for .NET and SQL Server uses a combination of pre-built rules and machine learning models to start identifying application problems. It creates automated dashboards on CloudWatch with the list of problems detected, and a detailed view for these problems along with related anomalies and errors.

Q: What is CloudWatch Metric Streams?

CloudWatch Metric Streams is a feature that enables you to continuously stream CloudWatch metrics to a destination of your choice with minimal setup and configuration. It is a fully managed solution, and doesn't require you to write any code or maintain any infrastructure. With a few clicks, you can configure a metric stream to destinations like Amazon Simple Storage Service (S3). You can also send your metrics to a selection of third-party service providers to keep your operational dashboards up to date.

Q: Why should I use CloudWatch Metric Streams?

Metric Streams provides an alternative way of obtaining metrics data from CloudWatch without the need to poll APIs. You can create a metric stream with just a few clicks, and your metrics data will start to flow to your destination. You can easily direct your metrics to your data lake on AWS such as on Amazon S3, and start analyzing usage or performance with tools such as Amazon Athena. Metrics Streams also makes it easier to send CloudWatch metrics to popular third-party service providers using an Amazon Kinesis Data Firehose HTTP endpoint. You can create a continuous, scalable stream including the most up-to-date CloudWatch metrics data to power dashboards, alarms, and other tools that rely on accurate and timely metric data.

Q: How can I create and manage CloudWatch Metric Streams?

You can create and manage Metric Streams through the CloudWatch Console or programmatically through the CloudWatch API, AWS SDK, AWS CLI, or AWS CloudFormation to provision and configure Metric Streams. You can also use AWS CloudFormation templates provided by third-party service providers to set up Metric Streams delivery to destinations outside AWS. For more information, see the documentation on CloudWatch Metric Streams.

Q: Can I manage metrics to be included in my CloudWatch Metric Stream?

Yes. It is possible to choose to send all metrics by default, or create filter rules to include and exclude groups of metrics defined by namespace, e.g. AWS/EC2. Metric Streams automatically detects new metrics matching filter rules and includes metric updates in the stream. When resources are terminated, Metric Streams will automatically stop sending updates for the inactive metrics.

Q: What formats does CloudWatch Metric Streams support?

Metric Streams can output in either OpenTelemetry or JSON format. You can select the output format when creating or managing metric streams.

Q: Can I monitor the cost and volume of data delivered by CloudWatch Metric Streams?

Yes. You can visit the monitoring section of the Metric Streams console page. You will see automatic dashboards for the volume of metric updates over time. These metrics are also available under the AWS/CloudWatch namespace and can be used to create alarms to send notifications in the case of an unusual spike in volume.

Log monitoring

Q: What log monitoring does Amazon CloudWatch provide?

CloudWatch Logs lets you monitor and troubleshoot your systems and applications using your existing system, application and custom log files.

With CloudWatch Logs, you can monitor your logs, in near real time, for specific phrases, values or patterns. For example, you could set an alarm on the number of errors that occur in your system logs or view graphs of latency of web requests from your application logs. You can then view the original log data to see the source of the problem. Log data can be stored and accessed for up to as long as you need in highly durable, low-cost storage so you don't have to worry about filling up hard drives.

Q: What are Amazon CloudWatch Vended Logs?

Amazon CloudWatch Vended logs are logs that are natively published by AWS services on behalf of the customer. VPC Flow logs is the first Vended log type that will benefit from this tiered model. However, more AWS Service log types will be added to Vended Log type in the future.

Q: Is CloudWatch Logs available in all regions?

Please refer to Regional Products and Services for details of CloudWatch Logs service availability by region.

Q: How much does CloudWatch Logs cost?

Please see our pricing page for the latest information.

Q: What kinds of things can I do with my logs and Amazon CloudWatch?

CloudWatch Logs is capable of monitoring and storing your logs to help you better understand and operate your systems and applications. When you use CloudWatch Logs with your logs, your existing log data is used for monitoring, so no code change are required. Here are two examples of what you can do with Amazon CloudWatch and your logs:

Real time Application and System Monitoring: You can use CloudWatch Logs to monitor applications and systems using log data in near real time. For example, CloudWatch Logs can track the number of errors that occur in your application logs and send you a notification whenever the rate of errors exceeds a threshold you specify. Amazon CloudWatch uses your log data for monitoring and consequently it doesn't involve any code changes from you.

Long Term Log Retention: You can use CloudWatch Logs to store your log data for as long as you need in highly durable and cost-effective storage without worrying about hard drives running out of space. The CloudWatch Logs Agent makes it easy to quickly move both rotated and non-rotated log files off of a host and into the log service. You can then access the raw log event data when you need it.
Q: What types of data can I send to Amazon CloudWatch Logs from my EC2 instances running Microsoft SQL Server and Microsoft Windows Server?

You can configure the EC2Config service to send a variety of data and log files to CloudWatch including: custom text logs, Event (Application, Custom, Security, System) logs, Event Tracing (ETW) logs, and Performance Counter (PCW) data. Learn more about the EC2Config service here.

Q: How frequently does the CloudWatch Logs Agent send data?

The CloudWatch Logs Agent will send log data every five seconds by default and is configurable by the user.

Q: What log formats does CloudWatch Logs support?

CloudWatch Logs can ingest, aggregate and monitor any text based common log data or JSON-formatted logs.

Q: What if I configure the CloudWatch Logs Agent to send non-text log data?

The CloudWatch Logs Agent will record an error in the event it has been configured to report non text log data. This error is recorded in the /var/logs/awslogs.log.

Q: How do I start monitoring my logs with CloudWatch Logs?

You can monitor log events as they are sent to CloudWatch Logs by creating Metric Filters. Metric Filters turn log data into Amazon CloudWatch Metrics for graphing or alarming. Metric Filters can be created in the Console or the CLI. Metric Filters search for and match terms, phrases or values in your log events. When a Metric Filter finds one of the terms, phrases or values in your log events, it counts it in an Amazon CloudWatch Metric that you choose. For example, you can create a Metric Filter to search for and count the occurrence of the word “Error” in your log events. Metric Filters can also extract values from space delimited log events, such as the latency of web requests. You can also use conditional operators and wildcards to create exact matches. The Amazon CloudWatch Console can help you test your patterns before creating Metric Filters.

Q: What is the syntax of Metric Filter patterns?

A Metric Filter pattern can contain search terms or a specification of your common log or JSON event format.

For example, if you want to search for the term Error, the pattern for the metric filter would just be the term Error. Multiple search terms can be included to search for multiple terms. For example, if you wanted to count events which contained the terms Error and Exception you would use the pattern Error Exception. If you wanted to match the term Error Exception exactly, you would put double quotes around the search term, "Error Exception". You can specify as many search terms as you like.

CloudWatch Logs can also be used to extract values from a log event in common log or JSON format. For example, you could track the bytes transferred from your Apache access logs. You can also use conditional operators and wildcards to match and extract the data you are interested in. To use the extraction feature of Metric Filters, log events must be space delimited and use a starting and ending double quote """, or, a starting square brace "[" and a closing square brace "]"square, to enclose fields. Alternatively, they can be JSON-formatted log events. For the full details of the syntax and examples, please see the Developer Guide for Metric Filters.

Q: How do I know that a Metric Filter pattern I specified will match my log events?

CloudWatch Logs lets you test the Metric Filter patterns you want before you create a Metric Filter. You can test your patterns against your own log data that is already in CloudWatch Logs

or you can supply your own log events to test. Testing your pattern will show you which log events matched the Metric Filter pattern and, if extracting values, what the extracted value is in the test data. Metric Filter testing is available for use in the console and the CLI.

Q: Can I use regular expressions with my log data?

Amazon CloudWatch Metric Filters does not support regular expressions. To process your log data with regular expressions, consider using Amazon Kinesis and connect the stream with a regular expression processing engine.

Log management

Q: How do I retrieve my log data?

You can retrieve any of your log data using the CloudWatch Logs console or through the CloudWatch Logs CLI. Log events are retrieved based on the Log Group, Log Stream and time with which they are associated. The CloudWatch Logs API for retrieving log events is GetLogEvents.

Q: How do I search my logs?

You can use the CLI to retrieve your log events and search through them using command line grep or similar search functions.

Q: How long does CloudWatch Logs store my log data?

You can store your log data in CloudWatch Logs for as long as you want. By default, CloudWatch Logs will store your log data indefinitely. You can change the retention for each Log Group at any time.

Log analytics

Q: What permissions do I need to access Logs Insights?

To access Logs Insights, your IAM policy must include permissions for logs:DescribeLogGroups and logs:FilterLogEvents.

Q: What logs can I query with CloudWatch Logs Insights?

You can use Logs Insights to query all logs being sent to CloudWatch. Logs Insights automatically discovers the logs fields from logs from AWS services such as Lambda, CloudTrail, Route53, and VPC Flow Logs; and any application log that generates log events in JSON format. Additionally, for all log types, it generates 3 system fields @message, @logStream, and @timestamp for all logs sent to CloudWatch. @message contains the raw unparsed log event, @logStream contains the name of the source that generated the log event, and @timestamp contains the time at which the log event was added to CloudWatch.

Q: Which query language does CloudWatch Logs Insights support?

Logs Insights introduces a new purpose-built query language for log processing. The query language supports a few simple, but powerful query commands. You can write commands to retrieve one or more log fields, find log events that match one or more search criteria, aggregate

your log data, and extract ephemeral fields from your text-based logs. The query language is easy to learn, and Logs Insights offers in-product help in the form of sample queries, command descriptions, and query auto-completion to help you get started. You can find additional details about the query language here.

Q: What are the service limits for CloudWatch Logs Insights?

The service limits are documented here.

Q: What regions is CloudWatch Logs Insights available in?

Logs Insights is available in US West (Oregon), US West (N. California), US East (Ohio), US East (N. Virginia), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), EU (Frankfurt), EU (Ireland), EU (London), EU (Paris), South America (São Paulo).

Q: What type of queries does CloudWatch Logs Insights support?

You can write queries containing aggregations, filters, regular expressions, and text searches. You can also extract data from log events to create ephemeral fields, which can be further processed by the query language to help you access the information you are looking for. The query language supports string, numeric, and mathematical functions, such as concat, strlen, trim, log, and sqrt, among others. You can also use boolean and logical expressions, and aggregate functions such as min, max, sum, average, and percentile, among others. You can find additional details about the query language and supported functions here.

Q: What query commands and functions can I use with CloudWatch Logs Insights?

You can find a list of query commands here. You can find a list of supported functions here.

Q: What data visualizations can I use with CloudWatch Logs Insights?

You can use visualizations to identify trends and patterns that occur over time within your logs. Logs Insights supports visualizing data using line charts and stacked area charts. It generates visualizations for all queries containing one or more aggregate functions, where data is grouped over a time interval specified using the bin() function. You can find additional details about visualizing timeseries data here.

Q: Can I use regular expressions with CloudWatch Logs Insights?

You can use Java-style regular expressions with Logs Insights. Regular expressions can be used in the filter command. You can find examples of queries with regular expressions using the in-product help or here.

Q: How do I escape special characters with CloudWatch Logs Insights queries?

You can use backticks to escape special characters. Log field names that contain characters other than alphanumeric characters, @, and . require escaping with backticks.

Q: Why do certain log fields have a “@” sign and others don’t?

System fields generated by Logs Insights begin with @. Logs Insights currently generates 3 system fields @message which contains the raw, unparsed log event as sent to CloudWatch, @logStream which contains the name of the source that generated the log event, and @timestamp which contains the time when the log event was added to CloudWatch.

Q: Can I query historical logs with CloudWatch Logs Insights?

Logs Insights enables you to query log data that was added to CloudWatch Logs on or after November 5, 2018.

Q: Can I search for log events from a specific log stream?

You can search for log events from a specific log stream by adding the query command filter @logStream = "log_stream_name" to your log query.

Q: Today I use an AWS Partner ISV solution to analyze my logs from CloudWatch. What does CloudWatch Logs Insights change for me?

CloudWatch Logs already supports integration options with other AWS Services such as Amazon Kinesis, Amazon Kinesis Data Firehose, Amazon Elasticsearch and AWS Partner ISV solutions such as Splunk, Sumo Logic, and DataDog, among others, to provide you with choice and flexibility across all environments, for your custom log processing, enrichment, analytics, and visualization needs. In addition, the query capabilities of CloudWatch Logs Insights are available for programmatic access through the AWS SDK, to facilitate AWS ISV Partners to build deeper integrations, advanced analytics, and additional value on top of CloudWatch Logs Insights.

Q: How will I benefit from having access to query capabilities of CloudWatch Logs Insights through an AWS ISV Partner solution?

ISV Partner integrations with CloudWatch Logs Insights enable you to bring in your log data into one place and have the ability to analyze using the tools and frameworks of your choice in a high performance, cost-effective way, without having to move large amounts of data. It also provides you with faster access to your logs by removing the associated data transfer latencies and eliminates the operational complexities of configuring and maintaining certain data transfers.

Alarms

Q: What types of CloudWatch Alarms can be created?

You can create an alarm to monitor any Amazon CloudWatch metric in your account. For example, you can create alarms on an Amazon EC2 instance CPU utilization, Amazon ELB request latency, Amazon DynamoDB table throughput, Amazon SQS queue length, or even the charges on your AWS bill.

You can also create an alarm on custom metrics that are specific to your custom applications or infrastructure. If the custom metric is a high-resolution metric, you have the option of creating high-resolution alarms that alert as soon as 10-second or 30-second periods.

With composite alarms, you can combine multiple alarms into alarm hierarchies. This reduces alarm noise by triggering just once when multiple alarms fire at the same time. You can provide an overall state for a grouping of resources like an application, AWS Region, or Availability Zone.

Q: What actions can I take from a CloudWatch Alarm?

When you create an alarm, you can configure it to perform one or more automated actions when the metric you chose to monitor exceeds a threshold you define. For example, you can set an alarm that sends you an email, publishes to an SQS queue, stops or terminates an Amazon EC2 instance, or executes an Auto Scaling policy. Since Amazon CloudWatch alarms are integrated with Amazon Simple Notification Service, you can also use any notification type supported by SNS. You can use the AWS Systems Manager OpsCenter action to automatically create an OpsItem when an alarm enters the ALARM state. This helps you to quickly diagnose and remediate issues with AWS resources from a single console.

Q: What thresholds can I set to trigger a CloudWatch Alarm?

When you create an alarm, you first choose the Amazon CloudWatch metric you want it to monitor. Next, you choose the evaluation period (e.g., five minutes or one hour) and a statistical value to measure (e.g., Average or Maximum). To set a threshold, set a target value and choose whether the alarm will trigger when the value is greater than (>), greater than or equal to (>=), less than (<), or less than or equal to (<=) that value.

Q: My CloudWatch Alarm is constantly in the Alarm state, what did I do wrong?

Alarms continue to evaluate metrics against your chosen threshold, even after they have already triggered. This allows you to view its current up-to-date state at any time. You may notice that one of your alarms stays in the ALARM state for a long time. If your metric value is still in breach of your threshold, the alarm will remain in the ALARM state until it no longer breaches the threshold. This is normal behavior. If you want your alarm to treat this new level as OK, you can adjust the alarm threshold accordingly.

Q: How long can I view my Alarm history?

Alarm history is available for 14 days. To view your alarm history, log in to CloudWatch in the AWS Management Console, choose Alarms from the menu at left, select your alarm, and click the History tab in the lower panel. There you will find a history of any state changes to the alarm as well as any modifications to the alarm configuration.

Dashboards

Q: What is CloudWatch Dashboards?

Amazon CloudWatch Dashboards allow you to create, customize, interact with, and save graphs of AWS resources and custom metrics.

Q: How do I get started with CloudWatch Dashboards?

To get started, visit the Amazon CloudWatch Console and select “Dashboards”. Click the “Create Dashboard” button. You can also copy the desired view from Automatic Dashboards by clicking on Options -> “Add to Dashboard”.

Q: What are the advantages of Automatic Dashboards?

Automatic Dashboards are pre-built with AWS service recommended best practices, remain resource aware, and dynamically update to reflect the latest state of important performance metrics. You can now filter and troubleshoot to a specific view without adding additional code to reflect the latest state of your AWS resources. Once you have identified the root cause of a performance issue, you can quickly act by going directly to the AWS resource.

Q: Do the dashboards support auto refresh?

Yes. Dashboards will auto refresh while you have them open.

Q: Can I share my dashboard?

Yes, a dashboard is available to anyone with the correct permissions for the account with the dashboard.

Events

Q: What is CloudWatch Events?

Amazon CloudWatch Events (CWE) is a stream of system events describing changes in your AWS resources. The events stream augments the existing CloudWatch Metrics and Logs streams to provide a more complete picture of the health and state of your applications. You write declarative rules to associate events of interest with automated actions to be taken.

Q: What services emit CloudWatch Events?

Currently, Amazon EC2, Auto Scaling, and AWS CloudTrail are supported. Via AWS CloudTrail, mutating API calls (i.e., all calls except Describe*, List*, and Get*) across all services are visible in CloudWatch Events.

Q: What can I do once an event is received?

When an event matches a rule you've created in the system, you can automatically invoke an AWS Lambda function, relay the event to an Amazon Kinesis stream, notify an Amazon SNS topic, or invoke a built-in workflow.

Q: Can I generate my own events?

Yes. Your applications can emit custom events by using the PutEvents API, with a payload uniquely suited to your needs.

Q: Can I do things on a fixed schedule?

CloudWatch Events is able to generate events on a schedule you set by using the popular Unix cron syntax. By monitoring for these events, you can implement a scheduled application.

Q: What is the difference between CloudWatch Events and AWS CloudTrail?

CloudWatch Events is a near real time stream of system events that describe changes to your AWS resources. With CloudWatch Events, you can define rules to monitor for specific events and perform actions in an automated manner. AWS CloudTrail is a service that records API calls for your AWS account and delivers log files containing API calls to your Amazon S3 bucket or a CloudWatch Logs log group. With AWS CloudTrail, you can look up API activity history related to creation, deletion and modification of AWS resources and troubleshoot operational or security issues.

Q: What is the difference between CloudWatch Events and AWS Config?

AWS Config is a fully managed service that provides you with an AWS resource inventory, configuration history, and configuration change notifications to enable security and governance. Config rules help you determine whether configuration changes are compliant. CloudWatch Events is for reacting in near real time to resource state changes. It doesn't render a verdict on whether the changes comply with policy or give detailed history like Config/Config Rules do. It is a general purpose event stream.

Container Monitoring

Q: What is CloudWatch Container Insights?

CloudWatch Container Insights is a feature for monitoring, troubleshooting, and alarming on your containerized applications and microservices. Container Insights simplifies the isolation and analysis of performance issues impacting your container environment. DevOps and systems engineers have access to automatic dashboards in the CloudWatch console, giving them end-to-end operational visibility of metrics, logs, and distributed traces summarizing the performance and health of their Amazon Elastic Container Service for Kubernetes (EKS), Amazon Elastic Container Service (ECS), AWS Fargate, and Kubernetes clusters by pods/tasks, containers, and services.

Q: How can I get started with CloudWatch Container Insights?

You can get started collecting detailed performance metrics, logs, and metadata from your containers and clusters in just a few clicks by following these steps in the CloudWatch Container Insights documentation.

Q: How is CloudWatch Container Insights priced?

CloudWatch Container Insights automatically collects custom metrics from performance events ingested as CloudWatch Logs from your container environment.

Q: What is Prometheus and why do I want to collect Prometheus metrics in CloudWatch?

Prometheus is a popular open source monitoring project, part of the Cloud Native Compute Foundation (CNCF). The open source community has built over 150 plugins and defined a framework that DevOps teams can use to expose custom metrics to be collected using a pull-based approach from their applications. With this new feature, DevOps teams can automatically discover services for containerized workloads such as AWS App Mesh, NGINX, and Java/JMX. They can then expose custom metrics on those services, and ingest the metrics in CloudWatch. By curating the collection and aggregation of Prometheus metrics, CloudWatch users can monitor, troubleshoot, and alarm on application performance degradation and failures faster while reducing the number of monitoring tools required.

Q: How does pricing work when ingesting Prometheus metrics from my container environments?

You will be charged for what you use for the following: (1) CloudWatch Logs ingested by the Gigabyte (GB), (2) CloudWatch Logs stored, and (3) CloudWatch custom metrics.

Q: Is the storage retention configurable for Prometheus metrics high cardinality events ingested as CloudWatch Logs?

Yes. Each Kubernetes (k8s) cluster has its own log group for the events (e.g., /aws/containerinsights/<cluster-name>/prometheus) with their own configurable retention period. For more details, please refer to the documentation on log group retention.

Q: How does metric storage retention work for Prometheus metrics?

Prometheus metrics are automatically ingested as CloudWatch custom metrics. The retention period is 15 months per metric data point with automatic roll up (<60secs available for 3 hours, one min available for 15 days, 5 min available for 63 days, one hour available for 15 months). To learn more, see the documentation on CloudWatch metrics retention.

Q: Are all Prometheus metrics types supported for the Public Beta?

No. Current metric types supported are Gauge and Counters. Histogram and Summary metrics are planned for an upcoming release.

Q: Do you support PromQL as a query language?

No. All metrics are ingested as CloudWatch Logs events and can be queried using CloudWatch Logs Insights queries. For more information, see the documentation on CloudWatch Logs Insights search language syntax.

Lambda Monitoring

Q: What is CloudWatch Lambda Insights?

CloudWatch Lambda Insights is a feature for monitoring, troubleshooting, and optimizing the performance and cost of your Lambda functions. Lambda Insights simplifies the isolation and analysis of performance issues impacting your Lambda environments. DevOps and systems engineers have access to automatic dashboards in the CloudWatch console, giving them end-to-end operational visibility of metrics, logs, and traces summarizing the performance and health of their AWS Lambda functions.

Q: How can I get started with CloudWatch Lambda Insights?

You can get started collecting detailed performance metrics, logs, and metadata from your Lambda functions by following these steps in the CloudWatch Lambda Insights documentation.

Q: How is CloudWatch Lambda Insights priced?

CloudWatch Lambda Insights automatically collects custom metrics from performance events ingested as CloudWatch logs from your Lambda functions.

Digital Experience Monitoring

Q: What is Amazon CloudWatch Digital Experience Monitoring (DEM)?

Amazon CloudWatch DEM lets you monitor how your end users experience your applications (including performance, availability, and usability).

Spot intermittent issues, get notified even when there is no user traffic, and monitor your endpoints and UI using CloudWatch Synthetic canaries. Complement synthetic monitoring with CloudWatch RUM to understand end user impact and to get better visibility of your digital experience. With CloudWatch Evidently, improve your end user's digital experience by experimenting and validating new designs and features.

Q: What is Amazon CloudWatch RUM?

Amazon CloudWatch RUM is a real user monitoring feature that gives you visibility into an application's client-side performance to help you reduce mean time to resolution (MTTR). With CloudWatch RUM, you can collect client-side data on web application performance in real time to identify and debug issues. It complements the CloudWatch Synthetics data to give you more visibility into the end-user's digital experience. You can visualize anomalies in performance and use the relevant debugging data (such as error messages, stack traces, and user sessions) to fix performance issues (such as JavaScript errors, crashes, and latencies). You can also understand the range of end-user impacts, including number of sessions, geolocations, or browsers. CloudWatch RUM aggregates data on your users' journey through your application, which can help you determine which features to launch and bug fixes to prioritize.

Q: How can I get started with CloudWatch RUM?

Create an app monitor in CloudWatch RUM and add the lightweight web client in the HTML header of your application. Then start using CloudWatch RUM's dashboards to receive user insights from different geolocations, devices, platforms, and browsers.

Q: What is Amazon CloudWatch Evidently?

Amazon CloudWatch Evidently allows you to conduct experiments and identify unintended consequences of new features before rolling them out for general use, thereby reducing risk related to new feature roll-outs. Evidently allows you to validate new features across the full application stack before release, which makes for a safer release. When launching new features, you can expose them to a smaller user base, monitor key metrics such as page load times or conversions, and then dial up traffic. Evidently also allows developers to try out different designs, collect user data, and release the most effective design in production. It assists you in interpreting and acting on experiment results without the need for advanced statistical knowledge. You can use the insights provided by Evidently's statistical engine (such as anytime p-value and confidence intervals) to make decisions while an experiment is in progress.

Q: How can I get started with CloudWatch Evidently?

You can use the CloudWatch RUM JavaScript code snippet to collect client-side user journeys and performance metrics. If desired, you can also add custom metrics like conversions using the Evidently API. Next, new features to be tested can be instrumented with the CloudWatch Evidently SDK, which provides the ability to control how users get exposed to new features. Now you can run launches and experiments, using either the AWS console or CLI.

Q: What is Amazon CloudWatch Synthetics?

Amazon CloudWatch Synthetics allows you to monitor application endpoints more easily. It runs tests on your endpoints every minute, 24x7, and alerts you as soon as your application endpoints don't behave as expected. These tests can be customized to check for availability, latency, transactions, broken or dead links, step by step task completions, page load errors, load latencies for UI assets, complex wizard flows, or checkout flows in your applications. You

can also use CloudWatch Synthetics to isolate alarming application endpoints and map them back to underlying infrastructure issues to reduce mean time to resolution.

Q: How can I get started with CloudWatch Synthetics?

It's easy to get started with CloudWatch Synthetics. You can write your first passing canary in minutes. To learn more, visit the documentation on Amazon CloudWatch Synthetics.

Q: When should I use Amazon CloudWatch Evidently and when should I use AWS AppConfig?

The two services can be used separately, but are even better together.

AppConfig is a capability of AWS Systems Manager that you can use to create, manage, and deploy feature flags and other application configuration. When developing new features, you can use AppConfig to deploy a new feature to production, but hide it behind a flag toggle. Once you are ready to launch, you simply update your configuration to release the feature instantly or gradually.

For more advanced feature management and experimentation, you can use Evidently, which is a new capability of Amazon CloudWatch. With Evidently you can run experiments on different feature variations and compare them with a baseline, or launch a feature variation on a schedule, while monitoring business metrics like visit duration and revenue. Evidently also integrates with CloudWatch RUM, which provides client-side application performance monitoring, so RUM metrics can be used directly in Evidently.

Metrics analytics

Q: What is Amazon CloudWatch Metrics Insights?

CloudWatch Metrics Insights is a high-performance query engine that helps you slice and dice your operational metrics in real time and create aggregations on the fly using standard SQL queries. Metrics Insights helps you understand the status of your application health and performance by giving you the ability to analyze your metrics at scale. It is integrated with CloudWatch Dashboards, so you can save your queries into your health and performance dashboards to proactively monitor and pinpoint issues quickly.

Q: How can I get started with CloudWatch Metrics Insights?

To get started, just click on the metrics tab on your CloudWatch console, and you will find Metrics Insights as a built-in query engine under Query tab at no additional cost. While Metrics Insights comes with standard SQL language, you can also get started with Metrics Insights by using the visual query builder. To use the query builder, you select your metrics of interest, namespaces and dimensions visually, and the console automatically constructs your SQL queries for you, based on your selections. You can use the query editor to type in your raw SQL queries anytime to dive deep and pinpoint issues to further granular detail. Metrics Insights also comes with a set of out of the box sample queries that can help you start monitoring and investigating your application performance instantly. Metrics Insights is also available programmatically through CloudFormation, the AWS SDK, and CLI.

AWS Auto Scaling

Application scaling to optimize performance and costs.

AWS Auto Scaling monitors your applications and automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost. Using AWS Auto Scaling, it's easy to setup application scaling for multiple resources across multiple services in minutes. The service provides a simple, powerful user interface that lets you build scaling plans for resources including Amazon EC2 instances and Spot Fleets, Amazon ECS tasks, Amazon DynamoDB tables and indexes, and Amazon Aurora Replicas. AWS Auto Scaling makes scaling simple with recommendations that allow you to optimize performance, costs, or balance between them. If you're already using Amazon EC2 Auto Scaling to dynamically scale your Amazon EC2 instances, you can now combine it with AWS Auto Scaling to scale additional resources for other AWS services. With AWS Auto Scaling, your applications always have the right resources at the right time.

How it works

Benefits

SETUP SCALING QUICKLY

AWS Auto Scaling lets you set target utilization levels for multiple resources in a single, intuitive interface. You can quickly see the average utilization of all of your scalable resources without having to navigate to other consoles. For example, if your application uses Amazon EC2 and Amazon DynamoDB, you can use AWS Auto Scaling to manage resource provisioning for all of the EC2 Auto Scaling groups and database tables in your application.

MAKE SMART SCALING DECISIONS

AWS Auto Scaling lets you build scaling plans that automate how groups of different resources respond to changes in demand. You can optimize availability, costs, or a balance of both. AWS Auto Scaling automatically creates all of the scaling policies and sets targets for you based on your preference. AWS Auto Scaling monitors your application and automatically adds or removes capacity from your resource groups in real-time as demands change.

AUTOMATICALLY MAINTAIN PERFORMANCE

Using AWS Auto Scaling, you maintain optimal application performance and availability, even when workloads are periodic, unpredictable, or continuously changing. AWS Auto Scaling continually monitors your applications to make sure that they are operating at your desired performance levels. When demand spikes, AWS Auto Scaling automatically increases the capacity of constrained resources so you maintain a high quality of service.

PAY ONLY FOR WHAT YOU NEED

AWS Auto Scaling can help you optimize your utilization and cost efficiencies when consuming AWS services so you only pay for the resources you actually need. When demand drops, AWS Auto Scaling will automatically remove any excess resource capacity so you avoid overspending. AWS Auto Scaling is free to use, and allows you to optimize the costs of your AWS environment.

What is Amazon EC2 Auto Scaling?

Amazon EC2 Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. You create collections of EC2 instances, called *Auto Scaling groups*. You can specify the minimum number of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes below this size. You can specify the maximum number of instances in each Auto Scaling group, and Amazon EC2 Auto Scaling ensures that your group never goes above this size. If you specify the

desired capacity, either when you create the group or at any time thereafter, Amazon EC2 Auto Scaling ensures that your group has this many instances. If you specify scaling policies, then Amazon EC2 Auto Scaling can launch or terminate instances as demand on your application increases or decreases.

For example, the following Auto Scaling group has a minimum size of one instance, a desired capacity of two instances, and a maximum size of four instances. The scaling policies that you define adjust the number of instances, within your minimum and maximum number of instances, based on the criteria that you specify.

Auto Scaling components

The following table describes the key components of Amazon EC2 Auto Scaling.

	Groups Your EC2 instances are organized into *groups* so that they can be treated as a logical unit for the purposes of scaling and management. When you create a group, you can specify its minimum, maximum, and, desired number of EC2 instances. For more information, see Auto Scaling groups.

	Configuration templates Your group uses a *launch template*, or a *launch configuration* (not recommended, offers fewer features), as a configuration template for its EC2 instances. You can specify information such as the AMI ID, instance type, key pair, security groups, and block device mapping for your instances. For more information, see Launch templates and Launch configurations.
	Scaling options Amazon EC2 Auto Scaling provides several ways for you to scale your Auto Scaling groups. For example, you can configure a group to scale based on the occurrence of specified conditions (dynamic scaling) or on a schedule. For more information, see Scaling options.

Supported resources

AWS Auto Scaling supports the use of scaling plans for the following services and resources:

- **Amazon Aurora** – Increase or decrease the number of Aurora read replicas that are provisioned for an Aurora DB cluster.
- **Amazon EC2 Auto Scaling** – Launch or terminate EC2 instances by increasing or decreasing the desired capacity of an Auto Scaling group.
- **Amazon Elastic Container Service** – Increase or decrease the desired task count in Amazon ECS.
- **Amazon DynamoDB** – Increase or decrease the provisioned read and write capacity of a DynamoDB table or a global secondary index.
- **Spot Fleet** – Launch or terminate EC2 instances by increasing or decreasing the target capacity of a Spot Fleet.

Scaling plan features and benefits

Scaling plans provide the following features and benefits:

- **Resource discovery** – AWS Auto Scaling provides automatic resource discovery to help find resources in your application that can be scaled.
- **Dynamic scaling** – Scaling plans use the Amazon EC2 Auto Scaling and Application Auto Scaling services to adjust capacity of scalable resources to handle changes in traffic or workload. Dynamic scaling metrics can be standard utilization or throughput metrics, or custom metrics.
- **Built-in scaling recommendations** – AWS Auto Scaling provides scaling strategies with recommendations that you can use to optimize for performance, costs, or a balance between the two.
- **Predictive scaling** – Scaling plans also support predictive scaling for Auto Scaling groups. This helps to scale your Amazon EC2 capacity faster when there are regularly occurring spikes.

Important

If you have been using scaling plans only for configuring predictive scaling for your Auto Scaling groups, we strongly recommend that you use the predictive scaling policies of Auto Scaling groups instead. This recently introduced option offers enhanced features, such as using metrics aggregations to create new custom metrics or retain historical metric data across blue/green deployments.

AWS Auto Scaling FAQs

General

Q. What is AWS Auto Scaling?

AWS Auto Scaling is a new AWS service that helps you optimize the performance of your applications while lowering infrastructure costs by easily and safely scaling multiple AWS resources. It simplifies the scaling experience by allowing you to scale collections of related resources that support your application with just a few clicks. AWS Auto Scaling helps you configure consistent and congruent scaling policies across the full infrastructure stack backing your application. AWS Auto Scaling will automatically scale resources as needed to align to your selected scaling strategy, so you maintain performance and pay only for the resources you actually need.

Q. What are the benefits of AWS Auto Scaling?

AWS Auto Scaling is a fast, easy way to optimize the performance and costs of your applications.

- Setup scaling quickly: AWS Auto Scaling provides a unified scaling experience for all of the scalable resources powering your application. You can see the average utilization

for all of your scalable resources and quickly define target utilization levels for each group of like resources from a single, intuitive interface.

- Make smart scaling decisions: AWS Auto Scaling lets you automate how groups of different resources respond to changes in demand. Easy-to-understand scaling strategies let you choose to optimize availability, costs, or a balance of both. AWS Auto Scaling automatically creates all of the scaling policies and sets targets for you based on your preference.
- Automatically maintain performance: AWS Auto Scaling continually monitors resources underlying your application to make sure that they are operating at your desired performance levels. When demand spikes, AWS Auto Scaling automatically increases the capacity of constrained resources so you maintain a high quality of service.
- Anticipate costs and avoid overspending: AWS Auto Scaling can help you optimize your utilization and cost efficiencies when consuming AWS services so you only pay for the resources you actually need. When demand drops, AWS Auto Scaling will automatically remove any excess resource capacity so you avoid overspending.

Q. When should I use AWS Auto Scaling?

You should use AWS Auto Scaling if you have an application that uses one or more scalable resources and experiences variable load. A good example would be an e-commerce web application that receives variable traffic through the day. It follows a standard three tier architecture with Elastic Load Balancing for distributing incoming traffic, Amazon EC2 for the compute layer, and DynamoDB for the data layer. In this case, AWS Auto Scaling will scale one or more EC2 Auto Scaling groups and DynamoDB tables that are powering the application in response to the demand curve.

Q. How can I get started with AWS Auto Scaling?

AWS Auto Scaling allows you to select your applications based on resource tags or AWS CloudFormation stacks. In just a few clicks, you can create a scaling plan for your application, which defines how each of the resources in your application should be scaled. For each resource, AWS Auto Scaling creates a target tracking scaling policy with the most popular metric for that resource type and keeps it at a target value based on your selected scaling strategy. To set target values for your resource metrics, you can choose from three predefined scaling recommendations that optimize availability, optimize costs, or balance the two. Or, if you prefer, you can define your own target values. AWS Auto Scaling also automatically sets the min/max values for the resources.

Scaling Options

Q. What are the different ways that I can scale AWS resources?

AWS customers have multiple options for scaling resources. Amazon EC2 Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for your application. EC2 Auto Scaling can also detect when an instance is unhealthy, terminate it, and launch an instance to replace it. When you use EC2 Auto Scaling, your applications gain better fault tolerance, availability, and cost management.

To scale a resource other than EC2, you can use the Application Auto Scaling API, which allows you to define scaling policies to automatically scale your AWS resources or schedule one-time or recurring scaling actions. Application Auto Scaling can scale Amazon ECS services, Amazon EC2 Spot fleets, Amazon EMR clusters, Amazon AppStream 2.0 fleets, provisioned read and

write capacity for Amazon DynamoDB tables and global secondary indexes, Amazon Aurora Replicas, and Amazon SageMaker endpoint variants.

To configure automatic scaling for multiple resources across multiple services, use AWS Auto Scaling to create a scaling plan for the resources underlying your application. AWS Auto Scaling is also used to create predictive scaling for EC2 resources.

Q. When should I use AWS Auto Scaling vs. Amazon EC2 Auto Scaling?

You should use AWS Auto Scaling to manage scaling for multiple resources across multiple services. AWS Auto Scaling lets you define dynamic scaling policies for multiple EC2 Auto Scaling groups or other resources using predefined scaling strategies. Using AWS Auto Scaling to configure scaling policies for all of the scalable resources in your application is faster than managing scaling policies for each resource via its individual service console. It's also easier, as AWS Auto Scaling includes predefined scaling strategies that simplify the setup of scaling policies. You should also use AWS Auto Scaling if you want to create predictive scaling for EC2 resources.

You should use EC2 Auto Scaling if you only need to scale Amazon EC2 Auto Scaling groups, or if you are only interested in maintaining the health of your EC2 fleet. You should also use EC2 Auto Scaling if you need to create or configure Amazon EC2 Auto Scaling groups, or if you need to set up scheduled or step scaling policies (as AWS Auto Scaling supports only target tracking scaling policies).

EC2 Auto Scaling groups must be created and configured outside of AWS Auto Scaling, such as through the EC2 console, Auto Scaling API or via CloudFormation. AWS Auto Scaling can help you configure dynamic scaling policies for your existing EC2 Auto Scaling groups.

Q. When should I use AWS Auto Scaling vs. Auto Scaling for individual services?

You should use AWS Auto Scaling to manage scaling for multiple resources across multiple services. AWS Auto Scaling enables unified scaling for multiple resources, and has predefined guidance that helps make it easier and faster to configure scaling. If you prefer, you can instead choose to use the individual service consoles, Auto Scaling API, or Application Auto Scaling API to scale individual AWS services. You should also use the individual consoles or API if you want to setup step scaling policies or scheduled scaling, as AWS Auto Scaling creates target tracking scaling policies only.

Q. What is Predictive Scaling?

Predictive Scaling is a feature of AWS Auto Scaling that looks at historic traffic patterns and forecasts them into the future to schedule changes in the number of EC2 instances at the appropriate times going forward. Predictive Scaling uses machine learning models to forecast daily and weekly patterns.

Auto Scaling enhanced with Predictive Scaling delivers faster, simpler, and more accurate capacity provisioning resulting in lower cost and more responsive applications. By predicting traffic changes, Predictive Scaling provisions EC2 instances in advance of changing traffic, making Auto Scaling faster and more accurate.

Q. Which services can I use Predictive Scaling with?

At this time, Predictive Scaling only generates schedules for EC2 instances.

Q. How can I use Predictive Scaling with target tracking?

Predictive Scaling works with in conjunction with target tracking to make your EC2 capacity changes more responsive to your incoming application traffic. While Predictive Scaling sets up the minimum capacity for your application based on forecasted traffic, target tracking changes the actual capacity based on the actual traffic at the moment. Target tracking works to track the desired capacity utilization levels over varying traffic conditions and addresses unpredicted traffic spikes and other fluctuations. Predictive Scaling and target tracking are configured together by a user to generate a scaling plan.

Q. What is a scaling plan?

A scaling plan is a collection of scaling instructions for multiple AWS resources. You configure a scaling plan by first selecting all the EC2 resources underlying your application in AWS Auto Scaling. Then you pick the resource utilization metric that you would like to track, such as CPU utilization, and set the value to track, for example 50%. Finally, you select the CloudWatch metric that represents your input traffic flow – you might have to set this up if you haven't already.

The resource utilization metric and the incoming traffic metric are the key parameters for the scaling plan. The incoming traffic metric is used by Predictive Scaling to generate traffic forecasts. Based on these forecasts, Predictive Scaling then schedules future scaling actions to configure minimum capacity. Dynamic Scaling uses the resource utilization metric and its target value to dynamically change the EC2 capacity for your application over time as traffic varies.

Q. Can I configure a scaling plan without Predictive Scaling?

Yes, you can configure a scaling plan with only Dynamic Scaling and opt-out of Predictive Scaling. Conversely, you can also enable just Predictive Scaling without configuring Dynamic Scaling.

Q. How much historical data does Predictive Scaling need to generate the scaling plan?

Predictive Scaling needs up to two weeks of historical data but can generate a predictive scaling schedule with as little as a day's worth of data.

Q. How much into the future does Predictive Scaling forecast the traffic?

Every 24 hours, Predictive Scaling forecasts traffic 48 hours into the future and schedules capacity changes for those 48 hours.

Q. Can I configure Predictive Scaling to provision instances before an actual spike in traffic?

Yes, you can optionally configure buffer time to provision instances at some time before a predicted traffic change. This is useful for applications whose EC2 instances need some "warm-up" time before they are ready to serve application traffic .

Q. How much does Predictive Scaling cost?

As with other Auto Scaling features, Predictive Scaling is free to use. You pay for the resources being utilized for running your applications.

Q. How is AWS Auto Scaling different than the scaling capabilities for individual services?

The following table provides a comparison of AWS scaling options.

	AWS Auto Scaling	Amazon EC2 Auto Scaling	Auto Scaling for Other Services
Resources you can scale	EC2 Auto Scaling groups EC2 Spot Fleets ECS services DynamoDB provisioned capacity for tables & GSIs Aurora Replicas	EC2 Auto Scaling groups	EC2 Spot Fleets ECS services DynamoDB provisioned capacity for tables & GSIs Aurora Replicas EMR clusters Appstream 2.0 fleet Sagemaker endpoint variants
Scaling method	Application-wide scaling using a unified interface	One Auto Scaling group at a time	One resource at a time
Predictive Scaling	Yes (EC2 Only)	No	No
Automatic discovery of all scalable resources in your application	Yes	No	No
Ability to scale multiple resources across multiple services with a unified interface	Yes	No	No
Guidance and recommendations for setting up scaling policies	Yes	No	No

Ability to create and setup Auto Scaling groups	No	Yes	Not applicable
Ability to use Auto Scaling only for EC2 Fleet Management	No	Yes	Not applicable
Setup intelligent, self-optimizing target tracking scaling policies*	Yes	Yes	Yes
Setup scheduled scaling actions	No	Yes	Yes
Setup step scaling policies	No	Yes	Yes
Configure a scaling policy with different metrics and thresholds for each resource	No	Yes	Yes

* Recommended versus step scaling policies

Features

Q. What can I scale with AWS Auto Scaling?

You can use AWS Auto Scaling to setup scaling for the following resources in your application through a single, unified interface:

- Amazon EC2 Auto Scaling groups
- Amazon Elastic Container Service (ECS) services (currently ECS services cannot be discovered using resource tags)
- Amazon EC2 Spot Fleets
- Amazon DynamoDB throughput capacity
- Aurora replicas for Amazon Aurora

Q. How does AWS Auto Scaling make scaling recommendations?

AWS Auto Scaling bases its scaling recommendations on the most popular scaling metrics and thresholds used for Auto Scaling. It also recommends safe guardrails for scaling by providing recommendations for the minimum and maximum sizes of the resources. This way you can get started quickly and can then fine tune your scaling strategy over time.

Q. How do I select an application stack within AWS Auto Scaling?

You can either select an AWS CloudFormation stack or select resources based on common resource tag(s). Please note that currently, ECS services cannot be discovered using tags.

Q. How does AWS Auto Scaling discover what resources can scale?

AWS Auto Scaling will scan your selected AWS CloudFormation stack or resources with the specified tags to identify the supported AWS resource types that can be scaled. Please note that currently, ECS services cannot be discovered using tags.

AWS Elastic Beanstalk

Deploy and scale web applications.

How it works

AWS Elastic Beanstalk deploys web applications so that you can focus on your business.

Use cases

Quickly launch web applications
Deploy scalable web applications in minutes without the complexity of provisioning and managing underlying infrastructure.

Create mobile API backends for your applications
Use your favorite programming language to build mobile API backends, and Elastic Beanstalk will manage patches and updates.

Replatform critical business applications

Migrate stateful applications off legacy infrastructure to Elastic Beanstalk and connect securely to your private network.

What is AWS Elastic Beanstalk?

Amazon Web Services (AWS) comprises over one hundred services, each of which exposes an area of functionality. While the variety of services offers flexibility for how you want to manage your AWS infrastructure, it can be challenging to figure out which services to use and how to provision them.

With Elastic Beanstalk, you can quickly deploy and manage applications in the AWS Cloud without having to learn about the infrastructure that runs those applications. Elastic Beanstalk reduces management complexity without restricting choice or control. You simply upload your application, and Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring.

Elastic Beanstalk supports applications developed in Go, Java, .NET, Node.js, PHP, Python, and Ruby. When you deploy your application, Elastic Beanstalk builds the selected supported platform version and provisions one or more AWS resources, such as Amazon EC2 instances, to run your application.

You can interact with Elastic Beanstalk by using the Elastic Beanstalk console, the AWS Command Line Interface (AWS CLI), or **eb**, a high-level CLI designed specifically for Elastic Beanstalk.

You can also perform most deployment tasks, such as changing the size of your fleet of Amazon EC2 instances or monitoring your application, directly from the Elastic Beanstalk web interface (console).

To use Elastic Beanstalk, you create an application, upload an application version in the form of an application source bundle (for example, a Java .war file) to Elastic Beanstalk, and then provide some information about the application. Elastic Beanstalk automatically launches an environment and creates and configures the AWS resources needed to run your code. After your environment is launched, you can then manage your environment and deploy new application versions. The following diagram illustrates the workflow of Elastic Beanstalk.

After you create and deploy your application, information about the application—including metrics, events, and environment status—is available through the Elastic Beanstalk console, APIs, or Command Line Interfaces, including the unified AWS CLI.

AWS Elastic Beanstalk FAQs

General

Q: What is AWS Elastic Beanstalk?

AWS Elastic Beanstalk makes it even easier for developers to quickly deploy and manage applications in the AWS Cloud. Developers simply upload their application, and Elastic Beanstalk automatically handles the deployment details of capacity provisioning, load balancing, auto-scaling, and application health monitoring.

Q: Who should use AWS Elastic Beanstalk?

Those who want to deploy and manage their applications within minutes in the AWS Cloud. You don't need experience with cloud computing to get started. AWS Elastic Beanstalk supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker web applications.

Q: Which languages and development stacks does AWS Elastic Beanstalk support?

AWS Elastic Beanstalk supports the following languages and development stacks:

Apache Tomcat for Java applications

Apache HTTP Server for PHP applications

Apache HTTP Server for Python applications

Nginx or Apache HTTP Server for Node.js applications

Passenger or Puma for Ruby applications

Microsoft IIS 7.5, 8.0, and 8.5 for .NET applications

Java SE

Docker

Go

See Supported Platforms for a complete, up-to-date list of supported language and development stacks.

Q: Will AWS Elastic Beanstalk support other languages?

Yes. AWS Elastic Beanstalk is designed so that it can be extended to support multiple development stacks and programming languages in the future. AWS is working with solution providers on the APIs and capabilities needed to create additional Elastic Beanstalk offerings.

Q: What can developers now do with AWS Elastic Beanstalk that they could not before?

AWS Elastic Beanstalk automates the details of capacity provisioning, load balancing, auto scaling, and application deployment, creating an environment that runs a version of your application. You can simply upload your deployable code (e.g., WAR file), and AWS Elastic Beanstalk does the rest. The AWS Toolkit for Visual Studio and the AWS Toolkit for Eclipse allow you to deploy your application to AWS Elastic Beanstalk and manage it without leaving your IDE. Once your application is running, Elastic Beanstalk automates management tasks–such as monitoring, application version deployment, a basic health check–and facilitates log file access. By using Elastic Beanstalk, developers can focus on developing their application and are freed from deployment-oriented tasks, such as provisioning servers, setting up load balancing, or managing scaling.

Q: How is AWS Elastic Beanstalk different from existing application containers or platform-as-a-service solutions?

Most existing application containers or platform-as-a-service solutions, while reducing the amount of programming required, significantly diminish developers' flexibility and control. Developers are forced to live with all the decisions predetermined by the vendor–with little to no opportunity to take back control over various parts of their application's infrastructure. However, with AWS Elastic Beanstalk, developers retain full control over the AWS resources powering their application. If developers decide they want to manage some (or all) of the elements of their infrastructure, they can do so seamlessly by using Elastic Beanstalk's management capabilities.

Q: What elements of my application can I control when using AWS Elastic Beanstalk?

With AWS Elastic Beanstalk, you can:

Select the operating system that matches your application requirements (e.g., Amazon Linux or Windows Server 2016)

Choose from several Amazon EC2 instances including On-Demand, Reserved instances, and Spot instances

Choose from several available database and storage options

Enable login access to Amazon EC2 instances for immediate and direct troubleshooting

Quickly improve application reliability by running in more than one Availability Zone

Enhance application security by enabling HTTPS protocol on the load balancer

Access built-in Amazon CloudWatch monitoring and getting notifications on application health and other important events

Adjust application server settings (e.g., JVM settings) and pass environment variables

Run other application components, such as a memory caching service, side-by-side in Amazon EC2

Access log files without logging in to the application servers

Q: What are the Cloud resources powering my AWS Elastic Beanstalk application?

AWS Elastic Beanstalk uses proven AWS features and services, such as Amazon EC2, Amazon RDS, Elastic Load Balancing, Auto Scaling, Amazon S3, and Amazon SNS, to create an environment that runs your application. The current version of AWS Elastic Beanstalk uses the Amazon Linux AMI or the Windows Server 2019.

Q: What kinds of applications are supported by AWS Elastic Beanstalk?

AWS Elastic Beanstalk supports Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker, and is ideal for web applications. However, due to Elastic Beanstalk's open architecture, non-web applications can also be deployed using Elastic Beanstalk. We expect to support additional application types and programming languages in the future. See Supported Platforms to learn more.

Q: Which operating systems does AWS Elastic Beanstalk use?

AWS Elastic Beanstalk runs on the Amazon Linux AMI and the Windows Server AMI. Both AMIs are supported and maintained by Amazon Web Services and are designed to provide a stable, secure, and high-performance execution environment for Amazon EC2 Cloud computing.

Getting started

Q: How do I sign up for AWS Elastic Beanstalk?

To sign up for AWS Elastic Beanstalk, choose the Sign Up Now button on the Elastic Beanstalk detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Elastic Beanstalk process. After signing up, please refer to the AWS Elastic Beanstalk Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for AWS Elastic Beanstalk?

AWS Elastic Beanstalk registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a few minutes and involves receiving an automated phone call during the registration process and entering a PIN number using the phone key pad.

Q: How do I get started after I have signed up?

The best way to get started with AWS Elastic Beanstalk is to work through the AWS Elastic Beanstalk Getting Started Guide, part of our technical documentation. Within a few minutes, you will be able to deploy and use a sample application or upload your own application.

Q: Is there a sample application that I can use to check out AWS Elastic Beanstalk?

Yes. AWS Elastic Beanstalk includes a sample application that you can use to test drive the offering and explore its functionality.

Databases and storage

Q: Does AWS Elastic Beanstalk store anything in Amazon S3?

Yes. AWS Elastic Beanstalk stores your application files and, optionally, server log files in Amazon S3. If you are using the AWS Management Console, the AWS Toolkit for Visual Studio, or AWS Toolkit for Eclipse, an Amazon S3 bucket will be created in your account for you and the files you upload will be automatically copied from your local client to Amazon S3. Optionally, you may configure Elastic Beanstalk to copy your server log files every hour to Amazon S3. You do this by editing the environment configuration settings.

Q: Can I use Amazon S3 to store application data, like images?

Yes. You can use Amazon S3 for application storage. The easiest way to do this is by including the AWS SDK as part of your application’s deployable file. For example, you can include the AWS SDK for Java as part of your application's WAR file.

Q: What database solutions can I use with AWS Elastic Beanstalk?

AWS Elastic Beanstalk does not restrict you to any specific data persistence technology. You can choose to use Amazon Relational Database Service (Amazon RDS) or Amazon DynamoDB, or use Microsoft SQL Server, Oracle, or other relational databases running on Amazon EC2.

Q: How do I set up a database for use with AWS Elastic Beanstalk?

Elastic Beanstalk can automatically provision an Amazon RDS DB instance. The information about connectivity to the DB instance is exposed to your application by environment variables. To learn more about how to configure RDS DB instances for your environment, see the Elastic Beanstalk Developer Guide.

Q: Does this mean I need to modify the application code when moving from test to production?

Not with AWS Elastic Beanstalk. With Elastic Beanstalk, you can specify the connection information in the environment configuration. By extracting the connection string from the application code, you can easily configure different Elastic Beanstalk environments to use different databases.

Security

Q: How do I make my application private?

By default, your application is available publicly at myapp.elasticbeanstalk.com for anyone to access. You can use Amazon VPC to provision a private, isolated section of your application in a virtual network that you define. This virtual network can be made private through specific security group rules, network ACLs, and custom route tables. You can also easily control what other incoming traffic, such as SSH, is delivered or not to your application servers by changing the EC2 security group settings.

Q: Can I run my application inside a Virtual Private Cloud (VPC)?

Yes, you can run your applications in a VPC. For more details, see the AWS Elastic Beanstalk Developer Guide.

Q: Where can I find more information about security and running applications on AWS?

For more information about security on AWS, please refer to our Amazon Web Services: Overview of Security Processes document and visit our Security Center.

Q: Is it possible to use Identity & Access Management (IAM) with AWS Elastic Beanstalk?

Yes. IAM users with the appropriate permissions can now interact with AWS Elastic Beanstalk.

Q: Why should I use IAM with AWS Elastic Beanstalk?

IAM allows you to manage users and groups in a centralized manner. You can control which IAM users have access to AWS Elastic Beanstalk, and limit permissions to read-only access to Elastic Beanstalk for operators who should not be able to perform actions against Elastic Beanstalk resources. All user activity within your account will be aggregated under a single AWS bill.

Q: How do I create IAM users?

You can use the IAM console, IAM command line interface (CLI), or IAM API to provision IAM users. By default, IAM users have no access to AWS services until permissions are granted.

Q: How do I grant an IAM user access to AWS Elastic Beanstalk?

You can grant IAM users access to services by using policies. To simplify the process of granting access to AWS Elastic Beanstalk, you can use one of the policy templates in the IAM console to help you get started. Elastic Beanstalk offers two templates: a read-only access template and a full-access template. The read-only template grants read access to Elastic Beanstalk resources. The full-access template grants full access to all Elastic Beanstalk operations, as well as permissions to manage dependent resources, such as Elastic Load Balancing, Auto Scaling, and Amazon S3. You can also use the AWS Policy Generator to create custom policies. For more details, see the AWS Elastic Beanstalk Developer Guide.

Q: Can I restrict access to specific AWS Elastic Beanstalk resources?

Yes. You can allow or deny permissions to specific AWS Elastic Beanstalk resources, such as applications, application versions, and environments.

Q: Who gets billed for the AWS resources that an IAM user creates?

All resources created by IAM users under a root account are owned and billed to the root account.

Q: Who has access to an AWS Elastic Beanstalk environment launched by an IAM user?

The root account has full access to all AWS Elastic Beanstalk environments launched by any IAM user under that account. If you use the Elastic Beanstalk template to grant read-only access to an IAM user, that user will be able to view all applications, application versions, environments, and any associated resources in that account. If you use the Elastic Beanstalk template to grant full access to an IAM user, that user will be able to create, modify, and terminate any Elastic Beanstalk resources under that account.

Q: Can an IAM user access the AWS Elastic Beanstalk console?

Yes. An IAM user can access the AWS Elastic Beanstalk console using their username and password.

Q: Can an IAM user call the AWS Elastic Beanstalk API?

Yes. An IAM user can use their access key and secret key to perform operations using the Elastic Beanstalk API.

Q: Can an IAM user use the AWS Elastic Beanstalk command line interface?

Yes. An IAM user can use their access key and secret key to perform operations using the AWS Elastic Beanstalk command line interface (CLI).

Managed platform updates

Q: How can I keep the underlying platform of the environment running my application automatically up-to-date?

You can opt-in to having your AWS Elastic Beanstalk environments automatically updated to the latest version of the underlying platform running your application during a specified maintenance window. Elastic Beanstalk regularly releases new versions of supported platforms (Java, PHP, Ruby, Node.js, Python, .NET, Go, and Docker) with operating system, web and application server, and language and framework updates.

Q: How can I get started with managed platform updates?

To let Elastic Beanstalk automatically manage your platform updates, you must enable managed platform updates in the Configuration tab of the Elastic Beanstalk console or use the EB CLI or API. After you have enabled the feature, you can configure which types of updates to allow and when updates can occur.

Q: What kinds of platform version updates will managed platform updates apply?

AWS Elastic Beanstalk can automatically perform platform updates for new patch and minor platform versions. Elastic Beanstalk will not automatically perform major platform version updates (e.g., Java 7 Tomcat 7 to Java 8 Tomcat 8) because they include backwards incompatible changes and require additional testing. In these cases, you must manually initiate the update.

Q: How does AWS Elastic Beanstalk distinguish between “major,” “minor,” and “patch” version releases?

AWS Elastic Beanstalk platforms are versioned using this pattern: MAJOR.MINOR.PATCH (e.g., 2.0.0). Each portion is incremented as follows:

MAJOR version when there are incompatible changes.

MINOR version when there is additional functionality added in a backward-compatible manner.

PATCH version when there are backward-compatible bug fixes.

Q: When and how can I perform major version updates?

You can perform major version updates at any time using the AWS Elastic Beanstalk management console, API, or CLI. You have the following options to perform a major version update:

Apply the update in-place on an existing environment. See Updating Your Elastic Beanstalk Environment's Platform Version.

Create a clone of an existing environment with the new platform version. See Clone an Environment to learn more.

Q: How does Elastic Beanstalk apply managed platform updates?

The updates are applied using an immutable deployment mechanism that ensures that no changes are made to the existing environment until a parallel fleet of Amazon EC2 instances, with the updates installed, is ready to be swapped with the existing instances, which are then terminated. In addition, if the Elastic Beanstalk health system detects any issues during the update, traffic is redirected to the existing fleet of instances, ensuring minimal impact to end users of your application.

Q: Will my application be available during the maintenance windows?

Since managed platform updates use an immutable deployment mechanism to perform the updates, your application will be available during the maintenance window and consumers of your application will not be impacted by the update.

Q: What does it cost to use managed platform updates?

There is no additional charge for the managed platform updates feature. You simply pay for the additional EC2 instances necessary to perform the update for the duration of the update.

Q: What is a maintenance window?

A maintenance window is a weekly two-hour-long time slot during which AWS Elastic Beanstalk will initiate platform updates if managed platform updates is enabled and a new version of the platform is available. For example, if you select a maintenance window that begins every Sunday at 2 AM, AWS Elastic Beanstalk will initiate the platform update sometime between 2-4 AM every Sunday. It is important to note that, depending on the configuration of your applications, updates could complete outside of the maintenance window.

The maintenance window is set on a per-environment basis, providing you the option to set different maintenance windows for your various application components or applications. This allows environment updates to be staggered if you do not want multiple pieces of your application to be updated at the same time. If you enable managed platform updates but do not specify a maintenance window, a default weekly 2-hour window will be assigned for your environment. If you want to change when maintenance is performed on your behalf, you can do so by modifying the managed update configuration in the AWS Management Console or by using the UpdateEnvironment API.

Q: How will I be notified of the availability of new platform versions?

You will be notified about the availability of new platform versions through the AWS Management Console, forum announcements, and release notes.

Q: Where can I find details of changes between platform versions?

Details on changes between platform versions can be found on the AWS Elastic Beanstalk Release Notes page.

Q: What operations can I perform on the environment while a managed update is in progress?

The only action available to you while a managed platform update is in-progress is ‘abort’. This will allow you to stop the update immediately and roll back to the previous version.

Q: Which platform version will my environment be updated to if there are multiple new versions released in between maintenance windows?

Your environment will always be updated to the latest version available based on the level (minor plus patch or patch only) you have selected.

Q: Where can I find details of all the managed platform updates that have been performed on my environment?

Details for every managed platform update are available on the events page and are tagged with an event type of “MAINTENANCE.”

Q: How often are platform version updates released?

The number of version releases in a given year varies based on the frequency and content of releases and patches from the language/framework’s vendor or core team, and the outcome of a thorough vetting of these releases and patches by our platform engineering team.

AWS Graviton support

Q: How do I deploy a new workload with the Graviton processor from the Elastic Beanstalk console?

To deploy your application with arm64-based processors on the Elastic Beanstalk console, you can select processor architecture and instance type from capacity tab in Configure more options settings.

Q: How do I deploy a new workload with the Graviton processor from the AWS CLI, Elastic Beanstalk CLI, or infrastructure as code services?

To deploy your application using the Elastic Beanstalk CLI, AWS CLI, CFN, or AWS CDK, refer to Elastic Beanstalk Developer Guide.

Q: Do I need to recompile my workload before migrating to Graviton?

If your workload is on an interpreted programming language such as Node.js, Python, Tomcat, PHP, or Ruby, you do not need to recompile your workload to use Graviton. If you are using Go or .Net Core for your workload, you need to update the build command for the arm64 instance type. You also need to recompile binary dependencies or use an arm64 compatible release of binary dependencies. If you are using Docker, your Docker image must be multi-architecture and support deploying to both x86 and arm64.

Q: Which platform branches are supported by Graviton on Elastic Beanstalk?

Elastic Beanstalk supports Graviton on 64bit Amazon Linux 2 for a variety of platform and branches. See the documentation for a full list.

Q: What are the use cases where I can use the Graviton processor?

You can easily transition your workload to Graviton and take advantage of performance and cost benefits in the following use cases: Linux-based workloads built primarily on open-source technologies; containerized and microservices-based applications such as Docker and MC Docker; applications written in portable programming languages such as Java, Python, .NET Core, node.js, and PHP; Compiled C/C++, Rust, or Go applications; .NET Core (v3.1+)

workloads running on Linux; multi-threaded workloads;non-uniform memory access (NUMA) sensitive workloads; and arm64-native software development and testing.

Amazon Lightsail

Build applications and websites fast with low-cost, pre-configured cloud resources.

Amazon Lightsail offers easy-to-use virtual private server (VPS) instances, containers, storage, databases, and more at a cost-effective monthly price.

Amazon Lightsail helps developers get started using Amazon Web Services (AWS) to build websites or web applications. It includes the features that you need to launch your project: instances (virtual private servers), managed databases, object storage, load balancers, content delivery network (CDN) distributions, SSD-based block storage, static IP addresses, DNS management of registered domains, and snapshots (backups).

Use cases

Launch simple web applications
Use pre-configured development stacks like LAMP, Nginx, MEAN, and Node.js. to get online quickly and easily.

Create custom websites
Build and personalize your blog, ecommerce, or personal website in just a few clicks, with pre-configured applications like WordPress, Magento, Prestashop, and Joomla.

Build small business applications
Launch line-of-business software such as file storage and sharing, backups, financial and accounting software, and more.

Spin up test environments
Easily create and delete development sandboxes and test environments where you can try out new ideas, risk free.

Amazon Lightsail FAQs

General

Q: What is Amazon Lightsail?

Amazon Lightsail is a virtual private server (VPS) provider and is the easiest way to get started with AWS for developers, small businesses, students, and other users who need a solution to build and host their applications on cloud. Lightsail provides developers compute, storage, and networking capacity and capabilities to deploy and manage websites and web applications in the cloud. Lightsail includes everything you need to launch your project quickly – virtual machines, containers, databases, CDN, load balancers, DNS management etc. – for a low, predictable monthly price.

Q: What is a Virtual Private Server?

A virtual private server, also known as an "instance", allows users to run websites and web applications in a highly secure and available environment, while being cost effective.

Q: What are the benefits of a VPS?

There are many benefits to using a virtual private server, including affordability, scalability, security, and customizable resources.

Q: What can I do with Lightsail?

You can choose among several preconfigured VPS plans that include everything you need to easily deploy and manage your application. Lightsail is best suited for projects that require a few virtual private servers and users who prefer a simple management interface. Common use cases for Lightsail include running websites, web applications, blogs, e-commerce sites, simple software, and more.

Q: What is a Lightsail instance plan?

Also referred to as a bundle, a Lightsail plan includes a virtual private server with a fixed amount of memory (RAM) and compute (vCPUs), SSD-based storage (disks), and a data transfer allowance. Lightsail plans also offer static IP addresses (5 per account) and DNS management (3 domain zones per account). Lightsail plans are charged on an hourly, on-demand basis, so you only pay for a plan when you're using it.

Q: Are Lightsail plans the same across AWS Regions?

As Lightsail expands its availability across more Regions, we’re evolving our infrastructure to provide a better overall customer experience. As a result, plans in certain Regions will have different specifications, such as an additional vCPU, compared to plans in other Regions.

Q: What is a Lightsail instance?

A Lightsail instance is a virtual private server (VPS) that lives in the AWS Cloud. Use your Lightsail instances to store your data, run your code, and build web-based applications or websites. Your instances can connect to each other and to other AWS resources through both

public (Internet) and private (VPC) networking. You can create, manage, and connect easily to instances right from the Lightsail console.

Q: What software can I run on my instance?

Lightsail offers a range of operating system and application templates that are automatically installed when you create a new Lightsail instance. Application templates include WordPress, Drupal, Joomla!, Ghost, Magento, Redmine, LAMP, Nginx (LEMP), MEAN, Node.js, Django, and more.

You can install additional software on your instances by using the in-browser SSH or your own SSH client.

Q: How do I create a Lightsail instance?

After logging in to Lightsail, you can use the Lightsail console, command line interface (CLI), or API to create and manage instances.

The first time you log in to the console, choose Create Instance. The create instance page is where you can choose the software, location, and name for your instance. Once you choose Create, your new instance will spin up automatically within minutes.

Q: Does Lightsail offer an API?

Yes. Everything you do in the Lightsail console is backed by a publicly available API. Learn how to install and use the Lightsail CLI and API.

Q: How do I sign up for Lightsail?

To start using Lightsail, choose Get Started and log in. You use your Amazon Web Services account to access Lightsail; if you don't already have one, you'll be prompted to create one.

Lightsail resources

Q: How do Lightsail instances perform?

Lightsail instances are specifically engineered by AWS for web servers, developer environments, and small database use cases. Such workloads don't use the full CPU often or consistently, but occasionally need a performance burst. Lightsail uses burstable performance instances that provide a baseline level of CPU performance with the additional ability to burst above the baseline. This design enables you to get the performance you need, when you need it, while protecting you from the variable performance or other common side effects that you might typically experience from over-subscription in other environments. For more information on burstable performance, go here.

If you need highly configurable environments and instances with consistently high CPU performance for applications such as video encoding or HPC applications, we recommend you use Amazon EC2.

Q: How do I connect to a Lightsail instance?

Lightsail offers a 1-click secure connection to your instance's terminal right from your browser, supporting SSH access for Linux/Unix-based instances and RDP access for Windows-based instances. To use 1-click connections, launch your instance management screens, click Connect using SSH or Connect using RDP, and a new browser window opens and automatically connects to your instance.

If you prefer to connect to your Linux/Unix-based instance using your own client, Lightsail will do the SSH key storing and management work for you, and provide you with a secure key to use in your SSH client.

Q: How do I use IPs in Lightsail?

Each Lightsail instance automatically gets a private IP address and a public IP address. You can use the private IP to transmit data between Lightsail instances and AWS resources privately, for free. You can use the public IP to connect to your instance from the Internet, such as through a registered domain name or through an SSH or RDP connection from your local computer. You can also attach a static IP to the instance, which substitutes the public IP with an IP address that doesn't change even if the instance is stopped and started.

Q: What is a static IP?

A static IP is a fixed, public IP that is dedicated to your Lightsail account. You can assign a static IP to an instance, replacing its public IP. If you decide to replace your instance with another one, you can reassign the static IP to the new instance. In this way, you don't have to reconfigure any external systems (like DNS records) to point to a new IP every time you want to replace your instance.

Q: What are DNS records?

DNS is a globally distributed service that translates human readable names like www.example.com into numeric IP addresses, like 192.0.2.1 that computers use to connect to each other. With Lightsail, you can easily map your registered domain names such as photos.example.com to the public IPs of your Lightsail instances. In this way, when users type human readable names like example.com into their browsers, Lightsail automatically translates the address into the IP of the instance you want to direct your users to. Each of these translations is referred to as a DNS query.

It's important to know that in order to use a domain in Lightsail, you must first register it. You can register new domains using domain support features on Lightsail, Amazon Route 53, or your preferred DNS registrar.

Q: Can I manage firewall settings for my instance?

Yes. You can control the data traffic for your instances by using the Lightsail firewall. From the Lightsail console, you can set rules about which ports of your instance are publicly accessible for different types of traffic.

Q: What are snapshots?

Snapshots are point-in-time backups of instances, databases, or block storage disks. You can create a snapshot of your resources at any time, or you can enable automatic snapshots on

instances and disks to have Lightsail create snapshots for you. You can use snapshots as baselines to create new resources or to back up your data. A snapshot contains all of the data that is needed to restore your resource (from the moment when the snapshot was taken). When you restore a resource by creating it from a snapshot, the new resource begins as an exact replica of the original resource that was used to create the snapshot.

You can manually take snapshots of your Lightsail instances, disks, and databases, or you can use automatic snapshots to instruct Lightsail to take daily snapshots of your instances and disks automatically. For more information, see Snapshots in Amazon Lightsail.

Q: What are automatic snapshots?

Automatic snapshots are a way to schedule daily snapshots of your Linux/Unix instances in Amazon Lightsail. You can pick a time of the day, and Lightsail will automatically take a snapshot for you each day at the time you chose and always keep your seven most recent automatic snapshots. Enabling snapshots is free – you only pay for the actual storage used by your snapshots.

Q: What are the differences between manual snapshots and automatic snapshots?

Automatic snapshots cannot be tagged or exported directly to Amazon EC2. However, automatic snapshots can be copied and converted into manual snapshots. To copy an automatic snapshot into a manual one, choose Keep from the automatic snapshot's context menu to copy it as a manual snapshot.
Q: How can I back up my instances?

If you want to back up your data, you can use the Lightsail console or API to take a snapshot of your instance. If there is a failure or bad code deployment, you can later use your instance snapshot to create a brand new instance. We recommend stopping your instance temporarily when taking a snapshot, to ensure your data is complete and not corrupted in any way.

Q: What is the difference between stopping and deleting my instance?

When you stop your instance, it is powered down at its current state and is available for you to start again at any time. Stopping your instance will release its public IP, so it is recommended that you use static IPs for instances that must retain the same IP after they stop.

When you delete your instance, you are performing a destructive action. Unless you have taken an instance snapshot, all of your instance data will be lost and you cannot recover it again. The instance's public and private IPs will also be released. If you were using a static IP with that instance, the static IP is detached, but remains in your account.

Q: Can I upgrade my plan?

Yes. You can take a snapshot of your instance, and use the API to launch a new, larger size instance. You can launch new instances from snapshot using the Lightsail console or the CLI. Find instructions on how to use the CLI here.

Q: How can I connect Lightsail instances to other resources in my AWS account?

You can connect your Lightsail instances to VPC resources in your AWS account privately, by using VPC peering. Just choose Enable VPC peering on your Lightsail account page, and

Lightsail does the work for you. Once VPC peering is enabled, you can address other AWS resources in your default AWS VPC by using their private IPs. Find instructions here.

Note that you need to have a default VPC set up in your AWS account in order for VPC peering with Lightsail to work. AWS accounts created before December 2013 do not have a default VPC, and you will need to set one up. Find out more about setting up your default VPC here.

Q: In which regions is Lightsail available?

Lightsail is currently available in all the Availability Zones in the following AWS Regions:

US East (N. Virginia)
US East (Ohio)
US West (Oregon)
Canada (Central)
Europe (Frankfurt)
Europe (Ireland)
Europe (London)
Europe (Paris)
Europe (Stockholm)
Asia Pacific (Mumbai)
Asia Pacific (Singapore)
Asia Pacific (Sydney)
Asia Pacific (Tokyo)
Asia Pacific (Seoul)

Q: What are Availability Zones?

Availability Zones are collections of data centers that run on physically distinct, independent infrastructure and are engineered to be highly reliable. Common points of failure such as generators and cooling equipment are not shared between Availability Zones. Additionally, Availability Zones are physically separate, so that even extremely uncommon disasters such as fires, tornados, or flooding can affect only a single Availability Zone.

Q: What are the Lightsail service quotas?

For the latest Lightsail service quotas, including which quotas can be increased, see the Lightsail service quotas in the AWS General Reference. If you need to increase a quota, please open a case with customer service.

Q: How can I get more help?

We're here for you. Quick Assist in Lightsail offers immediate helpful tips about your actions in the console. From the Lightsail console, you can also access a library of getting started guides, overviews, and how-to topics. And if you want to use the API or CLI, Lightsail has a full API reference for all supported programming languages. You can also use Lightsail support resources:

- If you have an issue with your account or billing, contact customer service online. You get free 24x7 access with your Lightsail account.
- If you have a general question about how to use Lightsail, search the Lightsail documentation and support forums.

Additionally, AWS Support offers an array of paid plans to cover your individual needs.

Q: What operating systems can I use with Amazon Lightsail?

Lightsail currently supports 6 Linux or Unix-like distributions – Amazon Linux, Debian, FreeBSD, OpenSUSE, and Ubuntu – and 3 Windows Server versions – 2012 R2, 2016, and 2019.

Q: What are tags?

A tag is a label that you assign to a Lightsail resource. Each tag consists of a key and a value, both of which you define. A tag value is optional, so you can choose to create “key-only” tags for filtering resources in the Lightsail console.

Q: What can I use tags to do in Lightsail?

Tags have multiple use cases - they enable you to group and filter your resources in the Lightsail console and API, track and organize your costs in your bill, and regulate who can see or modify your resources through access management rules. By tagging your resources you can:

- Organize - use the Lightsail console and API filters to view and manage resources based on their tags you have assigned them. This is useful when you have many resources of the same type—you can quickly identify a specific resource based on the tags you've assigned to it.
- Cost-allocate - track and allocate costs across different projects or users by tagging your resources and creating “cost allocation tags” in the billing console. For instance, you can split out your bill and understand your costs by project or by client.
- Manage access - control how users with access to your AWS account can edit, create, and delete Lightsail resources by using AWS Identity and Access Management policies. This allows you to more easily collaborate with others without needing to give them full access to your Lightsail resources.

Q: What resources can be tagged?

Lightsail current supports tagging for the following resources:

- Instances (Linux and Windows)
- Container services
- Block storage disks
- Load balancers
- Databases
- DNS zones
- Instance, disk, and database snapshots

Manual snapshots also support tags and are automatically given the same tags as the source resource. You can edit these tags when you use a snapshot to create a new instance, disk, or database.

Q: How can I tag my Lightsail snapshots?

The Lightsail console automatically tags manual snapshots with the same tags as its parent resource. However, tags are not automatically copied over from a resource to its automatic snapshots. If you use the Lightsail API to create a snapshot, you can choose the tags for the snapshot yourself.

Important: Database snapshots tags are not currently included in billing reports (cost allocation tags).

Q: What is the difference between key-value and key-only Tags?

Lightsail tags are key-value pairs, allowing you to organize resources such as instances across different categories, e.g. project:Blog, project:Game, project:Test. This allows you full control across all use cases such as resource organization, bill reporting, and access management. The Lightsail console also allows you to tag your resources with key-only tags for quick filtering in the console.

Q: Does Amazon Lightsail support monitoring and alerting?

Yes. With Amazon Lightsail you can collect metrics on various resources, including instances, load balancers and databases. For any individual resource you can setup to two alarms thresholds for each metric. If the alarm threshold is breached you will receive a notification in the Lightsail console, and, optionally, you can choose to receive an email message and/or SMS message. There is no additional charge for the alerting and monitoring feature in Lightsail, however, you may incur charges from your mobile carrier for SMS messaging.

Lightsail block storage

Q: What can I do with Lightsail block storage?

Lightsail block storage provides additional storage volumes (called “attached disks” in Lightsail) that you can attach to your Lightsail instance, similar to an individual hard drive. Attached disks are useful for applications or software that need to separate out specific data from their core service and to protect application data in case of a failure or other issue with your instance and system disk. Attached disks offers consistent performance and low latency needed for applications or software that frequently access their stored data.

Lightsail block storage uses solid-state drives (SSD). This type of block storage balances a low price and good performance and is intended to support the vast majority of workloads that run on Lightsail. For customers with applications that require sustained IOPS performance, high amounts of throughput per disk, or that are running large databases like MongoDB, Cassandra, etc., we recommend using EC2 with GP2 or Provisioned IOPS SSD storage instead of Lightsail.

Q: How large can I make my attached disk?

Each attached disk can be up to 16 TB.

Q: How many disks can I attach per Lightsail instance?

You can attach up to 15 disks per Lightsail instance.

Q: Can I attach a disk to more than 1 instance?

No, disks can only be attached to one instance at a time.

Q: Does my disk need to be attached to an instance?

No, you can choose not to attach a disk to an instance. The disk will remain in your account in an unattached state. There is no difference in price if your disk is not attached to an instance.

Q: Can I increase the size of my attached disk?

Yes, you can increase the size of a disk by taking a disk snapshot and then creating a new, larger disk from snapshot.

Q: Does Lightsail block storage offer encryption?

Yes, to help keep your data secure, all Lightsail attached disks and disk snapshots are encrypted at rest by default, using keys that Lightsail manages on your behalf. Lightsail also provides encryption of data as it moves between Lightsail instances and attached disks.

Q: What availability can I expect from Lightsail block storage?

Lightsail block storage is designed to be highly available and reliable. Each attached disk is automatically replicated within its Availability Zone to protect you from component failure. Although Lightsail does not guarantee SLAs, Lightsail block storage disks are designed for 99.99% availability and for an annual failure rate of less than 0.2%. Lightsail also supports disk snapshots to allow regular backups of your data.

Q: How do I back up my attached disk?

You can back up your disk by taking a disk snapshot. You can also backup your entire instance and any attached disks by taking an instance snapshot.

Q: How are attached disks different than the storage included in my Lightsail plan?

The system disk included with your Lightsail plan is your instance's root device. If you terminate your instance, the system disk will be deleted as well. If you experience an instance failure, the system disk could be impacted. You also cannot detach your system disk or back it up separately from your instance. Data stored on an attached disk persists independently of the instance. Attached disks can be detached and moved between instances and can be backed up independently from an instance using disk snapshot. To protect your data, we recommend that you use your Lightsail instance's system disk only for temporary data. For data requiring a higher level of durability, we recommend using attached disks and regularly backing up your disk using disk or instance snapshots.

Lightsail load balancer

Q: What can I do with Lightsail load balancers?

Lightsail load balancers allow you to build highly available websites and applications. By distributing traffic across instances in different Availability Zones and pointing traffic to only healthy target instances, Lightsail load balancers reduce the risk of your application going down due to an issue with your instance or to a datacenter outage. With Lightsail load balancers and multiple target instances, your website or application can also accommodate increases in web traffic and maintain good performance for your visitors during peak load times.

In addition, you can use Lightsail load balancers to build secure applications and accept HTTPS traffic. Lightsail takes the complexity out of requesting, provisioning, and maintaining SSL/TLS certificates. The built-in certificate management requests and renews certificates on your behalf and adds the certificate to your load balancer automatically.
Q: Can I use load balancers with instances in different AWS Regions or different Availability Zones?

You cannot use load balancers with instances running in different AWS Regions. You can, however, use target instances across different Availability Zones with your load balancer. In fact, we recommend that you distribute your target instances across Availability Zones to maximize the availability of your application.

Q: How does my Lightsail load balancer deal with traffic spikes?

Lightsail load balancers scale automatically to handle traffic spikes to your application without you having to manually adjust them. If your application experiences a transient spike in traffic, your Lightsail load balancer will automatically scale and continue to efficiently direct traffic to your Lightsail instances. While your Lightsail load balancer is designed to easily manage traffic spikes, applications that consistently experience very high volume levels of traffic may experience performance degradation or throttling. If you expect your application consistently to manage more than 5 GB/hour of data or consistently to have a large number of connections (>400k new connections/hour, >15k active, concurrent connections), we recommend using Amazon EC2 with Application Load Balancing instead.

Q: How do Lightsail load balancers route traffic to my target instances?

Lightsail load balancers direct traffic to your healthy target instances based on a round robin algorithm.

Q: How does Lightsail know if my target instances are healthy?

During load balancer creation, you will be asked to specify a path (a common file or webpage URL) for Lightsail to ping. If the target instance can be reached using this path, then Lightsail will route traffic there. If one of your target instances is unresponsive, Lightsail will not route traffic to that instance. You can update the Health check path if needed in the load balancer management screens.

Q: What is the difference between key-value and key-only Tags?

Lightsail tags are key-value pairs, allowing you to organize resources such as instances across different categories, e.g. project:Blog, project:Game, project:Test. This allows you full control across all use cases such as resource organization, bill reporting, and access management. The Lightsail console also allows you to tag your resources with key-only tags for quick filtering in the console.

Q: Can I assign one instance to multiple load balancers?

Yes, Lightsail supports adding instances as target instances for more than one load balancer, if desired.

Q: What happens to my target instances when I delete my load balancer?

If you delete your load balancer, the attached target instances will continue to run normally and will appear in the Lightsail console as regular Lightsail instances. Please note that you will likely need to update your DNS records to direct traffic to one of your former target instances after you delete the load balancer.

Q: What is session persistence?

Session persistence enables the load balancer to bind a visitor's session to a specific target instance. This ensures that all requests from the user during the session are sent to the same target instance. Lightsail supports session persistence for applications that require visitors to hit the same target instances for data consistency. For example, many applications that require user authentication can benefit from using session persistence. You can turn on session persistence for specific load balancer from the load balancer management screens after creation.

Q: What kind of connections do Lightsail load balancers support?

Lightsail load balancers support HTTP and HTTPS connections.

Certificate management

Q: How can I use Lightsail-provisioned certificates?

SSL/TLS certificates are used to establish the identity of your website or application and secure connections between browsers and your website. Lightsail provides a signed certificate to use with your load balancer, and the load balancer provides SSL/TLS termination before routing verified traffic to your target instances over the secure AWS network. Lightsail certificates can only be used with Lightsail load balancers, not with individual Lightsail instances.

Q: How do I validate my certificate?

Lightsail certificates are domain validated, meaning that you need to provide proof of identity by validating that you own or have access to your website’s domain before the certificate can be provisioned by the certificate authority. When you request a new certificate, Lightsail will prompt you to add a CNAME to the DNS zone(s) of the domain or domains you are validating. You will add this CNAME wherever you currently manage your DNS zones – either Lightsail DNS management or an external DNS hosting provider (e.g., Route 53, GoDaddy, Namecheap, etc.). Once your certificate is validated, you can remove the CNAME record from your DNS zone, if desired.

Q: What happens if I cannot validate my domain?

You must be able to validate that you own a domain for security purposes. This means if you or someone in your organization cannot add a DNS record to validate your certificate for any reason, you will not be able to use an HTTPS-enabled load balancer with Lightsail.

Q: How many domains and subdomains can I add to my certificate?

You can add up to 10 domains or subdomains per certificate. Lightsail does not currently support wild card domains.

Q: How can I change the domains associated with my certificate?

To change the domains (add/delete) associated with your certificate, you will need to resubmit the certificate and revalidate your ownership of the domain(s). Follow the steps in the certificate management screens to regenerate your certificate and add or remove domains when prompted.

Q: How do I renew my certificate?

Lightsail provides managed renewal for your SSL/TLS certificates. This means that Lightsail tries to renew the certificates automatically before they expire with no action required from you. Your Lightsail certificate must be actively associated load balancer before it can be automatically renewed.

Q: What happens to my certificate when I delete my load balancer?

If your load balancer is deleted, your certificate is deleted as well. If you need to use a certificate for the same domain(s) in the future, you will need to request and validate a new certificate.

Q: Can I download my certificate provided by Lightsail?

No, Lightsail certificates are bound to your Lightsail account and cannot be removed and used outside of Lightsail.

Upgrade to EC2

Q: What is Upgrade to EC2?

Upgrade to EC2 is a feature that allows you to create a copy of your Lightsail instance in Amazon EC2. When you upgrade to EC2, you can pick among the wide set of instance types, configurations, and pricing models that EC2 offers, and have even more fine-tuned control over your networking, storage, and compute environment.

Q: Why would I want to upgrade to EC2?

Lightsail offers you an easy way to run and scale a wide set of cloud-based applications, at a bundled, predictable, and low price. Lightsail also automatically sets up your cloud environment configurations such as networking and access management.

Upgrading to EC2 allows you to run your application on a wider set of instance types, ranging from virtual machines with more CPU power, memory, and networking capabilities, to specialized or accelerated instances with FPGAs and GPUs. In addition, EC2 performs less

automatic management and set-up, allowing you more control over how you configure your cloud environment, such as your VPC.

Q: How does it work?

To get started, you need to export your Lightsail instance manual snapshot. You'll then use the Upgrade to EC2 wizard to create an instance in EC2.

Customers who are comfortable with EC2 can then use the EC2 creation wizard or API to create a new EC2 instance as they would from an existing EC2 AMI. Alternatively, Lightsail also provides a guided Lightsail console experience to help you easily create a new EC2 instance.

Note: Exporting Ghost and Django instance manual snapshots to Amazon EC2 is not supported at this time. We apologize for the inconvenience.

Q: How am I billed?

Using the Upgrade to EC2 feature is free. Once you have exported your snapshots to EC2, you will be charged for the EC2 image separately and in addition to your Lightsail snapshot. Any new EC2 instances you launch will also be billed by EC2, including their EBS storage volume(s) and data transfer. Refer to the EC2 pricing page for details on the pricing for your new instance and resources. Lightsail resources that continue to run in your Lightsail account will continue to be billed at their regular rates until they are deleted.

Q: Can I export managed database or disk snapshots?

The upgrade feature allows you to export Lightsail disk manual snapshots but doesn't currently support managed database snapshots. Disk snapshots can be rehydrated as EBS volumes from the EC2 console or API.

Q: What Lightsail resources can I upgrade?

Lightsail's upgrade to EC2 feature is designed to support the export of Linux and Windows instances and their attached block storage (if applicable) to EC2. It also supports the export of unattached block storage disks to EBS. It does not currently support the export of load balancers, databases, static IPs or DNS records.

Managed databases

Q: What are Lightsail's managed databases?

Lightsail's managed databases are instances that are dedicated to running databases, instead of other workloads like webserver, mail server, etc. A Lightsail database can contain multiple user-created databases, and you can access it by using the same tools and applications that you use with a stand-alone database. Lightsail maintains the security and health of your database’s underlying infrastructure and operating system, so that you can run a database without deep expertise in infrastructure management.

Like regular Lightsail instances, Lightsail databases come with a fixed amount of memory, computing power, and SSD based storage in their plans that you can scale up over time. Lightsail will automatically install and configure your chosen database for you upon creation.

Q: What can I do with Lightsail's managed databases?

Lightsail managed databases provide an easy, low maintenance way to store your data in the cloud. You can run Lightsail databases either as a new databases or by migrating from an existing on-premises or hosted database to Lightsail.

They can also allow you to scale your application to accept larger amounts of traffic and more intensive loads, by separating out your database into a dedicated instance. Lightsail databases are especially useful for stateful applications – like WordPress and most common CMSs – that need data to be kept in sync when you scale beyond a single instance. Lightsail databases can be paired with a Lightsail load balancer and two or more Lightsail instances to create a powerful, scaled application. By using Lightsail High Availability plans, you can also create add redundancy to your database, helping to ensure high uptime for your application.

Q: What does Lightsail manage for me on my managed database?

Lightsail manages a range of maintenance activities and security for your database and its underlying infrastructure. Lightsail automatically backs up your database and allows point in time restore from the past 7 days using the database restore tool, to help protect against data loss or component failure. Lightsail also automatically encrypts your data at rest and in motion for increased security and stores your database password for easy and secure connections to your database. On the maintenance side, Lightsail runs maintenance on your database during your set maintenance window. This maintenance include automatic upgrades to the latest minor database version and all management of the underlying infrastructure and operating system.

Q: What managed database plans does Lightsail offer?

Lightsail offers 4 sizes of databases in Standard and High Availability plans. Each plan comes with a fixed amount of storage and a monthly allowance of data transfer. You can also scale up to larger plans over time, as needed, and switch between Standard and High Availability plans. High Availability plans mirror the same resources as Standard plans and additionally include a standby database running in a separate Availability Zones from your primary database for redundancy.

Q: What is a High Availability plan?

Managed databases are available in Standard and High Availability plans. Standard and High availability plans have identical plan resources, including memory, storage, and data transfer allowance. High Availability plans add redundancy and durability to your database, by automatically creating standby database in a separate Availability Zone from your primary database, synchronously replicating data to the standby database, and providing failover to the standby database in case of infrastructure failure and during maintenance so that you ensure uptime even when databases is being automatically upgraded/maintained by Lightsail. Use High Availability plans for running production applications or software where high uptime is required.

Q: How do I scale up or down my managed database?

You can scale up your database by taking a snapshot of your database and creating a new, larger database plan from snapshot or by creating a new, larger database using the emergency restore feature. You can also switch from Standard to High Availability plans and vice versa using either method. You cannot scale down your database. To learn more, see Creating a database from a snapshot in Amazon Lightsail.

Q: How can I back up my managed database?

Lightsail backs up your data automatically and allows restore of this data from a specific point in time to a new database. Automatic backup is a free service for your database but only saves the last 7 days of data. If you delete your database, all automatic backup records are deleted and point in time restore is no longer possible. To retain backups of data after deleting your database or to retain a backup for more than 7 days in the past, use manual snapshots.

You can take manual snapshots of your Lightsail managed databases from the database management pages. Manual snapshots contain all the data from your database and can be used as backups for data that you want to store permanently. You can also use manual snapshots to create a new, larger database or to switch between Standard and High Availability plans. Manual snapshots are stored until you delete them and are billed at $0.05 USD/GB-month.

Q: What happens to my data if I delete my managed database?

If you delete your managed database, both your database itself and all automatic backups will be deleted. There is no way to recover this data unless you take a manual snapshot before deleting your database. During deletion of your database, Lightsail provides a one-click option to take a manual snapshot, if desired, to help protect against accidental loss of data. Taking a manual snapshot before deletion is optional but highly recommended. You can delete your manual snapshot in the future when you no longer need the stored data.

Q: How do managed databases work with my Lightsail instances?

After you create your managed database, you can start using your database with your application immediately, using your Lightsail instances as web servers or other dedicated workloads for your app. To connect your Lightsail instance to a database, use your database endpoint and reference your securely stored password to configure the database as your data store in the code of your application. You can find connection data in the database management screens. The file name and location for your database configuration file will vary by application. Note that you can connect many instances to one database, either using the same tables or using different ones.

Lightsail CDN distributions

Q: What can I do with Lightsail CDN distributions?

Lightsail content delivery network (CDN) distributions make it easy for you to accelerate the delivery of content hosted on your Lightsail resources by storing and serving it on Amazon's global delivery network, powered by Amazon CloudFront. Distributions also help you enable your website to support HTTPS traffic by providing simple SSL certificate creation and hosting. Finally, distributions can help reduce the load on your Lightsail resources and help your website handle large traffic spikes. Like all of Lightsail's features, setup can be completed with just a few clicks, and you pay a simple monthly price.

Q: What types of resources can I use as the origin of my distributions?

Lightsail distributions allow you to use your Lightsail instances, containers, load balancers, or Lightsail buckets as origins.

Q: Do I need to attach a static IP to my Lightsail instance in order to use it as an origin for my Lightsail distribution?

Yes, static IPs are required to be attached to instances that are specified as origins.

Q: How do I setup a Lightsail distribution with my WordPress website?

Simply create your distribution, select your WordPress instance as the origin, choose your plan, and you're all set. Lightsail distributions automatically configure your distribution settings to optimize performance for most WordPress configurations.

Q: Can I attach multiple origins?

Although you cannot attach multiple origins to your Lightsail distribution, you can attach multiple instances to a Lightsail load balancer and specify it as the origin of your distribution.

Q: Do Lightsail distributions support certificate creation?

Yes. Lightsail distributions makes it easy to create, verify, and attach certificates directly from your distribution's management page.

Q: Is a certificate required?

A certificate is only required if you wish to use your custom domain name with your distribution. All Lightsail distributions are created with a unique Amazon CloudFront domain name that is HTTPS-enabled. However, if you wish to use your custom domain with your distribution, then you need to attach a certificate for your custom domain to your distribution.

Q: Is there a limit to the number of certificates I can create?

Yes, refer to Lightsail's service quotas for more information.

Q: How can I configure my distribution to redirect HTTP requests to HTTPS?

Lightsail distributions automatically redirect all HTTP requests to HTTPS to ensure that your content is served securely.

Q: How can I configure my apex domain to point to my Lightsail distribution?

In order to point your apex domain to your CDN distribution, you must create an ALIAS record in the domain name system (DNS) of your domain that maps your apex domain to your distribution's default domain. If your DNS hosting provider does not support ALIAS records, you can use Lightsail DNS zones to easily configure your apex domain to point to your distribution's domain.

Q: What are the differences between Lightsail's instance data transfer quotas and distribution data transfer quotas?

While data transfer IN and OUT count toward your instance's data transfer quota, only data transfer OUT counts toward you distribution's quota. In addition, all data transfer OUT in excess of your distribution's quota is charged an overage fee, whereas some types of data transfer OUT are free for instances. Finally, Lightsail distributions use a different regional overage model, though the majority of the rates are the same as those charged for instance data transfer overage.

Q: Can I change the plan associated with my distribution?

Yes, you can change your distribution's plan once per month. If you wish to change your plan a second time, you must wait until the beginning of the following month to do so.

Q: How do I know if my distribution is working?

Lightsail distributions provide you with a variety of metrics that track the performance of your distribution, including the total number of requests your distribution has received, the amount of data your distribution has sent to clients and to your origin, and the percentage of requests that have resulted in errors. Additionally, you can create alerts that are linked to distribution metrics.

Q: When should I use Lightsail distributions versus Amazon CloudFront distributions?

Lightsail distributions are designed specifically for users who are hosting websites or web applications on Lightsail resources, such as instances and load balancers. If you're using another service in AWS to host your website or app, have complex configuration needs, or have a workload that involves a high number of requests per second or large amount of video streaming, we recommend that you use Amazon CloudFront.

Q: How is Lightsail CDN intended to be used?

Amazon Lightsail CDN distributions are created using fixed-priced bundles of data transfer to make the cost of using the service simple and predictable. Distribution bundles are designed to cover a month's worth of usage. Using distribution bundles in a way to avoid incurring overage fees (including, but not limited to, frequently upgrading or downgrading bundles, or using an excessively large number of distributions with a single origin) is beyond the intended scope of use and is not permitted. In addition, workloads that involve a high number of requests per second or large amount of video streaming are not permitted. Engaging in these behaviors may result in throttling or suspension of your data services or account.

Q: Can I move my Lightsail content delivery network (CDN) distribution to Amazon CloudFront?

Yes, you can move your Lightsail distribution by creating a similarly configured distribution in CloudFront. All of the settings that can be configured in a Lightsail distribution can also be configured in a CloudFront distribution. Complete the following steps to move your distribution to CloudFront:

1. Take a snapshot of your Lightsail instance that is configured as your distribution's origin. Export the snapshot to Amazon EC2, and then create a new instance from the snapshot in EC2. For more information, see Exporting Amazon Lightsail snapshots.

   Note: Create an application load balancer in Elastic Load Balancing if you need to load balance your website or web application. For more information, see the Elastic Load Balancing User Guide.

2. Disable custom domains for your Amazon Lightsail distribution to detach certificates that you might have attached to it. For more information, see Disabling custom domains for your Amazon Lightsail distributions.

3. Using the AWS Command Line Interface (CLI), run the get-distributions command to get a list of your Lightsail distribution's settings. For more information, see get-distributions in the AWS CLI Reference.

4. Sign in to the Amazon CloudFront console and create a distribution with the same configuration settings as your Lightsail distribution. For more information, see Creating a Distribution in the Amazon CloudFront Developer Guide.

5. Create a certificate in AWS Certificate Manager (ACM) that you will attach to your CloudFront distribution. For more information, see Request a Public Certificate in the ACM User Guide.

Update your CloudFront distribution to use the ACM certificate you created. For more information, see Updating your CloudFront distribution in the CloudFront User Guide.

Containers

Q: What can I do with Lightsail container services?

Lightsail container services provide an easy way to run containerized applications in the cloud. You can run a variety of applications on a container service, ranging from simple web apps to multi-tiered micro services. You just specify the container image, power (CPU, RAM) and scale (number of nodes) required for your container service. Lightsail takes care of running the container service without you having to manage any underlying infrastructure. Lightsail will provide you with a load balanced TLS endpoint to access the application running on the container service.

Q: Can Lightsail container service run Docker containers?

Yes. Lightsail supports Linux-based Docker containers. Windows containers are currently not supported.

Q: How do I use my Docker container images with Lightsail container service?

You can use either container images on registries like Docker Hub or build your own custom image and push it to Lightsail in a few easy steps using the AWS CLI. For more information, see Pushing and managing container images on your Amazon Lightsail container services.

Q: Can I pull my container images from a private container registry?

Currently, only public container registries are supported by Lightsail container services. Alternately, you can push your custom container images from your local machine to Lightsail to keep them private.

Q: Can I change the power and scale of my service based on demand?

Yes, container service power and scale can be changed at any time even after the service is created.

Q: Can I customize the name of the HTTPS endpoint created by Lightsail container service?

Lightsail provides a HTTPS endpoint for every container service in the format “<service-name>.<random-guid>.<aws-region-name>.cs.amazonlightsail.com. Only the service name can be customized. Alternately, you can use a custom domain name. For more information, see Enabling and managing custom domains for your Amazon Lightsail container services.

Q: Can I use custom domains for the HTTPS endpoint of a Lightsail container service?

Yes. You can create and attach SSL/TLS certificates with custom domain names to your container service in Lightsail. The certificates need to be domain validated. After the certificate is attached, if you manage your domain via Lightsail DNS, you can easily add A/AAAA/CNAME records to route traffic for your domain to your container services. Alternately, you can use a DNS hosting provider who supports adding ALIAS records, to map the apex of your domain to the public domain (public DNS) of your Lightsail container service. For more information, see Enabling and managing custom domains for your Amazon Lightsail container service.

Q: What do Lightsail container services cost?

Lightsail container services are billed on an on-demand hourly rate, so you pay only for what you use. For every Lightsail container service you use, we charge you the fixed hourly price, up to the maximum monthly service price. Maximum monthly service price can be calculated by multiplying the base price of the power of your service with the scale of your service. For example, a service of Micro power and scale of 2 will cost a maximum of $10*2=$20/month. The least expensive Lightsail container service starts at $0.0094 USD/hour ($7 USD/month). Additional data transfer charges may apply for usage above the free-quota of 500 GB per month for each service.

Q: Will I be charged for the whole month even if I run my container service for a few days?

Your Lightsail container services are charged only when they're in the running or disabled state. If you delete your Lightsail container service before the end of the month, we charge you a prorated cost based on the total number of hours that you used your Lightsail container service. For example, if you use your Lightsail container service with a power of Micro and scale of 1 for 100 hours in a month, you will be charged $1.34 ($0.0134*100).

Q: Will I be charged for data transfer in and out of the container service?

Every container service comes with a data transfer quota (500 GB per month). This counts towards both the data transfer IN and OUT of your service. When you exceed the quota, you will get charged for data transfer OUT from a Lightsail container service to the Internet or to another AWS Region or to AWS resources in the same Region when using public IP addresses. The charge for these types of data transfer above the free allowance is as follows:

- US East (N. Virginia): $0.09 USD/GB
- US East (Ohio): $0.09 USD/GB
- US West (Oregon): $0.09 USD/GB
- Canada (Central): $0.09 USD/GB
- Europe (Frankfurt): $0.09 USD/GB
- Europe (Ireland): $0.09 USD/GB
- Europe (London): $0.09 USD/GB
- Europe (Paris): $0.09 USD/GB
- Europe (Stockholm): $0.09 USD/GB
- Asia Pacific (Mumbai): $0.13 USD/GB
- Asia Pacific (Singapore): $0.12 USD/GB
- Asia Pacific (Sydney): $0.17 USD/GB
- Asia Pacific (Tokyo): $0.14 USD/GB
- Asia Pacific (Seoul): $0.13 USD/GB

Q: What is the difference between stopping and deleting my container service?

When you disable your container service, your container nodes are in a disabled state and the public endpoint of the service returns a HTTP status code ‘503’. Enabling the service restores it to the last active deployment. Power and scale configurations are also retained. Public endpoint

name does not change after re-enabling. Deployment history and container images are preserved.

When you delete your container service, you are performing a destructive action. All the container nodes of the service will be permanently deleted. The HTTPS public endpoint address, container images, deployment history, and logs associated with your service will also be permanently deleted. You will not be able to recover the endpoint address.

Q: Will I be charged if my container service is in a disabled state?

Yes, you are charged according to the power and scale configuration of your container service, even when it is in a disabled state.

Q: Can I use container services as the origin to my Lightsail content delivery network (CDN) distributions?

Container services are currently not supported as origins for Lightsail CDN distributions.

Q: Can I use container services as targets for my Lightsail load balancer?

No. Container services are currently not available as targets for Lightsail load balancers. However, the public endpoints of container services come with built-in load balancing.

Q: Can I configure the public endpoint of my container service to redirect HTTP requests to HTTPS?

Lightsail container service public endpoints automatically redirect all HTTP requests to HTTPS to ensure that your content is served securely.

Q: Do container services support monitoring and alerting?

Container services provide metrics for CPU utilization and memory utilization across the nodes of your service. Alerting based on these metrics is currently not supported.

Object storage

Q: What can I do with Lightsail object storage?

Lightsail now provides you with the ability to store your static content such as images, videos or HTML files in an object storage that can be used for your websites and applications. Lightsail object storage can be associated to your Lightsail CDN distribution with a few simple clicks, making it quick and easy to accelerate the delivery of your content to a global audience. It can also be used as a low cost, secure backup solution.

Q: What does Lightsail object storage cost?

Lightsail object storage has three different fixed-priced bundles in all AWS Regions where Lightsail is available. The first bundle is $1 per month and is free for the first 12 months. This bundle includes 5 GB storage capacity and 25 GB data transfer. The second bundle is $3 per month and includes 100 GB storage capacity and 250 GB data transfer. Lastly, the third bundle is $5 per month and includes 250 GB of storage capacity and 500 GB of data transfer. Lightsail object storage includes unlimited data transfer into your bucket, as the bundled data transfer allowance is used only for data transfer out from your bucket.

Q: Does Lightsail object storage have overage charges?

When you exceed the monthly storage capacity or data transfer allowance of your object storage plan, you will get charged for the additional amount.

Q: How does my data transfer allowance work with object storage?

You can consume your data transfer allowance by transfering data into and out of Lightsail object storage, except for the following:

- Data transferred into Lightsail object storage from the internet
- Data transfer between Lightsail object storage resources
- Data transferred out from Lightsail object storage to another Lightsail resource in the same region (including to a different account in the same AWS Region)
- Data transferred out from Lightsail object storage to a Lightsail CDN distribution

Q: Can I change the plan associated with my Lightsail bucket?

Yes, you can change your Lightsail object storage plan one time within your monthly AWS billing cycle.

Q: Can I copy objects from Lightsail object storage to Amazon S3?

Yes, copying from Lightsail object storage to Amazon S3 is supported. Documentation on the proper AWS CLI command can be found here

Q: How do I get started with Lightsail object storage?

To use Lightsail object storage, you must first create a bucket that is used to store your data. For more information, see Creating buckets in Amazon Lightsail. After your bucket is up and running, you can start adding objects to your bucket by uploading files using the Lightsail console or by configuring your application to put content like logs or other application data in the bucket. Alternatively, you can also get started with Lightsail object storage through the use of AWS Command Line Interface (AWS CLI).

Q: Can I block public access to my bucket?

Lightsail buckets and objects are set to private by default, meaning that only users with appropriate permissions have access to the bucket and objects. A user can change this default setting and either make individual objects public and read only in a private bucket or opt to make the entire bucket public and read only. When a user makes a bucket or object public, anyone in the world can read its content.

Q: How do I provide programmatic access to my bucket?

You can use either access keys or roles for programmatic access to your bucket. First, select the bucket you want to programmatically connect to in the Lightsail console. Second, under the “Permissions” tab, create an access key or assign a role to your Lightsail instance and then configure your website or application code to use your bucket. This behavior may vary depending on how you plan to use object storage with your website or application. See more documentation here.

Q: How do I share a bucket with other AWS accounts?

Lightsail makes cross-account sharing easy by allowing you to share access to your bucket with the AWS account ID you specify in the “Cross-account access” in the Lightsail console. After you specify an AWS account ID, that account will have read-only access to the bucket.

Q: What is versioning?

Versioning allows you to preserve, retrieve, and restore every version of every object storage in your bucket, providing an additional level of protection from accidental overwrites and deletes. See more documentation on enabling versioning here.

Q: How do I associate my Lightsail bucket to my Lightsail CDN distribution?

Lightsail object storage can be associated to Lightsail CDN distributions with a few simple clicks, making it quick and easy to accelerate the delivery of your content to a global audience. To do so, create a Lightsail CDN distribution and simply select the Lightsail bucket as the origin of your Lightsail CDN distribution. See more documentation here.

Q: Is there a limit to the number of Lightsail object storage bundles I can create?

You can create up to 20 Lightsail object storage bundles per account. There is no limit to the number of objects that you can store in a bucket.

Q: Does Lightsail object storage support monitoring and alerting?

With Lightsail object storage, customers can easily view metrics on the total used space within a bucket and number of objects within the bucket. Alerting based on these metrics is also supported.

AWS Lambda

Run code without thinking about servers or clusters.

With AWS Lambda, you can run code without provisioning or managing servers. You pay only for the compute time that you consume—there's no charge when your code isn't running. You can run code for virtually any type of application or backend service—all with zero administration. Just upload your code and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

How it works

AWS Lambda is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers. You can trigger Lambda from over 200 AWS services and software as a service (SaaS) applications, and only pay for what you use.

File processing

Use Amazon Simple Storage Service (Amazon S3) to trigger AWS Lambda data processing in real time after an upload, or connect to an existing Amazon EFS file

system to enable massively parallel shared access for large-scale file processing.

Stream processing

Use AWS Lambda and Amazon Kinesis to process real-time streaming data for application activity tracking, transaction order processing, clickstream analysis, data cleansing, log filtering, indexing, social media analysis, IoT device data telemetry, and metering.

Web applications

Combine AWS Lambda with other AWS services to build powerful web applications that automatically scale up and down and run in a highly available configuration across multiple data centers.

IoT backends

Build serverless backends using AWS Lambda to handle web, mobile, Internet of Things (IoT), and third-party API requests

Mobile backends

Build backends using AWS Lambda and Amazon API Gateway to authenticate and process API requests. Use AWS Amplify to easily integrate your backend with your iOS, Android, Web, and React Native frontends.

Use cases

Process data at scale
Execute code at the capacity you need, as you need it. Scale to match your data volume automatically and enable custom event triggers.

Run interactive web and mobile backends
Combine AWS Lambda with other AWS services to create secure, stable, and scalable online experiences.

Enable powerful ML insights
Preprocess data before feeding it to your machine learning (ML) model. With Amazon Elastic File System (EFS) access, AWS Lambda handles infrastructure management and provisioning to simplify scaling.

Create event-driven applications
Build event-driven functions for easy communication between decoupled services. Reduce costs by running applications during times of peak demand without crashing or over-provisioning resources.

What is AWS Lambda?

Lambda is a compute service that lets you run code without provisioning or managing servers. Lambda runs your code on a high-availability compute infrastructure and performs all of the administration of the compute resources, including server and operating system maintenance, capacity provisioning and automatic scaling, and logging. With Lambda, you can run code for virtually any type of application or backend service. All you need to do is supply your code in one of the languages that Lambda supports.

You organize your code into Lambda functions. Lambda runs your function only when needed and scales automatically, from a few requests per day to thousands per second. You pay only for the compute time that you consume—there is no charge when your code is not running.

You can invoke your Lambda functions using the Lambda API, or Lambda can run your functions in response to events from other AWS services. For example, you can use Lambda to:

- Build data-processing triggers for AWS services such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
- Process streaming data stored in Amazon Kinesis.
- Create your own backend that operates at AWS scale, performance, and security.

Lambda is a highly available service.

When should I use Lambda?

Lambda is an ideal compute service for many application scenarios, as long as you can run your application code using the Lambda standard runtime environment and within the resources that Lambda provides.

When using Lambda, you are responsible only for your code. Lambda manages the compute fleet that offers a balance of memory, CPU, network, and other resources to run your code. Because Lambda manages these resources, you cannot log in to compute instances or customize the operating system on provided runtimes. Lambda performs operational and administrative activities on your behalf, including managing capacity, monitoring, and logging your Lambda functions.

If you need to manage your own compute resources, AWS has other compute services to meet your needs. For example:

- Amazon Elastic Compute Cloud (Amazon EC2) offers a wide range of EC2 instance types to choose from. It lets you customize operating systems, network and security settings, and the entire software stack. You are responsible for provisioning capacity, monitoring fleet health and performance, and using Availability Zones for fault tolerance.
- AWS Elastic Beanstalk enables you to deploy and scale applications onto Amazon EC2. You retain ownership and full control over the underlying EC2 instances.

Lambda features

The following key features help you develop Lambda applications that are scalable, secure, and easily extensible:

Concurrency and scaling controls

Concurrency and scaling controls such as concurrency limits and provisioned concurrency give you fine-grained control over the scaling and responsiveness of your production applications.

Functions defined as container images

Use your preferred container image tooling, workflows, and dependencies to build, test, and deploy your Lambda functions.

Code signing

Code signing for Lambda provides trust and integrity controls that let you verify that only unaltered code that approved developers have published is deployed in your Lambda functions.

Lambda extensions

You can use Lambda extensions to augment your Lambda functions. For example, use extensions to more easily integrate Lambda with your favorite tools for monitoring, observability, security, and governance.

Function blueprints

A function blueprint provides sample code that shows how to use Lambda with other AWS services or third-party applications. Blueprints include sample code and function configuration presets for Node.js and Python runtimes.

Database access

A database proxy manages a pool of database connections and relays queries from a function. This enables a function to reach high concurrency levels without exhausting database connections.

File systems access

You can configure a function to mount an Amazon Elastic File System (Amazon EFS) file system to a local directory. With Amazon EFS, your function code can access and modify shared resources safely and at high concurrency.

Related services

Lambda integrates with other AWS services to invoke functions based on events that you specify. For example:

- Use API Gateway to provide a secure and scalable gateway for web APIs that route HTTP requests to Lambda functions.

- For services that generate a queue or data stream (such as DynamoDB and Kinesis), Lambda polls the queue or data stream from the service and invokes your function to process the received data.
- Define Amazon S3 events that invoke a Lambda function to process Amazon S3 objects, for example, when an object is created or deleted.
- Use a Lambda function to process Amazon SQS messages or Amazon Simple Notification Service (Amazon SNS) notifications.
- Use AWS Step Functions to connect Lambda functions together into serverless workflows called state machines.

AWS Lambda FAQs

General

Q: What is AWS Lambda?

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume - there is no charge when your code is not running. With Lambda, you can run code for virtually any type of application or backend service - all with zero administration. Just upload your code, and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

Q: What is serverless computing?

Serverless computing allows you to build and run applications and services without thinking about servers. With serverless computing, your application still runs on servers, but all the server management is done by AWS. At the core of serverless computing is AWS Lambda, which lets you run your code without provisioning or managing servers.

Q: What events can trigger an AWS Lambda function?

Please see our documentation for a complete list of event sources.

Q: When should I use AWS Lambda versus Amazon EC2?

Amazon Web Services offers a set of compute services to meet a range of needs.

Amazon EC2 offers flexibility, with a wide range of instance types and the option to customize the operating system, network and security settings, and the entire software stack, allowing you to easily move existing applications to the cloud. With Amazon EC2 you are responsible for provisioning capacity, monitoring fleet health and performance, and designing for fault tolerance and scalability. AWS Elastic Beanstalk offers an easy-to-use service for deploying and scaling web applications in which you retain ownership and full control over the underlying EC2 instances. Amazon EC2 Container Service is a scalable management service that supports Docker containers and allows you to easily run distributed applications on a managed cluster of Amazon EC2 instances.

AWS Lambda makes it easy to execute code in response to events, such as changes to Amazon S3 buckets, updates to an Amazon DynamoDB table, or custom events generated by your applications or devices. With Lambda, you do not have to provision your own instances; Lambda performs all the operational and administrative activities on your behalf, including capacity provisioning, monitoring fleet health, applying security patches to the underlying compute resources, deploying your code, running a web service front end, and monitoring and logging your code. AWS Lambda provides easy scaling and high availability to your code without additional effort on your part.

Q: What kind of code can run on AWS Lambda?

AWS Lambda offers an easy way to accomplish many activities in the cloud. For example, you can use AWS Lambda to build mobile back-ends that retrieve and transform data from Amazon DynamoDB, handlers that compress or transform objects as they are uploaded to Amazon S3, auditing and reporting of API calls made to any Amazon Web Service, and server-less processing of streaming data using Amazon Kinesis.

Q: What languages does AWS Lambda support?

AWS Lambda natively supports Java, Go, PowerShell, Node.js, C#, Python, and Ruby code, and provides a Runtime API which allows you to use any additional programming languages to author your functions. Please read our documentation on using Node.js, Python, Java, Ruby, C#, Go, and PowerShell.

Q: Can I access the infrastructure that AWS Lambda runs on?

No. AWS Lambda operates the compute infrastructure on your behalf, allowing it to perform health checks, apply security patches, and do other routine maintenance.

Q: How does AWS Lambda isolate my code?

Each AWS Lambda function runs in its own isolated environment, with its own resources and file system view. AWS Lambda uses the same techniques as Amazon EC2 to provide security and separation at the infrastructure and execution levels.

Q: How does AWS Lambda secure my code?

AWS Lambda stores code in Amazon S3 and encrypts it at rest. AWS Lambda performs additional integrity checks while your code is in use.
Q: What AWS regions are available for AWS Lambda?

Please refer to the AWS Global Infrastructure Region Table.

AWS Lambda functions

Q: What is an AWS Lambda function?

The code you run on AWS Lambda is uploaded as a “Lambda function”. Each function has associated configuration information, such as its name, description, entry point, and resource requirements. The code must be written in a “stateless” style i.e. it should assume there is no affinity to the underlying compute infrastructure. Local file system access, child processes, and

similar artifacts may not extend beyond the lifetime of the request, and any persistent state should be stored in Amazon S3, Amazon DynamoDB, Amazon EFS, or another Internet-available storage service. Lambda functions can include libraries, even native ones.

Q: Will AWS Lambda reuse function instances?

To improve performance, AWS Lambda may choose to retain an instance of your function and reuse it to serve a subsequent request, rather than creating a new copy. To learn more about how Lambda reuses function instances, visit our documentation. Your code should not assume that this will always happen.

Q: What if I need scratch space on disk for my AWS Lambda function?

You can configure each Lambda function with its own ephemeral storage between 512MB and 10,240MB, in 1MB increments. The ephemeral storage is available in each function's /tmp directory.

Each function has access to 512MB of storage at no additional cost. When configuring your functions with more than 512MB of ephemeral storage, you will be charged based on the amount of storage you configure, and how long your function runs, metered in 1ms increments. For comparison, in the US East (Ohio) region, the AWS Fargate ephemeral storage price is $0.000111 per GB-hour, or $0.08 per GB-month. Amazon EBS gp3 storage volume pricing in US East (Ohio) is $0.08 per GB-month. AWS Lambda ephemeral storage pricing is $0.0000000309 per GB-second, or $0.000111 per GB-hour and $0.08 per GB-month.

Q: How do I configure my application to use AWS Lambda ephemeral storage?

You can configure each Lambda function with its own ephemeral storage between 512MB and 10,240MB, in 1MB increments by using the AWS Lambda console, AWS Lambda API, or AWS CloudFormation template during function creation or update.

Q: Is AWS Lambda ephemeral storage encrypted?

Yes. All data stored in ephemeral storage is encrypted at rest with a key managed by AWS.

Q: What metrics can I use to monitor my AWS Lambda ephemeral storage usage?

You can use AWS CloudWatch Lambda Insight metrics to monitor your ephemeral storage usage. To learn more, see the AWS CloudWatch Lambda Insights documentation.

Q: When should I use Amazon S3, Amazon EFS, or AWS Lambda ephemeral storage for my serverless applications?

If your application needs durable, persistent storage, consider using Amazon S3 or Amazon EFS. If your application requires storing data needed by code in a single function invocation, consider using AWS Lambda ephemeral storage as a transient cache. To learn more, please see Choosing between AWS Lambda data storage options in web apps.

Q: Can I use ephemeral storage while Provisioned Concurrency is enabled for my function?

Yes. However, if you application needs persistent storage, consider using Amazon EFS or Amazon S3. When you enable Provisioned Concurrency for your function, your function's initialization code runs during allocation and every few hours, as running instances of your function are recycled. You can see the initialization time in logs and traces after an instance processes a request. However, initialization is billed even if the instance never

processes a request. This Provisioned Concurrency initialization behavior may affect how your function interacts with data you store in ephemeral storage, even when your function isn't processing requests. To learn more about Provisioned Concurrency, please see the relevant documentation.

Q: How do I configure my application to use AWS Lambda ephemeral storage?

You can configure each Lambda function with its own ephemeral storage between 512MB and 10,240MB, in 1MB increments by using the AWS Lambda console, AWS Lambda API, or AWS CloudFormation template during function creation or update.

Q: Is AWS Lambda ephemeral storage encrypted?

Yes. All data stored in ephemeral storage is encrypted at rest with a key managed by AWS.
Q: What metrics can I use to monitor my AWS Lambda ephemeral storage usage?

You can use AWS CloudWatch Lambda Insight metrics to monitor your ephemeral storage usage. To learn more, see the AWS CloudWatch Lambda Insights documentation.

Q: Why must AWS Lambda functions be stateless?

Keeping functions stateless enables AWS Lambda to rapidly launch as many copies of the function as needed to scale to the rate of incoming events. While AWS Lambda's programming model is stateless, your code can access stateful data by calling other web services, such as Amazon S3 or Amazon DynamoDB.

Q: Can I use threads and processes in my AWS Lambda function code?

Yes. AWS Lambda allows you to use normal language and operating system features, such as creating additional threads and processes. Resources allocated to the Lambda function, including memory, execution time, disk, and network use, must be shared among all the threads/processes it uses. You can launch processes using any language supported by Amazon Linux.

Q: What restrictions apply to AWS Lambda function code?

Lambda attempts to impose as few restrictions as possible on normal language and operating system activities, but there are a few activities that are disabled: Inbound network connections are blocked by AWS Lambda, and for outbound connections, only TCP/IP and UDP/IP sockets are supported, and ptrace (debugging) system calls are blocked. TCP port 25 traffic is also blocked as an anti-spam measure.

Q: How do I create an AWS Lambda function using the Lambda console?

If you are using Node.js or Python, you can author the code for your function using code editor in the AWS Lambda console, which lets you author and test your functions, and view the results of function executions in a robust, IDE-like environment. Go to the console to get started.

You can also package the code (and any dependent libraries) as a ZIP and upload it using the AWS Lambda console from your local environment or specify an Amazon S3 location where the ZIP file is located. Uploads must be no larger than 50MB (compressed). You can use the AWS

Eclipse plugin to author and deploy Lambda functions in Java. You can use the Visual Studio plugin to author and deploy Lambda functions in C#, and Node.js.

Q: How do I create an AWS Lambda function using the Lambda CLI?

You can package the code (and any dependent libraries) as a ZIP and upload it using the AWS CLI from your local environment, or specify an Amazon S3 location where the ZIP file is located. Uploads must be no larger than 50MB (compressed). Visit the Lambda Getting Started guide to get started.

Q: Does AWS Lambda support environment variables?

Yes. You can easily create and modify environment variables from the AWS Lambda Console, CLI, or SDKs. To learn more about environment variables, see the documentation.

Q: Can I store sensitive information in environment variables?

For sensitive information, such as database passwords, we recommend you use client-side encryption using AWS Key Management Service and store the resulting values as ciphertext in your environment variable. You will need to include logic in your AWS Lambda function code to decrypt these values.

Q: How can I manage my AWS Lambda functions?

You can easily list, delete, update, and monitor your Lambda functions using the dashboard in the AWS Lambda console. You can also use the AWS CLI and AWS SDK to manage your Lambda functions. Visit the Lambda Developer Guide to learn more.

Q: Can I share code across functions?

Yes, you can package any code (frameworks, SDKs, libraries, and more) as a Lambda Layer and manage and share them easily across multiple functions.

Q: How do I monitor an AWS Lambda function?

AWS Lambda automatically monitors Lambda functions on your behalf, reporting real-time metrics through Amazon CloudWatch, including total requests, account-level and function-level concurrency usage, latency, error rates, and throttled requests. You can view statistics for each of your Lambda functions via the Amazon CloudWatch console or through the AWS Lambda console. You can also call third-party monitoring APIs in your Lambda function.

Q: How do I troubleshoot failures in an AWS Lambda function?

AWS Lambda automatically integrates with Amazon CloudWatch logs, creating a log group for each Lambda function and providing basic application lifecycle event log entries, including logging the resources consumed for each use of that function. You can easily insert additional logging statements into your code. You can also call third-party logging APIs in your Lambda function. Visit Troubleshooting Lambda functions to learn more. Amazon CloudWatch Logs rates will apply.

Q: How do I scale an AWS Lambda function?

You do not have to scale your Lambda functions – AWS Lambda scales them automatically on your behalf. Every time an event notification is received for your function, AWS Lambda quickly locates free capacity within its compute fleet and runs your code. Since your code is stateless, AWS Lambda can start as many copies of your function as needed without lengthy deployment and configuration delays. There are no fundamental limits to scaling a function. AWS Lambda will dynamically allocate capacity to match the rate of incoming events.

Q: How are compute resources assigned to an AWS Lambda function?

In the AWS Lambda resource model, you choose the amount of memory you want for your function, and are allocated proportional CPU power and other resources. For example, choosing 256MB of memory allocates approximately twice as much CPU power to your Lambda function as requesting 128MB of memory and half as much CPU power as choosing 512MB of memory.

You can set your memory from 128MB to 10,240MB.

Q: When should I use AWS Lambda functions with more than 3008 MB of memory?

Customers running memory or compute-intensive workloads can now use more memory for their functions. Larger memory functions help multithreaded applications run faster, making them ideal for data and computationally intensive applications like machine learning, batch and ETL jobs, financial modeling, genomics, HPC, and media processing.

Q: How long can an AWS Lambda function execute?

AWS Lambda functions can be configured to run up to 15 minutes per execution. You can set the timeout to any value between 1 second and 15 minutes.

Q: How will I be charged for using AWS Lambda functions?

AWS Lambda is priced on a pay-per-use basis.

Q: Can I save money on AWS Lambda with a Compute Savings Plan?

Yes. In addition to saving money on Amazon EC2 and AWS Fargate, you can also use Compute Savings Plans to save money on AWS Lambda. Compute Savings Plans offer up to 17% discount on Duration, Provisioned Concurrency, and Duration (Provisioned Concurrency). Compute Savings Plans do not offer a discount on Requests in your Lambda bill. However, your Compute Savings Plans commitment can apply to Requests at regular rates.

Q: Does AWS Lambda support versioning?

Yes. By default, each AWS Lambda function has a single, current version of the code. Clients of your Lambda function can call a specific version or get the latest implementation. Please read our documentation on versioning Lambda functions.

Q: How long after uploading my code will my AWS Lambda function be ready to call?

Deployment times may vary with the size of your code, but AWS Lambda functions are typically ready to call within seconds of upload.

Q: Can I use my own version of a supported library?

Yes. You can include your own copy of a library (including the AWS SDK) in order to use a different version than the default one provided by AWS Lambda.

Q: How does tiered pricing work?

AWS Lambda offers discounted pricing tiers for monthly on-demand function duration above certain thresholds. Tiered pricing is available for functions running on both x86 and Arm architectures. Lambda pricing tiers are applied to aggregate monthly on-demand duration of your functions running on the same architecture (x86 or Arm, respectively), in the same region, within the account. If you're using consolidated billing in AWS Organizations, pricing tiers are applied to the aggregate monthly duration of your functions running on the same architecture, in the same region, across the accounts in the organization. For example, if you are running x86 Lambda functions in the US East (Ohio) region, you will pay $0.0000166667 for every GB-second for the first 6 billion GB-seconds per month, $0.0000150000 for every GB-second for the next 9 billion GB-seconds per month, and $0.0000133334 for every GB-second over 15 billion GB-seconds per month, in that region. Pricing for Requests, Provisioned Concurrency, and Provisioned Concurrency Duration remains unchanged. For more information, please see AWS Lambda Pricing.

Q: Can I take advantage of both tiered pricing, and Compute Savings Plans?

Yes. Lambda usage that is covered by your hourly savings plan commitment is billed at the applicable CSP rate and discount. The remaining usage that is not covered by this commitment will be billed at the rate corresponding to the tier your monthly aggregate function duration falls in.

Using AWS Lambda to process AWS events

Q: What is an event source?

An event source is an AWS service or developer-created application that produces events that trigger an AWS Lambda function to run. Some services publish these events to Lambda by invoking the cloud function directly (for example, Amazon S3). Lambda can also poll resources in other services that do not publish events to Lambda. For example, Lambda can pull records from an Amazon Kinesis stream or an Amazon SQS queue and execute a Lambda function for each fetched message.

Many other services, such as AWS CloudTrail, can act as event sources simply by logging to Amazon S3 and using S3 bucket notifications to trigger AWS Lambda functions.

Q: What event sources can be used with AWS Lambda?

Please see our documentation for a complete list of event sources.

Q: How are events represented in AWS Lambda?

Events are passed to a Lambda function as an event input parameter. For event sources where events arrive in batches, such as Amazon SQS, Amazon Kinesis, and Amazon DynamoDB Streams, the event parameter may contain multiple events in a single call, based on the batch size you request. To learn more about Amazon S3 event notifications, visit Configuring Notifications for Amazon S3 Events. To learn more about Amazon DynamoDB Streams, visit the DynamoDB Stream Developers Guide. To learn more about invoking Lambda functions using Amazon SNS, visit the Amazon SNS Developers Guide. For more information on Amazon Cognito events, visit Amazon Cognito. For more information on AWS CloudTrail logs and auditing API calls across AWS services, see AWS CloudTrail.

Q: How do I make an AWS Lambda function respond to changes in an Amazon S3 bucket?

From the AWS Lambda console, you can select a function and associate it with notifications from an Amazon S3 bucket. Alternatively, you can use the Amazon S3 console and configure the bucket's notifications to send to your AWS Lambda function. This same functionality is also available through the AWS SDK and CLI.

Q: How do I make an AWS Lambda function respond to updates in an Amazon DynamoDB table?

You can trigger a Lambda function on DynamoDB table updates by subscribing your Lambda function to the DynamoDB Stream associated with the table. You can associate a DynamoDB Stream with a Lambda function using the Amazon DynamoDB console, the AWS Lambda console, or Lambda's registerEventSource API.

Q: How do I use an AWS Lambda function to process records in an Amazon Kinesis stream?

From the AWS Lambda console, you can select a Lambda function and associate it with an Amazon Kinesis stream owned by the same account. This same functionality is also available through the AWS SDK and CLI.

Q: How does AWS Lambda process data from Amazon Kinesis streams and Amazon DynamoDB Streams?

The Amazon Kinesis and DynamoDB Streams records sent to your AWS Lambda function are strictly serialized, per shard. This means that if you put two records in the same shard, Lambda guarantees that your Lambda function will be successfully invoked with the first record before it is invoked with the second record. If the invocation for one record times out, is throttled, or encounters any other error, Lambda will retry until it succeeds (or the record reaches its 24-hour expiration) before moving on to the next record. The ordering of records across different shards is not guaranteed, and processing of each shard happens in parallel.

Q: How should I choose between AWS Lambda and Amazon Kinesis Data Analytics for my analytics needs?

AWS Lambda allows you to perform time-based aggregations (such as count, max, sum, average, etc.) over a short window of up to 15 minutes for your data in Amazon Kinesis or Amazon DynamoDB Streams over a single logical partition such as a shard. This gives you the option to easily set up simple analytics for your event-based application without adding architectural complexity, as your business and analytics logic can be located in the same function. Lambda allows aggregations over a maximum of a 15-minute tumbling window, based

on the event timestamp. Amazon Kinesis Data Analytics allows you to build more complex analytics applications that support flexible processing choices and robust fault-tolerance with exactly-once processing without duplicates, and analytics that can be performed over an entire data stream across multiple logical partitions. With KDA, you can analyze data over multiple types of aggregation windows (tumbling window, stagger window, sliding window, session window) using either the event time or the processing time.

	AWS Lambda	**Amazon KDA**
Tumbling Window	Yes	Yes
Stagger Window	No	Yes
Sliding Window	No	Yes
Session Window	No	Yes
Enrichment	No	Yes
Joint input and reference tables	No	Yes
Split input stream	No	Yes
Exactly-once processing	No	Yes
Maximum time window	15 mins	No limit
Aggregation scope	Partition/shard	Stream
Time semantics	Event time	Event time, Processing time

Q: How do I use an AWS Lambda function to respond to notifications sent by Amazon Simple Notification Service (SNS)?

From the AWS Lambda console, you can select a Lambda function and associate it with an Amazon SNS topic. This same functionality is also available through the AWS SDK and CLI.
Q: How do I use an AWS Lambda function to respond to emails sent by Amazon Simple Email Service (SES)?

From the Amazon SES Console, you can set up your receipt rule to have Amazon SES deliver your messages to an AWS Lambda function. The same functionality is available through the AWS SDK and CLI.
Q: How do I use an AWS Lambda function to respond to Amazon CloudWatch alarms?

First, configure the alarm to send Amazon SNS notifications. Then from the AWS Lambda console, select a Lambda function and associate it with that Amazon SNS topic. See the Amazon CloudWatch Developer Guide for more on setting up Amazon CloudWatch alarms.
Q: How do I use an AWS Lambda function to respond to changes in user or device data managed by Amazon Cognito?

From the AWS Lambda console, you can select a function to trigger when any datasets associated with an Amazon Cognito identity pool are synchronized. This same functionality is also available through the AWS SDK and CLI. Visit Amazon Cognito for more information on using Amazon Cognito to share and synchronize data across a user's devices.
Q: How can my application trigger an AWS Lambda function directly?

You can invoke a Lambda function using a custom event through AWS Lambda's invoke API. Only the function owner or another AWS account that the owner has granted permission can invoke the function. Visit the Lambda Developers Guide to learn more.
Q: What is the latency of invoking an AWS Lambda function in response to an event?

AWS Lambda is designed to process events within milliseconds. Latency will be higher immediately after a Lambda function is created, updated, or if it has not been used recently.
Q: How do I create a mobile backend using AWS Lambda?

You upload the code you want AWS Lambda to execute and then invoke it from your mobile app using the AWS Lambda SDK included in the AWS Mobile SDK. You can make both direct (synchronous) calls to retrieve or check data in real time, as well as asynchronous calls. You can also define a custom API using Amazon API Gateway and invoke your Lambda functions through any REST compatible client. To learn more about the AWS Mobile SDK, visit the AWS Mobile SDK page. To learn more about Amazon API Gateway, visit the Amazon API Gateway page.
Q: How do I invoke an AWS Lambda function over HTTPS?

You can invoke a Lambda function over HTTPS by defining a custom RESTful API using Amazon API Gateway. This gives you an endpoint for your function which can respond to REST calls like GET, PUT, and POST. Read more about using AWS Lambda with Amazon API Gateway.
Q: How can my AWS Lambda function customize its behavior to the device and app making the request?

When called through the AWS Mobile SDK, AWS Lambda functions automatically gain insight into the device and application that made the call through the 'context' object.
Q: How can my AWS Lambda function personalize its behavior based on the identity of the end-user of an application?

When your app uses the Amazon Cognito identity, end users can authenticate themselves using a variety of public login providers such as Amazon, Facebook, Google, and other OpenID Connect-compatible services. User identity is then automatically and secured presented to your Lambda function in the form of an Amazon Cognito id, allowing it to access user data from Amazon Cognito, or as a key to store and retrieve data in Amazon DynamoDB or other web services.
Q: How do I create an Alexa skill using AWS Lambda?

AWS Lambda is integrated with the Alexa Skills Kit, a collection of self-service APIs, tools, documentation, and code samples that make it easy for you to create voice-driven capabilities (or "skills") for Alexa. You simply upload the Lambda function code for the new Alexa skill you are creating, and AWS Lambda does the rest, executing the code in response to Alexa voice interactions and automatically managing the compute resources on your behalf. Read the Alexa Skills Kit documentation for more details.

Q: What happens if my function fails while processing an event?

For Amazon S3 bucket notifications and custom events, AWS Lambda will attempt execution of your function three times in the event of an error condition in your code or if you exceed a service or resource limit.

For ordered event sources that AWS Lambda polls on your behalf, such as Amazon DynamoDB Streams and Amazon Kinesis streams, Lambda will continue attempting execution in the event of a developer code error until the data expires. You can monitor progress through the Amazon Kinesis and Amazon DynamoDB consoles and through the Amazon CloudWatch metrics that AWS Lambda generates for your function. You can also set Amazon CloudWatch alarms based on error or execution throttling rates.

Using AWS Lambda to build applications

Q: What is a serverless application?

Lambda-based applications (also referred to as serverless applications) are composed of functions triggered by events. A typical serverless application consists of one or more functions triggered by events such as object uploads to Amazon S3, Amazon SNS notifications, or API actions. These functions can stand alone or leverage other resources such as DynamoDB tables or Amazon S3 buckets. The most basic serverless application is simply a function.
Q: How do I deploy and manage a serverless application?

You can deploy and manage your serverless applications using the AWS Serverless Application Model (AWS SAM). AWS SAM is a specification that prescribes the rules for expressing serverless applications on AWS. This specification aligns with the syntax used by AWS CloudFormation today and is supported natively within AWS CloudFormation as a set of resource types (referred to as "serverless resources"). These resources make it easier for AWS customers to use CloudFormation to configure and deploy serverless applications using existing CloudFormation APIs.
Q: How can I discover existing serverless applications developed by the AWS community?

You can choose from a collection of serverless applications published by developers, companies, and partners in the AWS community with the AWS Serverless Application Repository. After finding an application, you can configure and deploy it straight from the Lambda console.
Q: How do I automate deployment for a serverless application?

You can automate your serverless application release process using AWS CodePipeline and AWS CodeDeploy. CodePipeline is a continuous delivery service that enables you to model, visualize and automate the steps required to release your serverless application. CodeDeploy provides a deployment automation engine for your Lambda-based applications. CodeDeploy lets you orchestrate deployments according to established best-practice methodologies such as canary and linear deployments, and helps you establish the necessary guardrails to verify that newly-deployed code is safe, stable, and ready to be fully released to production.

To learn more about serverless CI/CD, visit our documentation.
Q: How do I get started on building a serverless application?

To get started, visit the AWS Lambda console and download one of our blueprints. The file you download will contain an AWS SAM file (which defines the AWS resources in your application) and a .ZIP file (which includes your function code). You can then use AWS CloudFormation commands to package and deploy the serverless application that you just downloaded. For more details, visit our documentation.
Q: How do I coordinate calls between multiple AWS Lambda functions?

You can use AWS Step Functions to coordinate a series of AWS Lambda functions in a specific order. You can invoke multiple Lambda functions sequentially, passing the output of one to the other, and/or in parallel, and Step Functions will maintain state during executions for you.
Q: How do I troubleshoot a serverless application?

You can enable your Lambda function for tracing with AWS X-Ray by adding X-Ray permissions to your Lambda function execution role and changing your function "tracing mode" to "active. " When X-Ray is enabled for your Lambda function, AWS Lambda will emit tracing information to X-Ray regarding the Lambda service overhead incurred when invoking your function. This will provide you with insights such as Lambda service overhead, function init time, and function execution time. In addition, you can include the X-Ray SDK in your Lambda deployment package to create your own trace segments, annotate your traces, or view trace segments for downstream calls made from your Lambda function. X-Ray SDKs are currently available for Node.js and Java. Visit Troubleshooting Lambda-based applications to learn more. AWS X-Ray rates will apply.
Q. Can I build serverless applications that connect to relational databases?

Yes. You can build highly scalable, secure, Lambda-based serverless applications that connect to relational databases using Amazon RDS Proxy, a highly available database proxy that manages thousands of concurrent connections to relational databases. Currently, RDS Proxy supports MySQL and Aurora databases. You can begin using RDS Proxy through the Amazon RDS console or the AWS Lambda console. Serverless applications that use fully managed connection pools from RDS Proxy will be billed according to RDS Proxy Pricing.
Q: How is AWS SAM licensed?

The specification is open sourced under Apache 2.0, which allows you and others to adopt and incorporate AWS SAM into build, deployment, monitoring, and management tools with a commercial-friendly license. You can access the AWS SAM repository on GitHub here.

Container Image Support

Q: What is Container Image Support for AWS Lambda?

AWS Lambda now enables you to package and deploy functions as container images. Customers can leverage the flexibility and familiarity of container tooling, and the agility and operational simplicity of AWS Lambda to build applications.

Q: How can I use Container Image Support for AWS Lambda?

You can start with either an AWS provided base images for Lambda or by using one of your preferred community or private enterprise images. Then, simply use Docker CLI to build the image, upload it to Amazon ECR, and then create the function by using all familiar Lambda interfaces and tools, such as the AWS Management Console, the AWS CLI, the AWS SDK, AWS SAM, and AWS CloudFormation.

Q: Which container image types are supported?

You can deploy third-party Linux base images (e.g. Alpine or Debian) to Lambda in addition to the Lambda provided images. AWS Lambda will support all images based on the following image manifest formats: Docker Image Manifest V2 Schema 2 (used with Docker version 1.10 and newer) or Open Container Initiative (OCI) Spec (v1.0 and up). Lambda supports images with a size of up to 10GB.

Q: What base images can I use?

AWS Lambda provides a variety of base images customers can extend, and customers can also use their preferred Linux-based images with a size of up to 10GB.

Q: What container tools can I use to package and deploy functions as container images?

You can use any container tooling as long as it supports one of the following container image manifest formats: Docker Image Manifest V2 Schema 2 (used with Docker version 1.10 and newer) or Open Container Initiative (OCI) Specifications (v1.0 and up). For example, you can use native container tools (i.e. docker run, docker compose, Buildah and Packer) to define your functions as a container image and deploy to Lambda.

Q: What AWS Lambda features are available to functions deployed as container images?

All existing AWS Lambda features, with the exception of Lambda layers and Code Signing, can be used with functions deployed as container images. Once deployed, AWS Lambda will treat an image as immutable. Customers can use container layers during their build process to include dependencies.

Q: Will AWS Lambda patch and update my deployed container image?

Not at this time. Your image, once deployed to AWS Lambda, will be immutable. The service will not patch or update the image. However, AWS Lambda will publish curated base images for all supported runtimes that are based on the Lambda managed environment. These published images will be patched and updated along with updates to the AWS Lambda managed runtimes. You can pull and use the latest base image from DockerHub or Amazon ECR Public, re-build your container image and deploy to AWS Lambda via Amazon ECR. This allows you to build and test the updated images and runtimes, prior to deploying the image to production.

Q: What are the differences between functions created using ZIP archives vs. container images?

There are three main differences between functions created using ZIP archives vs. container images:

1. Functions created using ZIP archives have a maximum code package size of 250 MB unzipped, and those created using container images have a maximum image size of 10 GB.
2. Lambda uses Amazon ECR as the underlying code storage for functions defined as container images, so a function may not be invocable when the underlying image is deleted from ECR.
3. ZIP functions are automatically patched for the latest runtime security and bug fixes. Functions defined as container images are immutable, and customers are responsible for the components packaged in their function. Customers can leverage the AWS provided base images which are regularly updated by AWS for security and bug fixes, using the most recent patches available.

Q: Is there a performance difference between functions defined as zip and container images?

No - AWS Lambda ensures that the performance profiles for functions packaged as container images are the same as for those packaged as ZIP archives, including typically sub-second start up times.

Q: How will I be charged for deploying Lambda functions as container images?

There is no additional charge for packaging and deploying functions as container images to AWS Lambda. When you invoke your function deployed as a container image, you pay the regular price for requests and execution duration.

Q: What is the Lambda Runtime Interface Emulator (RIE)?

The Lambda Runtime Interface Emulator is a proxy for the Lambda Runtime API,which allows customers to locally test their Lambda function packaged as a container image. It is a lightweight web server that converts HTTP requests to JSON events and emulates the Lambda Runtime API. It allows you to locally test your functions using familiar tools such as cURL and the Docker CLI (when testing functions packaged as container images). It also simplifies running your application on additional compute services. You can include the Lambda Runtime Interface Emulator in your container image to have it accept HTTP requests natively instead of the JSON events required for deployment to Lambda. This component does not emulate the Lambda orchestrator, or security and authentication configurations. The Runtime Interface Emulator is open sourced on GitHub. You can get started by downloading and installing it on your local machine.

Q: Why do I need the Lambda Runtime Interface Emulator (RIE) during local testing?

The Lambda Runtime API in the running Lambda service accepts JSON events and returns responses. The Lambda Runtime Interface Emulator allows the function packaged as a container image to accept HTTP requests during local testing with tools like cURL, and surface them via the same interface locally to the function. It allows you to use the docker run or docker-compose up command to locally test your lambda application.

Q: What function behaviors can I test locally with the emulator?

You can use the emulator to test if your function code is compatible with the Lambda environment, runs successfully, and provides the expected output. For example, you can mock test events from different event sources. You can also use it to test extensions and agents built into the container image against the Lambda Extensions API.

Q: How does the Runtime Interface Emulator (RIE) help me run my Lambda compatible image on additional compute services?

Customers can add the Runtime Interface Emulator as the entry point to the container image or package it as a sidecar to ensure the container image now accepts HTTP requests instead of JSON events. This simplifies the changes required to run their container image on additional compute services. Customers will be responsible for ensuring they follow all security, performance, and concurrency best practices for their chosen environment. RIE is pre-packaged into the AWS Lambda provided images, and is available by default in AWS SAM CLI. Base image providers can use the documentation to provide the same experience for their base images.

Q: How can I deploy my existing containerized application to AWS Lambda?

You can deploy a containerized application to AWS Lambda if it meets the below requirements:

1. The container image must implement the Lambda Runtime API. We have open-sourced a set of software packages, Runtime Interface Clients (RIC), that implement the Lambda Runtime API, allowing you to seamlessly extend your preferred base images to be Lambda compatible.
2. The container image must be able to run on a read-only filesystem. Your function code can access a writable /tmp directory storage of 512 MB. If you are using an image that requires a writable root directory, configure it to write to the /tmp directory.

3. The files required for the execution of function code can be read by the default Lambda user. Lambda defines a default Linux user with least-privileged permissions that follows security best practices. You need to verify that your application code does not rely on files that are restricted by other Linux users for execution.
4. It is a Linux based container image.

Provisioned Concurrency

Q: What is AWS Lambda Provisioned Concurrency?

Provisioned Concurrency gives you greater control over the performance of your serverless applications. When enabled, Provisioned Concurrency keeps functions initialized and hyper-ready to respond in double-digit milliseconds.
Q: How do I set up and manage Provisioned Concurrency?

You can configure concurrency on your function through the AWS Management Console, the Lambda API, the AWS CLI, and AWS CloudFormation. The simplest way to benefit from Provisioned Concurrency is by using AWS Auto Scaling. You can use Application Auto Scaling to configure schedules, or have Auto Scaling automatically adjust the level of Provisioned Concurrency in real time as demand changes. To learn more about Provisioned Concurrency, see the documentation.
Q: Do I need to change my code if I want to use Provisioned Concurrency?

You don't need to make any changes to your code to use Provisioned Concurrency. It works seamlessly with all existing functions and runtimes. There is no change to the invocation and execution model of Lambda when using Provisioned Concurrency.
Q: How will I be charged for Provisioned Concurrency?

Provisioned Concurrency adds a pricing dimension, of 'Provisioned Concurrency', for keeping functions initialized. When enabled, you pay for the amount of concurrency that you configure and for the period of time that you configure it. When your function executes while Provisioned Concurrency is configured on it, you also pay for Requests and execution Duration.

Q: When should I use Provisioned Concurrency?

Provisioned Concurrency is ideal for building latency-sensitive applications, such as web or mobile backends, synchronously invoked APIs, and interactive microservices. You can easily configure the appropriate amount of concurrency based on your application's unique demand. You can increase the amount of concurrency during times of high demand and lower it, or turn it off completely, when demand decreases.
Q: What happens if a function receives invocations above the configured level of Provisioned Concurrency?

If the concurrency of a function reaches the configured level, subsequent invocations of the function have the latency and scale characteristics of regular Lambda functions. You can restrict your function to only scale up to the configured level. Doing so prevents the function from exceeding the configured level of Provisioned Concurrency. This is a mechanism to prevent undesired variability in your application when demand exceeds the anticipated amount.

AWS Lambda functions powered by Graviton2 processors

Q: What are AWS Lambda functions powered by Graviton2 processors?

AWS Lambda allows you to run your functions on either x86-based or Arm-based processors. AWS Graviton2 processors are custom built by Amazon Web Services using 64-bit Arm Neoverse cores to deliver increased price performance for your cloud workloads. Customers get the same advantages of AWS Lambda, running code without provisioning or managing servers, automatic scaling, high availability, and only paying for the resources you consume.

Q: Why should I use AWS Lambda functions powered by Graviton2 processors?

AWS Lambda functions powered by Graviton2, using an Arm-based processor architecture designed by AWS, are designed to deliver up to 34% better price performance compared to functions running on x86 processors, for a variety of serverless workloads, such as web and mobile backends, data, and stream processing. With lower latency, up to 19% better performance, a 20% lower cost, and the highest power-efficiency currently available at AWS, Graviton2 functions can power mission critical serverless applications. Customers can configure both existing and new functions to target the Graviton2 processor. They can deploy functions running on Graviton2 as either zip files or container images.

Q: How do I configure my functions to run on Graviton2 processors?

You can configure functions to run on Graviton2 through the AWS Management Console, the AWS Lambda API, the AWS CLI, and AWS CloudFormation by setting the architecture flag to 'arm64' for your function.

Q: How do I deploy my application built using functions powered by Graviton2 processors?

There is no change between x86-based and Arm-based functions. Simply upload your code via the AWS Management Console, zip file, or container image, and AWS Lambda automatically runs your code when triggered, without requiring you to provision or manage infrastructure.

Q: Can an application use both functions powered by Graviton2 processors and x86 processors?

An application can contain functions running on both architectures. AWS Lambda allows you to change the architecture ('x86_64' or 'arm64') of your function's current version. Once you create a specific version of your function, the architecture cannot be changed.

Q: Does AWS Lambda support multi-architecture container images?

No. Each function version can only use a single container image.

Q: Can I create AWS Lambda Layers that target functions powered by AWS Graviton2 processors?

Yes. Layers and extensions can be targeted to 'x86_64' or 'arm64' compatible architectures. The default architecture for functions and layers is 'x86_64'.

Q: What languages and runtimes are supported by Lambda functions running on Graviton2 processors?

At launch, customers can use Python, Node.js, Java, Ruby, .Net Core, Custom Runtime (provided.al2), and OCI Base images. To learn more, please see the AWS Lambda Runtimes.

Q: What is the pricing of AWS Lambda functions powered by AWS Graviton2 processors? Does the AWS Lambda free tier apply to functions powered by Graviton2?

AWS Lambda functions powered by AWS Graviton2 processors are 20% cheaper compared to x86-based Lambda functions. The Lambda free tier applies to AWS Lambda functions powered by x86 and Arm-based architectures.

Q: How do I choose between running my functions on Graviton2 processors or x86 processors?

Each workload is unique and we recommend customers test their functions to determine the price performance improvement they might see. To do that, we recommend using the AWS Lambda Power Tuning tool. We recommend starting with web and mobile backends, data, and stream processing when testing your workloads for potential price performance improvements.

Q: Do I need an Arm-based development machine to create, build, and test functions powered by Graviton2 processors locally?

Interpreted languages like Python, Java, and Node generally do not require recompilation unless your code references libraries that use architecture specific components. In those cases, you would need to provide the libraries targeted to arm64. For more details, please see the Getting started with AWS Graviton page. Non-interpreted languages will require compiling your code to target arm64. While more modern compilers will produce compiled code for arm64, you will need to deploy it into an arm-based environment to test. To learn more about using Lambda functions with Graviton2, please see the documentation.

Amazon EFS for AWS Lambda

Q: What is Amazon EFS for AWS Lambda?

With Amazon Elastic File System (Amazon EFS) for AWS Lambda, customers can securely read, write and persist large volumes of data at virtually any scale using a fully managed elastic NFS file system that can scale on demand without the need for provisioning or capacity management. Previously, developers added code to their functions to download data from S3 or databases to local temporary storage, limited to 512MB. With EFS for Lambda, developers don't need to write code to download data to temporary storage in order to process it.

Q: How do I set up Amazon EFS for Lambda?

Developers can easily connect an existing EFS file system to a Lambda function via an EFS Access Point by using the console, CLI, or SDK. When the function is first invoked, the file system is automatically mounted and made available to function code. You can learn more in the documentation.

Q: Do I need to configure my function with VPC settings before I can use my Amazon EFS file system?

Yes. Mount targets for Amazon EFS are associated with a subnet in a VPC. The AWS Lambda function needs to be configured to access that VPC.

Q: Who should use Amazon EFS for Lambda?

Using EFS for Lambda is ideal for building machine learning applications or loading large reference files or models, processing or backing up large amounts of data, hosting web content, or developing internal build systems. Customers can also use EFS for Lambda to keep state between invocations within a stateful microservice architecture, in a Step Functions workflow, or sharing files between serverless applications and instance or container-based applications.

Q: Will my data be encrypted in transit?

Yes. Data encryption in transit uses industry-standard Transport Layer Security (TLS) 1.2 to encrypt data sent between AWS Lambda functions and the Amazon EFS file systems.

Q: Is my data encrypted at rest?

Customers can provision Amazon EFS to encrypt data at rest. Data encrypted at rest is transparently encrypted while being written, and transparently decrypted while being read, so you don't have to modify your applications. Encryption keys are managed by the AWS Key Management Service (KMS), eliminating the need to build and maintain a secure key management infrastructure.

Q: How will I be charged for Amazon EFS for AWS Lambda?

There is no additional charge for using Amazon EFS for AWS Lambda. Customers pay the standard price for AWS Lambda and for Amazon EFS. When using Lambda and EFS in the same availability zone, customers are not charged for data transfer. However, if they use VPC peering for Cross-Account access, they will incur data transfer charges.

Q: Can I associate more than one Amazon EFS file system with my AWS Lambda function?

No. Each Lambda function will be able to access one EFS file system.

Q: Can I use the same Amazon EFS file system across multiple functions, containers, and instances?

Yes. Amazon EFS supports Lambda functions, ECS and Fargate containers, and EC2 instances. You can share the same file system and use IAM policy and Access Points to control what each function, container, or instance has access to.

Lambda Extensions

Q: What is AWS Lambda Extensions?

AWS Lambda Extensions lets you integrate Lambda with your favorite tools for monitoring, observability, security, and governance. Extensions enable you and your preferred tooling vendors to plug into Lambda's lifecycle and integrate more deeply into the Lambda execution environment.

Q: How do Lambda extensions work?

Extensions are companion processes that run within Lambda's execution environment which is where your function code is executed. In addition, they can run outside of the function invocation - i.e. they start before the function is initialized, run in parallel with the function, can run after the function execution is complete, and can also run before the Lambda service shuts down the execution environment.

Q: What can I use Lambda extensions for?

You can use extensions for your favorite tools for monitoring, observability, security, and governance from AWS as well as the following partners: AppDynamics, Coralogix, Datadog, Dynatrace, Epsagon, HashiCorp, Honeycomb, Imperva, Lumigo, Check Point CloudGuard, New Relic, Thundra, Splunk, Sentry, Site24x7, Sumo Logic, AWS AppConfig, Amazon CodeGuru Profiler, Amazon CloudWatch Lambda Insights, AWS Distro for OpenTelemetry. To learn more about these extensions, visit the launch blog post.

Q: How do I set up and manage Lambda extensions?

You can deploy extensions, using Layers, on one or more Lambda functions using the Console, CLI, or Infrastructure as Code tools such as CloudFormation, the AWS Serverless Application Model, and Terraform. To get started, visit the documentation.

Q: What runtimes can I use AWS Lambda extensions with?

You can view the list of runtimes that support extensions here.

Q: Do Extensions count towards the deployment package limit?

Yes, the total unzipped size of the function and all Extensions cannot exceed the unzipped deployment package size limit of 250 MB.

Q: Is there a performance impact of using an extension?

Extensions may impact the performance of your function because they share resources such as CPU, memory, and storage with the function, and because extensions are initialized before function code. For example, if an extension performs compute-intensive operations, you may see your function execution duration increase because the extension and your function code share the same CPU resources. Because Lambda allocates proportional CPU based on the memory setting you choose, you may see increased execution and initialization duration at lower memory settings as more processes compete for the same CPU resources.

You can use the *PostRuntimeExecutionDuration* metric to measure the extra time the extension takes after the function execution, and you can use the MaxMemoryUsed metric to measure the increase in memory used. To understand the impact of a specific extension, you can also use the Duration metric. Currently, the function execution response is returned after function execution and extension execution have been completed. To learn more, visit the Lambda developer documentation.

Q: How will I be charged for using Lambda extensions?

Extensions share the same billing model as Lambda functions. When using Lambda functions with extensions, you pay for requests served and the combined compute time used to run your code and all extensions, in 1ms increments. You will be charged for compute time as per existing Lambda duration pricing.

The Lambda lifecycle is made up of three distinct phases: ‘init’, when AWS Lambda initializes the function, dependencies, and extensions; ‘invoke’, when Lambda executes function and extension code in response to triggers; and ‘shut down’, after function execution has completed, but extension code could still be executing, and which can last up to two seconds. You will be charged for compute time used to run your extension code during all three phases of the Lambda lifecycle. To learn more about the Lambda lifecycle, see the documentation on the Lambda Execution Environment.

There is no additional cost for installing extensions, although partner offerings may be chargeable. See the relevant partner website for details.

Q: Can I create my own custom Lambda extensions?

Yes, by using the AWS Lambda Runtime Extensions API. Visit the documentation to learn more.

Q: How do extensions work while Provisioned Concurrency is enabled?

Provisioned Concurrency keeps functions initialized and ready to respond in double-digit milliseconds. When enabled, Provisioned Concurrency will also initialize extensions and keep them ready to execute alongside function code.

Q: What permissions do extensions have?

Because Extensions are executed within the same environment as a Lambda function, they have access to the same resources as the function, and permissions are shared between the function and the extension. Therefore they share credentials, role, and environment variables. Extensions have read-only access to function code, and can read and write in /tmp.
Q: What is the AWS Lambda Telemetry API?

The AWS Lambda Telemetry API enables you to use extensions to capture enhanced monitoring and observability data directly from Lambda and send it to a destination of your choice.

Q: How does the Telemetry API work?

The Lambda service automatically captures and streams telemetry data to Amazon CloudWatch and AWS X-Ray. The Telemetry API provides a simple HTTP or TCP interface for extensions to receive the same telemetry data along with Lambda execution environment lifecycle events, and function invocation-level metrics. Extensions can use the Telemetry API to consume these telemetry streams directly from Lambda, and then process, filter, and send them to any preferred destination.

Q: How do I get started with using the Telemetry API?

You can deploy Telemetry API enabled extensions for your Lambda functions using AWS Lambda Console, AWS CLI, or Infrastructure as Code tools such as AWS CloudFormation, AWS Serverless Application Model (SAM), and Terraform. You do not have to make code changes to use a Telemetry API enabled extension with your Lambda function. Simply add an extension from the tooling provider of your choice to your Lambda function. To get started with extensions from APN Partners, follow the links provided in the launch blog post. You can also build your own extension that uses Telemetry API. To learn how, visit the AWS Lambda Developer Guide.

Q: Is there a performance impact of using the Telemetry API?

You can only use the Telemetry API from within AWS Lambda Extensions. Extensions may impact the performance of your function because they share resources such as CPU, memory, and storage with the function. Memory usage increases linearly as the number of Telemetry API subscriptions increase because each subscription opens a new memory buffer to store the telemetry data. However, you can optimize memory usage by adjusting the buffering configuration in the Telemetry API subscription request. We recommend extension vendors to publish expected resource consumption to make it easier for function developers to choose a suitable extension. Please refer to your extension vendor's documentation to understand the potential performance overhead of using their extension.

Q: How will I be charged for using the Telemetry API?

There is no additional charge for using the AWS Lambda Telemetry API. Extensions that use the Telemetry API share the same billing model as other extensions and Lambda functions.

Q: Does using the Telemetry API disable sending logs to Amazon CloudWatch Logs?

No. By default, the Lambda service sends all telemetry data to CloudWatch Logs, and using the Telemetry API does not disable egress to CloudWatch Logs.

Lambda function URLs

Q: Do AWS Lambda functions support HTTP(S) endpoints?

Yes. Lambda functions can be configured with a function URL, a built-in HTTPS endpoint that can be invoked using the browser, curl, and any HTTP client. Function URLs are an easy way to get started building HTTPS accessible functions.

Q: How do I configure a Lambda function URL for my function?

You can configure a function URL for your function through the AWS Management Console, the AWS Lambda API, the AWS CLI, AWS CloudFormation, and the AWS Serverless Application Model. Function URLs can be enabled on the $LATEST unqualified version of your function, or on any function alias. To learn more about configuring a function URL, see the documentation.

Q: How do I secure my Lambda function URL?

Lambda function URLs are secured with IAM authorization by default. You can choose to disable IAM authorization to create a public endpoint or if you plan to implement custom authorization as part of the function's business logic.

Q: How do I invoke my function with a Lambda function URL?

You can easily invoke your function from your web browser by navigating to the Lambda URL, from your client application's code using an HTTP library, or from the command line using curl.

Q: Do Lambda function URLs work with function versions and aliases?

Yes. Lambda function URLs can be enabled on a function or function alias. If no alias is specified, the URL will point to $LATEST by default. Function URLs cannot target an individual function version.

Q: Can I enable custom domains for my Lambda function URL?

Custom domain names are not currently supported with function URLs. You can use a custom domain with your function URL by creating an Amazon CloudFront distribution and a CNAME to map your custom domain to your CloudFront distribution name. Then, map your CloudFront distribution domain name to be routed to your function URL as an origin.

Q: Can Lambda function URLs be used to invoke a function in a VPC?

Yes, function URLs can be used to invoke a Lambda function in a VPC.

Q: What is the pricing for using Lambda function URLs?

There is no additional charge for using function URLs. You pay the standard price for AWS Lambda.

Lambda@Edge

Q: What is Lambda@Edge?

Lambda@Edge allows you to run code across AWS locations globally without provisioning or managing servers, responding to end-users at the lowest network latency. You just upload your Node.js or Python code to AWS Lambda and configure your function to be triggered in response to Amazon CloudFront requests (i.e., when a viewer request lands, when a request is forwarded to or received back from the origin, and right before responding back to the end-

user). The code is then ready to execute across AWS locations globally when a request for content is received, and scales with the volume of CloudFront requests globally. Learn more in our documentation.
Q: How do I use Lambda@Edge?

To use Lambda@Edge, you just upload your code to AWS Lambda and associate a function version to be triggered in response to Amazon CloudFront requests. Your code must satisfy the Lambda@Edge service limits. Lambda@Edge supports Node.js and Python for global invocation by CloudFront events at this time. Learn more in our documentation.
Q: When should I use Lambda@Edge?

Lambda@Edge is optimized for latency-sensitive use cases where your end viewers are distributed globally. All the information you need to make a decision should be available at the CloudFront edge, within the function and the request. This means that use cases where you are looking to make decisions on how to serve content based on user characteristics (e.g., location, client device, etc.) can now be executed and served close to your users without having to be routed back to a centralized server.
Q: Can I deploy my existing Lambda functions for global invocation?

You can associate existing Lambda functions with CloudFront events for global invocation if the function satisfies the Lambda@Edge service requirements and limits. Read more here on how to update your function properties.
Q: What Amazon CloudFront events can be used to trigger my functions?

Your functions will automatically trigger in response to the following Amazon CloudFront events:

- Viewer Request - This event occurs when an end-user or a device on the Internet makes an HTTP(S) request to CloudFront, and the request arrives at the edge location closest to that user.
- Viewer Response - This event occurs when the CloudFront server at the edge is ready to respond to the end user or the device that made the request.
- Origin Request - This event occurs when the CloudFront edge server does not already have the requested object in its cache, and the viewer request is ready to be sent to your backend origin web server (e.g. Amazon EC2, or Application Load Balancer, or Amazon S3).
- Origin Response - This event occurs when the CloudFront server at the edge receives a response from your backend origin web server.

Q: How is AWS Lambda@Edge different from using AWS Lambda behind Amazon API Gateway?

The difference is that API Gateway and Lambda are regional services.
Using Lambda@Edge and Amazon CloudFront allows you to execute logic across multiple AWS locations based on where your end viewers are located.

Scalability and availability

Q: How available are AWS Lambda functions?

AWS Lambda is designed to use replication and redundancy to provide high availability for both the service itself and for the Lambda functions it operates. There are no maintenance windows or scheduled downtimes for either.
Q: Do my AWS Lambda functions remain available when I change my code or its configuration?

Yes. When you update a Lambda function, there will be a brief window of time, typically less than a minute, when requests could be served by either the old or the new version of your function.
Q: Is there a limit to the number of AWS Lambda functions I can execute at once?

No. AWS Lambda is designed to run many instances of your functions in parallel. However, AWS Lambda has a default safety throttle for the number of concurrent executions per account per region (visit here for info on default safety throttle limits). You can also control the maximum concurrent executions for individual AWS Lambda functions, which you can use to reserve a subset of your account concurrency limit for critical functions, or cap traffic rates to downstream resources.

If you wish to submit a request to increase the throttle limit, you can visit our Support Center, click "Open a new case," and file a service limit increase request.
Q: What happens if my account exceeds the default throttle limit on concurrent executions?

On exceeding the throttle limit, AWS Lambda functions being invoked synchronously will return a throttling error (429 error code). Lambda functions being invoked asynchronously can absorb reasonable bursts of traffic for approximately 15-30 minutes, after which incoming events will be rejected as throttled. In case the Lambda function is being invoked in response to Amazon S3 events, events rejected by AWS Lambda may be retained and retried by S3 for 24 hours. Events from Amazon Kinesis streams and Amazon DynamoDB streams are retried until the Lambda function succeeds or the data expires. Amazon Kinesis and Amazon DynamoDB Streams retain data for 24 hours.
Q: Is the default limit applied on a per function level?

No, the default limit only applies at an account level.
Q: What happens if my Lambda function fails while processing an event?

On failure, Lambda functions being invoked synchronously will respond with an exception. Lambda functions being invoked asynchronously are retried at least 3 times. Events from Amazon Kinesis streams and Amazon DynamoDB streams are retried until the Lambda function succeeds or the data expires. Kinesis and DynamoDB Streams retain data for a minimum of 24 hours.
Q: What happens if my Lambda function invocations exhaust the available policy?

On exceeding the retry policy for asynchronous invocations, you can configure a “dead letter queue” (DLQ) into which the event will be placed; in the absence of a configured DLQ the event may be rejected. On exceeding the retry policy for stream based invocations, the data would have already expired and therefore rejected.
Q: What resources can I configure as a dead letter queue for a Lambda function?

You can configure an Amazon SQS queue or an Amazon SNS topic as your dead letter queue.

Security and access control

Q: How do I allow my AWS Lambda function access to other AWS resources?

You grant permissions to your Lambda function to access other resources using an IAM role. AWS Lambda assumes the role while executing your Lambda function, so you always retain full, secure control of exactly which AWS resources it can use. Visit Setting up AWS Lambda to learn more about roles.
Q: How do I control which Amazon S3 buckets can call which AWS Lambda functions?

When you configure an Amazon S3 bucket to send messages to an AWS Lambda function, a resource policy rule will be created that grants access. Visit the Lambda Developer Guide to learn more about resource policies and access controls for Lambda functions.
Q: How do I control which Amazon DynamoDB table or Amazon Kinesis stream an AWS Lambda function can poll?

Access controls are managed through the Lambda function role. The role you assign to your Lambda function also determines which resource(s) AWS Lambda can poll on its behalf. Visit the Lambda Developer Guide to learn more.
Q: How do I control which Amazon SQS queue an AWS Lambda function can poll?

Access controls can be managed by the Lambda function role or a resource policy setting on the queue itself. If both policies are present, the more restrictive of the two permissions will be applied.
Q: Can I access resources behind Amazon VPC with my AWS Lambda function?

Yes. You can access resources behind Amazon VPC.
Q: How do I enable and disable the VPC support for my Lambda function?

To enable VPC support, you need to specify one or more subnets in a single VPC and a security group as part of your function configuration. To disable VPC support, you need to update the function configuration and specify an empty list for the subnet and security group. You can change these settings using the AWS APIs, CLI, or AWS Lambda Management Console.
Q: Can a single Lambda function have access to multiple VPCs?

No. Lambda functions provide access only to a single VPC. If multiple subnets are specified, they must all be in the same VPC. You can connect to other VPCs by peering your VPCs.
Q: Can Lambda functions in a VPC also be able to access the internet and AWS Service endpoints?

Lambda functions configured to access resources in a particular VPC will not have access to the internet as a default configuration. If you need access to external endpoints, you will need to create a NAT in your VPC to forward this traffic and configure your security group to allow this outbound traffic.
Q: What is Code Signing for AWS Lambda?

Code Signing for AWS Lambda offers trust and integrity controls that enable you to verify that only unaltered code from approved developers is deployed in your Lambda functions. You can use AWS Signer, a fully-managed code signing service, to digitally sign code artifacts and configure your Lambda functions to verify the signatures at deployment. Code Signing for AWS Lambda is currently only available for functions packaged as ZIP archives.

Q: How do I create digitally signed code artifacts?

You can create digitally signed code artifacts using a Signing Profile through the AWS Signer console, the Signer API, SAM CLI or AWS CLI. To learn more, please see the documentation for AWS Signer.

Q: How do I configure my Lambda functions to enable code signing?

You can enable code signing by creating a Code Signing Configuration through the AWS Management Console, the Lambda API, the AWS CLI, AWS CloudFormation, and AWS SAM. Code Signing Configuration helps you specify the approved signing profiles and configure whether to warn or reject deployments if signature checks fail. Code Signing Configurations can be attached to individual Lambda functions to enable the code signing feature. Such functions now start verifying signatures at deployment.

Q: What signature checks does AWS Lambda perform on deployment?

AWS Lambda can perform the following signature checks at deployment:

• Corrupt signature - This occurs if the code artifact has been altered since signing.
• Mismatched signature - This occurs if the code artifact is signed by a signing profile that is not approved.
• Expired signature - This occurs if the signature is past the configured expiry date.
• Revoked signature - This occurs if the signing profile owner revokes the signing jobs.

To learn more, please see the AWS Lambda documentation.

Q: Can I enable code signing for existing functions?

Yes, you can enable code signing for existing functions by attaching a code signing configuration to the function. You can do this using the AWS Lambda console, the Lambda API, the AWS CLI, AWS CloudFormation, and AWS SAM.

Q: Is there any additional cost for using Code Signing for AWS Lambda?

There is no additional cost when using Code Signing for AWS Lambda. You pay the standard price for AWS Lambda.

AWS Lambda functions in Java

Q: How do I compile my AWS Lambda function Java code?

You can use standard tools like Maven or Gradle to compile your Lambda function. Your build process should mimic the same build process you would use to compile any Java code that depends on the AWS SDK. Run your Java compiler tool on your source files and include the AWS SDK 1.9 or later with transitive dependencies on your classpath. For more details, see our documentation.
Q: What is the JVM environment that Lambda uses for executing my function?

Lambda provides the Amazon Linux build of openjdk 1.8.

AWS Lambda functions in Node.js

Q: Can I use packages with AWS Lambda?

Yes. You can use NPM packages as well as custom packages. Learn more here.
Q: Can I execute other programs from within my AWS Lambda function written in Node.js?

Yes. Lambda's built-in sandbox lets you run batch ("shell") scripts, other language runtimes, utility routines, and executables. Learn more here.
Q: Is it possible to use native modules with AWS Lambda functions written in Node.js?

Yes. Any statically linked native module can be included in the ZIP file you upload, as well as dynamically linked modules compiled with an rpath pointing to your Lambda function root directory. Learn more here.
Q: Can I execute binaries with AWS Lambda written in Node.js?

Yes. You can use Node.js' child_process command to execute a binary that you included in your function or any executable from Amazon Linux that is visible to your function. Alternatively several NPM packages exist that wrap command line binaries such as node-ffmpeg. Learn more here.
Q: How do I deploy AWS Lambda function code written in Node.js?

To deploy a Lambda function written in Node.js, simply package your Javascript code and dependent libraries as a ZIP. You can upload the ZIP from your local environment, or specify an Amazon S3 location where the ZIP file is located. For more details, see our documentation.

AWS Lambda functions in Python

Q: Can I use Python packages with AWS Lambda?

Yes. You can use pip to install any Python packages needed.

AWS Lambda functions in C#

Q: How do I package and deploy an AWS Lambda function in C#?

You can create a C# Lambda function using the Visual Studio IDE by selecting "Publish to AWS Lambda" in the Solution Explorer. Alternatively, you can directly run the "dotnet lambda publish" command from the dotnet CLI, which has the [# Lambda CLI tools patch] installed, which creates a ZIP of your C# source code along with all NuGet dependencies as well as your own published DLL assemblies, and automatically uploads it to AWS Lambda using the runtime parameter “dotnetcore1.0”

AWS Lambda functions in PowerShell

Q: How do I deploy AWS Lambda function code written in PowerShell?

A PowerShell Lambda deployment package is a ZIP file that contains your PowerShell script, PowerShell modules that are required for your PowerShell script, and the assemblies needed to host PowerShell Core. You then use the *AWSLambdaPSCore* PowerShell module that you can install from the PowerShell Gallery to create your PowerShell Lambda deployment package.

Q: How do I deploy AWS Lambda function code written in PowerShell? A PowerShell Lambda deployment package is a ZIP file that contains your PowerShell script, PowerShell modules that are required for your PowerShell script, and the assemblies needed to host PowerShell Core. You then use the *AWSLambdaPSCore* PowerShell module that you can install from the PowerShell Gallery to create your PowerShell Lambda deployment package.
Q: How do I deploy AWS Lambda function code written in PowerShell? A PowerShell Lambda deployment package is a ZIP file that contains your PowerShell script, PowerShell modules that are required for your PowerShell script, and the assemblies needed to host PowerShell Core. You then use the *AWSLambdaPSCore* PowerShell module that you can install from the PowerShell Gallery to create your PowerShell Lambda deployment package.

AWS Lambda functions in Go

Q: How do I package and deploy an AWS Lambda function in Go?

Upload your Go executable artifact as a ZIP file through the AWS CLI or Lambda console and select the go1.x runtime. With Lambda, you can use Go's native tools to build and package your code. For more details, read our documentation.

AWS Lambda functions in Ruby

Q: How do I deploy AWS Lambda function code written in Ruby?

To deploy a Lambda function written in Ruby, package your Ruby code and gems as a ZIP. You can upload the ZIP from your local environment, or specify an Amazon S3 location where the ZIP file is located.

Other topics

Q: Which versions of Amazon Linux, Node.js, Python, JDK, .NET Core, SDKs, and additional libraries does AWS Lambda support?

You can view the list of supported versions here.
Q: Can I change the version of Amazon Linux or any language runtime?

No. AWS Lambda offers a single version of the operating system and managed language runtime to all users of the service. You can bring your own language runtime to use in Lambda.
Q: How can I record and audit calls made to the AWS Lambda API?

AWS Lambda is integrated with AWS CloudTrail. AWS CloudTrail can record and deliver log files to your Amazon S3 bucket describing the API usage of your account.
Q: How do I coordinate calls between multiple Lambda functions?

You can use Amazon Step Functions to coordinate multiple invoking Lambda functions. You can invoke multiple Lambda functions serially, passing the output of one to the other, or in parallel. See our documentation for more details.
Q: Does AWS Lambda support Advanced Vector Extensions 2 (AVX2)?

Yes, AWS Lambda supports the Advanced Vector Extensions 2 (AVX2) instruction set. To learn more about how to compile your application code to target this instruction set for improved performance, visit the AWS Lambda developer documentation.

AWS Serverless Application Repository

Discover, deploy, and publish serverless applications.

The AWS Serverless Application Repository is a managed repository for serverless applications. It enables teams, organizations, and individual developers to store and share reusable applications, and easily assemble and deploy serverless architectures in powerful new ways. Using the Serverless Application Repository, you don't need to clone, build, package, or publish source code to AWS before deploying it. Instead, you can use pre-built applications from the Serverless Application Repository in your serverless architectures, helping you and your teams reduce duplicated work, ensure organizational best practices, and get to market faster. Integration with AWS Identity and Access Management (IAM) provides resource-level control of each application, enabling you to publicly share applications with everyone or privately share them with specific AWS accounts. To share an application you've built, publish it to the AWS Serverless Application Repository.

Each application is packaged with an AWS Serverless Application Model (SAM) template that defines the AWS resources used. Publicly shared applications also include a link to the application's source code. There is no additional charge to use the Serverless Application Repository - you only pay for the AWS resources used in the applications you deploy.

How it works: Deploying applications

BENEFITS

Develop More Powerful Apps

Quickly assemble serverless architectures in powerful new ways. Discover and share reusable serverless application patterns, privately or publicly, and compose new serverless architectures using the simplified syntax of AWS SAM.

Easily Manage Applications

Use pre-built applications in your serverless deployments, eliminating the need to clone, build, package, and publish source code to AWS before deploying it. It also supports SAM and semantic versioning to enable simple application management.

Reuse, Don't Rebuild

Develop and publish serverless applications once, store them in the Serverless Application Repository, and use them privately across teams or with the greater community to reduce duplicated efforts and accelerate development workflows.

Ensure Best Practices

Find and distribute serverless applications for common use cases. Build organizational best practices into your serverless architectures to help ensure consistency across teams—use permissions to share applications with specific AWS accounts.

What Is the AWS Serverless Application Repository?

The AWS Serverless Application Repository makes it easy for developers and enterprises to quickly find, deploy, and publish serverless applications in the AWS

Cloud. For more information about serverless applications, see Serverless Computing and Applications on the AWS website.

You can easily publish applications, sharing them publicly with the community at large, or privately within your team or across your organization. To publish a serverless application (or app), you can use the AWS Management Console, the AWS SAM command line interface (AWS SAM CLI), or AWS SDKs to upload your code. Along with your code, you upload a simple manifest file, also known as an AWS Serverless Application Model (AWS SAM) template. For more information about AWS SAM, see the *AWS Serverless Application Model Developer Guide*.

The AWS Serverless Application Repository is deeply integrated with the AWS Lambda console. This integration means that developers of all levels can get started with serverless computing without needing to learn anything new. You can use category keywords to browse for applications such as web and mobile backends, data processing applications, or chatbots. You can also search for applications by name, publisher, or event source. To use an application, you simply choose it, configure any required fields, and deploy it with a few clicks.

In this guide, you can learn about the two ways to work with the AWS Serverless Application Repository:

- Publishing Applications – Configure and upload applications to make them available to other developers, and publish new versions of applications.
- Deploying Applications – Browse for applications and view information about them, including source code and readme files. Also install, configure, and deploy applications of your choosing.

AWS Serverless Application Repository FAQs and Terms

GENERAL

What are serverless applications?

Serverless applications eliminate the need to provision, deploy, or manage servers or other infrastructure. They come with built-in high availability and they scale continuously and automatically. You can use one of the many fully managed AWS services to build and run serverless applications, including AWS Lambda for compute, Amazon API Gateway for APIs, and Amazon DynamoDB for databases.

What is the AWS Serverless Application Repository?

The AWS Serverless Application Repository offers a growing collection of serverless applications published by AWS and other AWS partners and developers. The Serverless Application Repository makes it easy to deploy applications for common use cases like web and mobile back-ends, stream processing, machine learning, and more, so you can quickly get started with the AWS Serverless platform.

In which regions is the AWS Serverless Application Repository available?

It is currently available in the following regions: US East (Ohio), US East (N. Virginia), US West (N. California), US West (Oregon), Asia Pacific (Tokyo), Asia Pacific (Seoul), Asia Pacific (Mumbai), Asia Pacific (Singapore), Asia Pacific (Sydney), Canada (Central), EU (Frankfurt), EU (Ireland), EU (London), and South America (São Paulo).

What kinds of applications are available in the AWS Serverless Application Repository?

The AWS Serverless Application Repository includes applications for Alexa Skills, chatbots, data processing, IoT, real time stream processing, web and mobile back-ends, social media trend analysis, image resizing, and more from publishers on AWS.

Does it cost anything to use the AWS Serverless Application Repository?

There is no charge to browse applications, deploy applications, or publish applications in the AWS Serverless Application Repository. Standard AWS charges apply to the resources used by the applications you deploy.

How are applications in the AWS Serverless Application Repository licensed?

Applications provided by AWS are available under the MIT open source license, while applications publicly provided by others can be made available under one or more open source license(s) approved by the Open Source Initiative (OSI). You can find licensing details by navigating to the licensing tab of an application on the AWS website.

Are applications in the AWS Serverless Application Repository verified by AWS?

All applications published by AWS are vetted for license adherence and code quality. Applications published by third parties are validated for correct use of permissions to ensure that consumers know which resources can be modified or accessed by an application. You can review those details, along with the number of times an application has been deployed by AWS customers, prior to using an application.

Can I use the AWS Serverless Application Repository in combination with GitHub?

Yes. The AWS Serverless Application Repository enables developers to publish serverless applications developed in a GitHub repository. Using AWS CodePipeline to link a GitHub source with the AWS Serverless Application Repository can make the publishing process even easier, and the process can be set up in minutes.

DEPLOYING APPLICATIONS

How do I manage serverless applications deployed to my account?

Serverless applications are deployed as AWS CloudFormation stacks, which make it easy for you to manage an application as a single unit. Each resource is tagged with the application's uniquely identifiable Amazon Resource Name (ARN), which helps you locate the resources using

the AWS Tag Editor console. You can also use existing AWS and third-party tools to manage each resource separately.

PUBLISHING APPLICATIONS

How do I publish a serverless application to the AWS Serverless Application Repository?

To publish a serverless application, describe the application using the AWS Serverless Application Model (SAM) format, package it using the AWS CLI, and publish it using the AWS Management Console, AWS CLI, or AWS SDKs. You must have a valid AWS account, and you will be required to provide a name, description, source code link, and a LICENSE.txt for your application. Applications that are intended to be made available to any AWS customer must be published in either the US East (N. Virginia) or US East (Ohio) regions.

Who can deploy the applications I publish to the AWS Serverless Application Repository?

Publishers control who can find and deploy their applications in the Repository. You can limit access to your team (a group under the same AWS account), share an application with other accounts, or make it publicly available to any AWS customer. To learn more about the permissions and visibility of applications you publish, review our Resource-Based Policies documentation.

Can I charge a fee for serverless applications I publish?

If you would like to charge a fee for the use of a serverless application, you can integrate AWS Lambda behind Amazon API Gateway, and then sell the API as a SaaS product through the AWS Marketplace. To learn more, review the steps to list on AWS Marketplace and how to monetize your API using API Gateway.

USING NESTED APPLICATIONS

What is a nested application?

A nested serverless application is a component that is deployed as part of another serverless application. As serverless architectures grow, common patterns emerge in which the same components are defined in multiple application templates. You can now separate out common patterns as dedicated applications and then nest them as part of new or existing application templates.

How are nested applications deployed?

Nested applications are deployed as AWS CloudFormation nested stacks.

How do I include a nested application in my SAM template?

Use the AWS:Serverless:Application resource type to add applications that you have developed locally or that have been shared with you via the Serverless Application Repository. You can define inputs for the nested application or reference outputs using simple AWS SAM syntax. For more details, see our documentation.

How do I handle resource name conflicts when nesting applications?

Specify a unique name for each nested application in your SAM template. Resources created by the nested application will be scoped to the application's unique name.

How do I package a nested application?

Use the SAM CLI sam package command to package your nested application. Publish your packaged application to the Serverless Application Repository using APIs or the AWS SDK. You can use the SAM CLI sam deploy command to deploy your application immediately.

Can I package a hierarchy of nested applications using SAM CLI?

Yes. The SAM CLI sam package command will recursively package a hierarchy of nested applications. You can have a maximum of 199 applications nested in a single top-level application template.

Consuming Nested Applications

How do I manage serverless applications deployed to my account?

Serverless applications are deployed as AWS CloudFormation stacks, which make it easy for you to manage an application as a single unit. Each resource is tagged with the application's uniquely identifiable Amazon Resource Name (ARN), which helps you locate the resources using the AWS Tag Editor console. You can also use existing AWS and third-party tools to manage each resource separately.

How do I nest applications shared with me via the Serverless Application Repository?

Configure the AWS:Serverless:Application SAM resource with the application ARN and any necessary input parameters required by the application. For more details, see our documentation.

Q: How do I package a nested app that I used from the Serverless Application Repository?

Nested applications from the Serverless Application Repository are already packaged and ready for you to use. You can use the existing SAM CLI sam package command to ensure that nested applications are still available to you before you deploy the application in your account.

What happens if an application that I nested is no longer available?

Applications that you nest which get deleted or that you no longer have access to will have no impact on the existing deployments of your application. You will be required to update your application if you attempt to deploy it with dependencies that no longer exist.

How can I tell if an application contains other nested applications?

Applications that contain one or more nested applications use CAPABILITY_AUTO_EXPAND.

Publishing Nested Applications

Can I publish nested applications to the Serverless Application Repository?

Yes. You can publish nested applications to the Serverless Application Repository just like you publish stand-alone applications today. Use the AWS console, AWS CLI, SAM CLI, or Serverless Application Repository APIs to publish nested applications.

Sharing Nested Applications

How do I share an application that contains nested applications?

Serverless applications follow the same model used by Lambda functions today. Access to serverless applications in the Serverless Application Repository can be controlled using AWS IAM resource-based policies. You can use policies to keep your application private, grant cross-account access, or make it publicly available.

How does sharing work when there is a hierarchy of nested applications?

With a hierarchy of nested applications, each nested application has to be made available to the AWS account with whom you are sharing the top-level application.

SHARING APPLICATIONS

Can I share applications with all accounts that belong to an AWS Organization?

Yes. You can now share application with all accounts that belong to an AWS Organization. Serverless apps follow the same model used by Lambda functions today. Access to serverless applications in the AWS Serverless Application Repository can be controlled using AWS IAM resource-based policies. You can use policies to keep your app private, grant cross-account access, grant organization access or make it available publicly. To learn more about resource-based policies for serverless applications, see here.

How do I update resource-based policies to share an app with an AWS organization?

To get started, find the application in the AWS Serverless Application Repository console, click on the 'Share' tab and then click on 'Create Statement'. Here you can add the AWS Organization principal to grant the organization with access to the app. You can also use the AWS Serverless Application Repository APIs to update resource-based policies to share apps across an organization. For more details on resource based polices, see our documentation here.

Can I share an application with multiple AWS Organizations?

No. You can share an application only with accounts that belong to the same AWS Organization as your account.

Can I share an application with a particular organizational unit within my organization?

Sharing of applications within an organizational unit isn't supported. You can use policies to keep your app private, grant cross-account access, grant organization access or make it available publicly.

How do I share an application with specific accounts in an organization?

To provide access to your application for certain accounts in an organization, simply update the resource-based policy to include AWS accounts along with the AWS organization ID with whom you would like to share the application.

Can I update permissions for an organization to only view an application?

Yes. You can set actions on your resource-based policy which can restrict the type of operations someone can take on an application that you have shared. Updating actions for a resource-based policy is supported through the AWS Serverless Application Repository APIs and console. For more details on actions you can set for resource-based polices, see our documentation here.

AWS Outposts

Run AWS infrastructure and services on premises for a truly consistent hybrid experience.

AWS Outposts Family

AWS Outposts Rack

AWS Outposts Servers

What is AWS Outposts?

AWS Outposts is a fully managed service that extends AWS infrastructure, services, APIs, and tools to customer premises. By providing local access to AWS managed infrastructure, AWS Outposts enables customers to build and run applications on premises using the same programming interfaces as in AWS Regions, while using local compute and storage resources for lower latency and local data processing needs.

An Outpost is a pool of AWS compute and storage capacity deployed at a customer site. AWS operates, monitors, and manages this capacity as part of an AWS Region. You can create subnets on your Outpost and specify them when you create AWS resources such as EC2 instances, EBS volumes, ECS clusters, and RDS instances. Instances in Outpost subnets communicate with other instances in the AWS Region using private IP addresses, all within the same VPC.

Key concepts

These are the key concepts for AWS Outposts.

- **Outpost site** – The customer-managed physical buildings where AWS will install your Outpost. A site must meet the facility, networking, and power requirements for your Outpost.

- **Outpost configurations** – Configurations of Amazon EC2 compute capacity, Amazon EBS storage capacity, and networking support. Each configuration has unique power, cooling, and weight support requirements.
- **Outpost capacity** – Compute and storage resources available on the Outpost. You can view and manage the capacity for your Outpost from the AWS Outposts console.
- **Outpost equipment** – Physical hardware that provides access to the AWS Outposts service. The hardware includes racks, servers, switches, and cabling owned and managed by AWS.
- **Outpost racks** – An Outpost form factor that is an industry-standard 42U rack. Outpost racks include rack-mountable servers, switches, a network patch panel, a power shelf and blank panels.
- **Outpost servers** – An Outpost form factor that is an industry-standard 1U or 2U server, which can be installed in a standard EIA-310D 19 compliant 4 post rack. Outpost servers provide local compute and networking services to sites that have limited space or smaller capacity requirements.
- **Service link** – Network route that enables communication between your Outpost and its associated AWS Region. Each Outpost is an extension of an Availability Zone and its associated Region.
- **Local gateway** – A logical interconnect virtual router that enables communication between an Outpost rack and your on-premises network.
- **Local network interface** – A network interface that enables communication from an Outpost server and your on-premises network.

How AWS Outposts works

AWS Outposts is designed to operate with a constant and consistent connection between your Outpost and an AWS Region. To achieve this connection to the Region, and to the local workloads in your on-premises environment, you must connect your Outpost to your on-premises network. Your on-premises network must provide wide area network (WAN) access back to the Region and to the internet. It must also provide LAN or WAN access to the local network where your on-premises workloads or applications reside.

The following diagram illustrates both Outpost form factors.

Network components

AWS Outposts extends an Amazon VPC from an AWS Region to an Outpost with the VPC components that are accessible in the Region, including internet gateways, virtual private gateways, Amazon VPC Transit Gateways, and VPC endpoints. An Outpost is homed to an Availability Zone in the Region and is an extension of that Availability Zone that you can use for resiliency.

The following diagram shows the network components for your Outpost.

- An AWS Region and an on-premises network
- A VPC with multiple subnets in the Region
- An Outpost in the on-premises network
- A local gateway for racks, or a local network interface for servers

VPCs and subnets

A virtual private cloud (VPC) spans all Availability Zones in its AWS Region. You can extend any VPC in the Region to your Outpost by adding an Outpost subnet. To add an Outpost subnet to a VPC, specify the Amazon Resource Name (ARN) of the Outpost when you create the subnet.

Outposts support multiple subnets. You can specify the EC2 instance subnet when you launch the EC2 instance in your Outpost. You cannot specify the underlying hardware where the instance is deployed, because the Outpost is a pool of AWS compute and storage capacity.

Each Outpost can support multiple VPCs that can have one or more Outpost subnets. For information about VPC quotas, see Amazon VPC Quotas in the *Amazon VPC User Guide*.

You create Outpost subnets from the VPC CIDR range of the VPC where you created the Outpost. You can use the Outpost address ranges for resources, such as EC2 instances that reside in the Outpost subnet.

Routing

By default, every Outpost subnet inherits the main route table from its VPC. You can create a custom route table and associate it with an Outpost subnet.

The route tables for Outpost subnets work as they do for Availability Zone subnets. You can specify IP addresses, internet gateways, local gateways, virtual private gateways, and peering connections as destinations. For example, each Outpost subnet, either through the inherited main route table, or a custom table, inherits the VPC local route. This means that all traffic in the VPC, including the Outpost subnet with a destination in the VPC CIDR remains routed in the VPC. You cannot configure a more specific range than the VPC CIDR local route on the Outpost for Outpost subnets.

Outpost subnet route tables can include the following destinations:

- **VPC CIDR range** – AWS defines this at installation. This is the local route and applies to all VPC routing, including traffic between Outpost instances in the same VPC.
- **AWS Region destinations** – This includes prefix lists for Amazon Simple Storage Service (Amazon S3), Amazon DynamoDB gateway endpoint, AWS Transit Gateways, virtual private gateways, internet gateways, and VPC peering.

  If you have a peering connection with multiple VPCs on the same Outpost, the traffic between the VPCs remains in the Outpost and does not use the service link back to the Region.

DNS

For network interfaces connected to a VPC, EC2 instances in Outposts subnets can use the Amazon Route 53 DNS Service to resolve domain names to IP addresses. Route 53 supports DNS features, such as domain registration, DNS routing, and health checks for instances running in your Outpost. Both public and private hosted Availability Zones are supported for routing traffic to specific domains. Route 53 resolvers are hosted in the AWS Region. Therefore, service link connectivity from the Outpost back to the AWS Region must be up and running for these DNS features to work.

You might encounter longer DNS resolution times with Route 53, depending on the path latency between your Outpost and the AWS Region. In such cases, you can

use the DNS servers installed locally in your on-premises environment. To use your own DNS servers, you must create DHCP option sets for your on-premises DNS servers and associate them with the VPC. You must also ensure that there is IP connectivity to these DNS servers. You might also need to add routes to the local gateway routing table for reachability but this is only an option for Outpost racks with local gateway. Because DHCP option sets have a VPC scope, instances in both the Outpost subnets and the Availability Zone subnets for the VPC will try to use the specified DNS servers for DNS name resolution.

Query logging is not supported for DNS queries originating from an Outpost.

Service link

The service link is a connection from your Outpost back to your chosen AWS Region or Outposts home Region. The service link is an encrypted set of VPN connections that are used whenever the Outpost communicates with your chosen home Region. You use a virtual LAN (VLAN) to segment traffic on the service link. The service link VLAN enables communication between the Outpost and the AWS Region for both management of the Outpost and intra-VPC traffic between the AWS Region and Outpost.

Your service link is created when your Outpost is provisioned. If you have a server form factor, you create the connection. if you have a rack, AWS creates the service link. For more information, see Service link.

Local gateways

Outpost racks include a local gateway to provide connectivity to your on-premises network. If you have an Outpost rack, you can include a local gateway as target where the destination is your on-premises network. Local gateways are only available for Outpost racks and can only be used in VPC and subnet route tables that are associated with an Outpost rack. For more information, see Local gateway.

Local network interfaces

Outpost servers include a local network interface to provide connectivity to your on-premises network. A local network interface is available only for Outposts servers running on an Outpost subnet. You cannot use a local network interface from an EC2 instance on an Outpost rack or in the AWS Region. The local network interface

is meant only for on-premises locations. For more information, see Local network interface in the AWS Outposts User Guide for Outpost servers.

AWS Outposts Family

AWS Outposts Family overview

AWS Outposts is a family of fully managed solutions delivering AWS infrastructure and services to virtually any on-premises or edge location for a truly consistent hybrid experience. Outposts solutions allow you to extend and run native AWS services on premises, and is available in a variety of form factors, from 1U and 2U Outposts servers to 42U Outposts racks, and multiple rack deployments.

With AWS Outposts, you can run some AWS services locally and connect to a broad range of services available in the local AWS Region. Run applications and workloads on premises using familiar AWS services, tools, and APIs. Outposts supports workloads and devices requiring low latency access to on-premises systems, local data processing, data residency, and application migration with local system interdependencies.

Benefits

Run AWS Services on premises
Extend AWS compute, networking, security, and other services on premises for low latency, local data processing, and data residency needs.

Fully managed infrastructure
Reduce the time, resources, operational risk, and maintenance downtime required to manage IT infrastructure with a fully managed experience.

Truly consistent hybrid experience
Use the same hardware infrastructure, APIs, tools, and management controls available in the cloud to provide a truly consistent developer and IT operations experience.

AWS Outposts Family members

AWS Outposts rack

The AWS Outposts rack is an industry standard 42U form factor. It provides the same AWS infrastructure, services, APIs, and tools to virtually any datacenter or co-location space. Outposts rack provides AWS compute, storage, database, and other services locally, while still allowing you to access the full range of AWS services available in the Region for a truly consistent hybrid experience. Scale from a single 42U rack to multiple rack deployments of up to 96 racks to create pools of compute and storage capacity.

AWS Outposts servers

The AWS Outposts servers come in a 1U or 2U form factor. They provide the same AWS infrastructure, services, APIs, and tools to on-premises and edge locations that have limited space or smaller capacity requirements, such as retail stores, branch offices, healthcare provider locations, or factory floors. Outposts servers provide local compute and networking services.

Feature comparison

	AWS Outposts rack	AWS Outposts servers
Form factors	The Outposts rack is 80 inches tall, 24 inches wide, and 48 inches deep. Inside are hosts, switches, a network patch panel, a power shelf, and blank panels.	The Outposts rack-mountable servers fit inside 19" width, EIA-310 cabinets. The 1U high server is 24” deep, and uses AWS Graviton2 processors. The 2U high server is 30” deep and uses 3rd generation Intel Xeon Scalable processors.
Installation	AWS delivers Outposts racks fully assembled and ready to be rolled into final position. Racks are installed by AWS and simply need to be plugged into power and network.	AWS delivers Outposts servers directly to you, installed by either onsite personnel install or a 3rd-party vendor. Once connected to your network, AWS will remotely provision compute and storage resources.
Locally supported services	Amazon Elastic Compute Cloud (EC2), Amazon Elastic Container Service (ECS), Amazon Elastic Kubernetes Service (EKS), Amazon Elastic Block Store (EBS), Amazon EBS Snapshots, Amazon Simple Storage Service (S3), Amazon Relational Database	Amazon EC2, Amazon ECS, AWS IoT Greengrass, or Amazon Sagemaker Edge Manager. Seamlessly extend Amazon Virtual Private Cloud on premises and run select AWS services locally on Outposts servers, and connect to a broad range of services available in the AWS Region.

	Service (RDS), Amazon Elasticache, Amazon EMR, Application Load Balancer (ALB), CloudEndure, and VMware Cloud. Seamlessly extend Amazon Virtual Private Cloud on premises and run select AWS services locally on Outposts rack, and connect to a broad range of services available in the AWS Region.	
Networking	• 1 Gbps, 10 Gbps, 40 Gbps, and 100 Gbps uplink speeds. • Includes integrated networking gear. • Supports Local Gateway, which requires Border Gateway Protocol (BGP) over a routed network.	• 1 Gbps and 10 Gbps uplink speeds. • Does not include integrated networking gear. • Supports a simplified network integration experience providing a local Layer 2 presence.
Power	• Supports three power configurations: 5 kVA, 10 kVA, or 15 kVA. The configuration of the power shelf depends on the total power draw of the Outpost capacity. • Centralized redundant power conversion unit and a direct current (DC) distribution system in the backplane handled by line mate connectors.	• Requires 1-2 kVA of power. • Supports standard alternating current (AC) and direct current (DC) power options.

Use cases

Low latency compute
Deliver high-quality gaming experiences for interactive applications, like real-time multiplayer games, to players all over the world. When the nearest public cloud servers are not close enough to meet single-digit millisecond latency requirements, AWS Outposts can help run business applications where you need them for manufacturing execution systems (MES), high-frequency trading, or medical diagnostics.

Data residency
Data sometimes needs to remain in a particular country, state, or municipality for regulatory, contractual, or information security reasons. This is often the case with financial services, healthcare, oil and gas, and other highly regulated industries. With AWS Outposts, you can control where your workloads run and where your data resides, with low-friction movement between cloud and edge locations to easily adapt to regulatory changes.

Migration and modernization
Legacy on-premises applications often have latency-sensitive system dependencies, making them difficult to migrate. AWS Outposts allows you to segment migrations into smaller pieces on premises, maintaining latency sensitive connectivity between application components until you are ready to migrate.

Local data processing
Process data locally for use cases such as data lakes and machine learning (ML) model training, or set up a consistent hybrid architecture to process local, difficult-to-migrate data sets with cost, size, or bandwidth constraints and move data to the cloud for long-term archival.

AWS Outposts Rack

How it works

AWS Outposts rack is a fully managed service that extends AWS infrastructure, services, APIs, and tools on premises for a truly consistent hybrid experience.

Use cases

Deliver low-latency compute

Run applications where you need them. Deliver single-digit millisecond latency for on-premises applications such as manufacturing execution systems (MES), high-frequency trading, and medical diagnostics.

Securely store data locally to meet data residency requirements
Control where your applications run and where your data resides. Maintain data control on premises to meet legal, industry, or contractual requirements.

Process data locally
Set up a consistent hybrid cloud architecture to process data on premises due to cost, size, bandwidth, or timing constraints.

Ease and accelerate cloud migration
Support on-premises hybrid cloud migration of applications with local system interdependencies or applications with data residency requirements.

AWS Outposts Servers

How it works

AWS Outposts servers deliver compute and networking services for locations with space and capacity constraints.

Use cases

Improve manufacturing operations performance
Reduce manufacturing errors by leveraging low-latency compute, AWS IoT, and machine learning (ML) services.

Uncover insights at retail locations
Run low-latency sales systems at retail locations to perform real-time analyses of store operations.

Power next-generation wireless networks
Manage, update, and scale high-performance and low-latency 5G networks.

Accelerate medical diagnoses
Analyze real-time equipment data and leverage ML inference to identify trends and conditions in patient monitoring devices.

AWS Outposts rack FAQs

General

Q: Why would I use AWS Outposts rack instead of operating in an AWS Region?

You can use Outposts rack to support your applications that have low latency or local data processing requirements. These applications may need to generate near real-time responses to end user applications or need to communicate with other on-premises systems or control on-site equipment. These can include workloads running on factory floors for automated operations in manufacturing, real-time patient diagnosis or medical imaging, and content and media streaming. You can use Outposts rack to securely store and process customer data that needs to remain on premises or in countries where there is no AWS Region. You can run data intensive workloads on Outposts rack and process data locally when transmitting data to AWS Regions is expensive and wasteful and for better control on data analysis, backup and restore.

Q: In which countries and territories will Outposts rack be available?

Outposts rack can be shipped to and installed in the following countries and territories.

- NA - US, Canada, Mexico
- EMEA - All EU countries, United Kingdom (UK), Switzerland, Norway, Bahrain, United Arab Emirates (UAE), Kingdom of Saudi Arabia (KSA)*, Israel, South Africa, Gibraltar, Morocco, Nigeria, Kenya, Oman, Kazakhstan, Serbia
- APAC - Australia, New Zealand, Japan, South Korea, Hong Kong Special Administrative Region, Macau, Taiwan, Singapore, Indonesia, Malaysia, Thailand, the Philippines, Brunei, India, Vietnam, Bangladesh
- SA - Brazil, Colombia**, Argentina***, Chile, Peru, Ecuador
- CA - Puerto Rico, Costa Rica, Panama

Support for more countries and territories are coming soon.

* Outposts rack availability in KSA is pending authorization.
** Outposts rack is available in Bogota and Valle del Cauca with support for additional departments coming soon.
*** Outposts rack is available in Buenos Aires and Santa Fe provinces with support for additional provinces coming soon.

Q: Can I order an Outpost to a country or territory where Outposts rack has not launched and link it back to a supported Region?

No, we can deliver and install Outposts rack only in countries and territories where Outposts rack can be delivered and supported.

Q: Can I use Outposts rack when it is not connected to the AWS Region or in a disconnected environment?

An Outpost relies on connectivity to the parent AWS Region. Outposts rack is not designed for disconnected operations or environments with limited to no connectivity. We recommend that customers have highly available networking connections back to their AWS Region. If interested in leveraging AWS services in disconnected environments such as cruise ships or remote mining locations, learn more about AWS services such as Snowball Edge, which is optimized to operate in environments with limited to no connectivity.

Q: Can I reuse my existing servers in an Outpost?

No, AWS Outposts rack leverages AWS designed infrastructure, and is only supported on AWS-designed hardware that is optimized for secure, high-performance, and reliable operations.

Q: Is there a software-only version of AWS Outposts rack?

No, AWS Outposts rack is a fully managed service that provides you with native access to AWS services.

Q: Can I order my own hardware that can be installed as part of my Outposts rack?

No, AWS Outposts rack provides fully integrated AWS designed configurations with built in top-of-rack switches and redundant power supply to ensure an ideal AWS experience. You can order as much compute and storage infrastructure as you need by selecting from the range of available Outposts rack options, or work with us to create a custom combination with your desired Amazon Elastic Compute Cloud (EC2), Amazon Elastic Block Store (EBS), and Amazon Simple Storage Service (S3) capacity. These are pre-validated and tested to ensure that you can get started quickly with no additional effort or configuration required on-site.

AWS services

Q: Can I create EC2 instances using an EBS backed AMI on my Outposts?

Yes, you can launch EC2 instances using the AMIs backed with EBS gp2 volume types.

Q: Where are EBS snapshots stored?

EBS snapshots of EBS Volumes on Outposts rack are stored by default on Amazon S3 in the Region. If the Outpost is provisioned with Amazon S3 on Outposts you have the option to store your snapshots locally on your Outpost. EBS snapshots are incremental, which means that only the blocks on your Outpost that have changed after your most recent snapshot are saved. You can at any time restore (hydrate) EBS Volume on Outposts from the stored snapshots. To learn more, visit the EBS Snapshots documentation.

Q: How do I enforce data residency on Outposts rack using EBS Local Snapshots on Outposts?

You can set resource-level IAM policies and permissions on your Outpost for EBS Local Snapshots on Outposts and AMI images to enforce data residency. Each account (or IAM user or role) can set up a data residency enforcement policy for single or multiple Outposts. This will block the creation or copy of local snapshots and images as well as API/CLI calls outside the specified Outposts rack ARN. All of the create, update, and copy operations are logged in CloudTrail audit logs, so you can track that local snapshots reside in your location. To learn more, visit the EBS Snapshots documentation.

Q: What use cases are best suited to run on S3 on Outposts?

S3 on Outposts is ideal for customers with data residency requirements or those in regulated industries that need to securely store and process customer data that needs to remain on premises or in locations where there is no AWS region. Additionally, customers can use S3 on Outposts to run data intensive workloads to process data locally and store on-premises. S3 on Outposts will also help if you have applications that need to communicate with other on-premises systems or control on-site equipment, such as within a factory, hospital, or research facility.

Q: How can I establish network connectivity between my Outpost and the AWS Region?

You can choose to establish Outposts rack service link VPN connection to the parent AWS Region via an AWS Direct Connect private connection, a public virtual interface, or the public Internet.

Q: Is Application Load Balancer available on Outposts rack?

Yes, Application Load Balancer is available on Outposts rack in all AWS Regions where Outposts rack is available, except the AWS GovCloud (US) Regions.

Q: Can Outposts rack support real-time applications with low-latency requirements?

Yes, with Amazon RDS on Outposts you can run managed Microsoft SQL Server, MySQL, and PostgreSQL databases on premises for low latency workloads that need to be run in close proximity to on premises data and applications. You can manage RDS databases both in the cloud and on premises using the same AWS Management Console, APIs, and CLI. For ultra low-latency applications, ElastiCache on Outposts enables sub-millisecond responses for real-time applications, including workloads running on factory floors for automated operations in manufacturing, real-time patient diagnosis, and media streaming.

Q: Can Outposts rack be used to meet data residency requirements?

Yes. Customer data can be configured to remain on Outposts rack using Amazon Elastic Block Store (EBS) and Amazon Simple Storage Service (S3) on Outposts, in the customer's on-premises location or specified co-location facility. Well-architected applications using Outposts rack and AWS services and tools address the data residency requirements we most commonly hear from our customers. AWS Identity and Access Management (IAM) lets you control access to AWS resources. You can use IAM and granular data control rules to specify which types of data must remain on Outposts rack and cannot be replicated to the AWS Region. S3 on Outposts stores data on your Outpost by default, and you may choose to replicate some or all of your data to AWS Regions based on your specific residency requirements. ElastiCache on Outposts allows you to securely process customer data locally on the Outposts rack. Some limited meta-data (e.g. instance IDs, monitoring metrics, metering records, tags, bucket names, etc.) will flow back to the AWS Region.

To ensure your unique data residency requirements are met, we recommend you confirm and work closely with your compliance and security teams.

Q: Is Resource Sharing available on AWS Outposts rack?

Yes. AWS Resource Access Manager (RAM) is a service that enables you to share your AWS resources with any AWS account or within your AWS Organization. RAM support lets you, the Outpost owner, create and manage Outpost resources — EC2 instances, EBS volumes, subnets, and local gateways (LGWs) centrally — and share the resources across multiple AWS accounts within the same AWS organization. This allows others to configure VPCs, launch and run instances, and create EBS volumes on the shared Outpost.

Q: Which EC2 instances are available on Outposts rack?

EC2 instances built on the AWS Nitro System, for general purpose, compute optimized, memory optimized, storage optimized, and GPU optimized with Intel Xeon Scalable processors are supported on AWS Outposts rack, and Graviton processors based EC2 instances are coming soon.

Getting started with ordering & installation

Q: Are there any prerequisites for deploying Outposts 42U racks at my location?

Your site must support the basic power, networking and space requirements to host an Outpost. An Outposts rack needs 5-15 kVA, can support 1/10/40/100 Gbps uplinks, and space for a 42U rack (80” X 24” X 48” dimensions). As Outposts rack requires reliable network connectivity to the AWS Region, you should plan for a public internet connection. Customers must have Enterprise Support or Enterprise On-Ramp Support, which provides 24x7 remote support within 15 minutes or within 30 minutes depending on the Support plan selected.

Security & compliance

Q: Do the same compliance certifications for AWS Services today apply for services on Outposts rack?

AWS Outposts rack itself is HIPAA eligible, PCI, SOC, ISMAP, IRAP and FINMA compliant, ISO, CSA STAR, and HITRUST certified, and we expect to add more compliance certifications in coming months. You can see the latest certification status for AWS Services on Outposts rack on our Services in Scope page. AWS Services on Outposts rack like RDS or Elasticache Redis that have achieved certifications like PCI are also considered certified on Outposts rack. As AWS Outposts rack runs at the customer’s data center, under the AWS Shared Responsibility model customers own the responsibility for physical security and access controls around the Outpost for compliance certification.

Q: Is AWS Outposts rack GxP Compatible?

Yes, AWS Outposts rack is GxP compatible. AWS Outposts rack extends AWS services to AWS-managed infrastructure that is physically located at a customer site. Outposts rack capacity can be accessed locally over a local gateway that is mapped to the customer’s local network, in addition to having a connection path back to the AWS Region. You can learn more about using the AWS Cloud, including Outposts rack, for GxP systems here. GxP-regulated life sciences organizations using AWS services are responsible for designing and verifying their GxP compliance.

Q: Who is responsible for the physical security of the Outposts rack at my datacenter?

AWS provides services that allow data to be encrypted at rest and in-transit and other granular security controls and auditing mechanisms. In addition, customer data is wrapped to a physical Nitro Security Key. Destroying the device is equivalent to destroying the data. As part of the shared responsibility model, customers are responsible for attesting to physical security and access controls around the Outpost, as well as environmental requirements for facility, networking, and power as published here. Prior to returning the Outposts rack hardware, the Nitro Security Key will be removed to ensure customer content is crypto shredded.

Q: What is AWS Outposts FedRAMP authorization status?

FedRAMP has granted authorization to AWS Outposts, excluding the hardware components of the service. The physical and environmental controls for the AWS Outposts protection while at the customers’ sites are the customers’ responsibilities. Customers and agencies can make a risk-based decision to use AWS Outposts or conduct a Type Accreditation to review the hardware components of AWS Outposts to process FedRAMP workloads. Assessment results previously conducted by the AWS Third Party Assessment Organization can be accessed via CapLinked.

Support & maintenance

Q: How does AWS maintain AWS Outposts rack infrastructure?

When your Outpost is installed and is visible in the AWS Management Console, AWS will monitor it as part of the public Region and will automatically execute software upgrades and patches.

If there is a need to perform physical maintenance, AWS will reach out to schedule a time to visit your site. AWS may replace a given module as appropriate but will not perform any host or network switch servicing on customer premises.

Q: What happens when my facility's network connection goes down?

EC2 instances and EBS volumes on the Outpost will continue to operate normally and can be accessed locally via the local gateway. Similarly, AWS service resources such as ECS worker nodes continue to run locally. However, API availability will be degraded, for instance run/start/stop/terminate APIs may not work. Instance metrics and logs will continue to be cached locally for a few hours, and will be pushed to the AWS Region when connectivity returns. Disconnection beyond a few hours however may result in loss of metrics and logs. At this time, DNS queries on the Outpost to the Route 53 Resolver (aka AmazonProvidedDNS) also rely on the network link to the AWS Region, so default DNS resolution will stop working. If you expect to lose network connectivity, we strongly recommend regularly testing your workload to ensure it behaves properly in this state when an Outpost is disconnected. For S3 on Outposts, if the network connection to your Outpost is lost, you will not be able to access your objects. Requests to store and retrieve objects are authenticated using the regional AWS Identity and Access Management (IAM) service, and if the Outpost has no connectivity to the home AWS Region, you are not able to access your data. Your data remains safely stored on your Outpost during periods of disconnect, and once connectivity is restored authentication and requests can resume.

Q: What type of control plane information flows back to the parent AWS Region?

As an example, information about instance health, instance activity (launched, stopped), and the underlying hypervisor system may be sent back to the parent AWS Region. This information enables AWS to provide alerting on instance health and capacity, and apply patches and updates to the Outpost. Your team does not need to implement your own tooling to manage these elements, or to actively push security updates and patches for your Outpost. For S3 on Outposts, certain data management and telemetry data, such as bucket names and metrics, may be stored in the AWS Region for reporting and management. When disconnected, this information cannot be sent back to the parent Region.

Q. How does AWS support adding capacity to existing Outposts?

There are two mechanisms to increase your compute and storage capacity of your AWS Outposts rack. First, you can increase capacity by adding additional Outposts racks from the Outposts rack catalog. Second, if your existing Outposts racks have available power and positions within the rack, you can increase from a “small” to a “medium” or “large” configuration, or from a” medium” to a “large” configuration. You will be able to add compute and storage capacity a maximum of twice within a rack that supports 10KVA – 15KVA power consumption. Note: The 1U and 2U Outposts servers cannot be installed in the 42U Outposts form factor.

AWS Outposts servers FAQs

General

Q: What are AWS Outposts 1U and 2U servers?

AWS Outposts 1U and 2U rack-mountable servers provide local compute and networking services to edge locations that have limited space or smaller capacity requirements. Outposts servers are delivered directly to you, and installed by either your onsite personnel install or a 3rd-party vendor. Once connected to your network, AWS will remotely provision compute and storage resources. You can use the same AWS APIs, control plane, and tools that you're familiar with to manage servers across hundreds or thousands of on-premises locations.

The 1U high server is suitable for 19″ wide cabinets, is 24″ deep, and uses an AWS Graviton2 processor to provide 64 vCPUs, 128 GiB memory, and 4 TB of local NVMe storage. The 2U high server is suitable for standard 19″ wide, 30″ deep cabinets, and uses a 3rd generation Intel Xeon Scalable processor to provide up to 128 vCPUs, 256 GiB memory, and 8 TB of local NVMe.
Q: How do I deploy and manage applications on AWS Outposts servers?

You can deploy and manage applications on servers using the same tools for workloads running on Outposts racks or in the AWS Region. Launch resources by targeting the unique identifier (Outpost ID) that is associated with the virtualized capacity on a server. Outposts servers support local Amazon Elastic Compute Cloud (EC2) for compute, Amazon Elastic Container Service (ECS) for containers, and Amazon Virtual Private Cloud (VPC) for networking.

For fleet management, Outposts servers generate the same capacity and network monitoring metrics as Outposts racks. As you scale across sites, you can group devices and target deployments by using resource tags to associate devices within your fleet as you would for resources running in the AWS Region. You can also use deployment management tools such as AWS CloudFormation, AWS CodeDeploy, and Terraform.
Q: In which countries and territories are Outposts servers available?

Outposts can be shipped to and installed in the following countries and territories.

- NA - US, Canada
- EMEA - All EU countries, United Kingdom (UK), Switzerland, Norway, Bahrain, Gibraltar, and Morocco
- APAC - Australia, New Zealand, Japan, Hong Kong Special Administrative Region, Macau, Singapore, Indonesia, Malaysia, the Philippines, and Brunei
- CA - Puerto Rico

Support for more countries and territories is coming soon.
Q: What type of storage is available on Outposts servers?

Outposts servers support EC2 instance storage, which is block-level storage located on local SSD NVMe disks. The size of instance storage varies by instance type. Instance storage is tied to an instance's lifecycle. It does disappear if an EC2 instance is terminated. To learn more about implementing termination protection to prevent this, see Terminate your instance.
Q: Are Outposts servers suited to run near employees or customers?

The noise from an Outposts server will range from 66 to 78 dB, which is louder than common consumer-grade equipment. Outposts enterprise-grade servers will work well in a data-center environment, or protected IT spaces like an IT room, closet, or telco site.
Q: How do installations and repairs work on AWS Outposts servers?

After you receive your AWS Outposts servers, you can have either onsite personnel or a third-party IT provider install them. For Outposts servers that require maintenance, AWS will send you a replacement device.

AWS services

Q: What services are available on Outposts servers?

AWS Outposts servers have two form factors: the 1U server comes with Amazon EC2 C6gd compute optimized instances that are powered by Arm-based AWS Graviton2 processors, and the 2U server comes with Amazon EC2 C6id compute optimized instances that are powered by 3rd generation Intel Xeon Scalable processors for a wide variety of workloads. You can launch AWS services including Amazon EC2 for compute, Amazon VPC for networking, and Amazon ECS for containers. Amazon Elastic Kubernetes Service (EKS) for containers is coming soon. You can also manage Outposts servers applications and capacity using in-Region governance services like AWS Organizations and Resource Access Manager, CloudFormation and CodeDeploy, CloudWatch and CloudTrail. The pricing information for Outposts servers includes delivery, infrastructure service maintenance, and software patches and upgrades.

Ordering and installation

Q: What happens after I order my Outposts servers in the console or by contacting sales?

AWS will deliver Outposts servers directly to you, and you can either have your onsite personnel install them or an AWS preferred third-party contractor. After the Outposts servers are connected to your network, AWS will remotely provision compute and storage resources so you can start launching applications.

Security and compliance

Q: Who is responsible for the physical security of the Outposts at my datacenter?

AWS provides services that allow data to be encrypted at rest and in-transit and other granular security controls and auditing mechanisms. In addition, customer data is wrapped to a physical Nitro Security Key. Destroying the device is equivalent to destroying the data. As part of the shared responsibility model, customers are responsible for attesting to physical security and access controls around the Outpost, as well as environmental requirements for facility, networking, and power. Prior to returning the Outpost hardware, the Nitro Security Key will be removed to ensure customer content is crypto shredded.

Q: What is AWS Outposts FedRAMP authorization status?

FedRAMP has granted authorization to AWS Outposts, excluding the hardware components of the service. The physical and environmental controls for the AWS Outposts protection while at the customers' sites are the customers' responsibilities. Customers and agencies can make a risk-based decision to use AWS Outposts or conduct a Type Accreditation to review the hardware components of AWS Outposts to process FedRAMP workloads. Assessment results previously conducted by the AWS Third Party Assessment Organization can be accessed via CapLinked.

Support and maintenance

Q: How does AWS maintain Outposts servers infrastructure?

When your Outposts server is installed and is visible in the AWS Management Console, AWS will monitor it as part of the public Region and will automatically execute software upgrades and patches.

If there is a need to replace physical equipment, AWS will notify you and ship an appropriate module to your site. There is no additional charge for this, as hardware is included as part of the Outposts service price.

AWS Snow

Move petabytes of data to and from AWS, or process data at the edge.

AWS Snowcone

AWS Snowball Edge

AWS Snowball

AWS Snowmobile

AWS Snow Family

AWS Snow Family key features

Each feature listed below are standard features across each device type.

On-board computing

Snow Family devices have computing resources to collect and process data at the edge. Devices run specific Amazon EC2 instances with processing and storage available to support your applications and AWS IoT Greengrass functions.

End-to-end tracking

Each device uses an E-Ink shipping label for easy tracking and automatic label updates for return shipping using Amazon Simple Notification Service (SNS), text messages, and via the AWS Console.

Simple management and monitoring

AWS OpsHub is a complimentary graphical user interface (GUI) available to makes it easy to setup and manage Snow devices. Rapidly deploy edge computing workloads and migrate data to the cloud. Download AWS OpsHub here.

Encryption

All data moved to AWS Snow Family devices is automatically encrypted with 256-bit encryption keys that are managed by the AWS Key Management Service (KMS). Encryption keys are never stored on the device so your data stays secure during transit.

Secure erasure

Once the data migration job is complete and verified, AWS performs a software erasure of the device that follows the National Institute of Standards and Technology (NIST) guidelines for media sanitization.

NFS endpoint

Applications can work with Snow Family devices as an NFS mount point. NFS v3 and v4.1 are supported so you can easily use Snow devices with your existing on-premises servers and file-based applications.

Anti-tamper & Tamper-evident

AWS Snow devices feature a Trusted Platform Module (TPM) that provides a hardware root of trust. Each device is inspected after each use to ensure the integrity of the device and helps preserve the confidentiality of your data.

AWS Snow Family service models

AWS Snowcone

AWS Snowcone is the most compact and portable device. Weighing in at 4.5 pounds (2.1 kg) and available with SSD or HDD options, Snowcone is ruggedized, secure, and purpose-built for use outside of a traditional data center.

AWS Snowball

AWS Snowball is available as a Compute Optimized device or a Storage Optimized device. Explore the options best suited for your needs. All devices are suited for extreme conditions, tamper proof, and highly secure.

AWS Snowmobile

AWS Snowmobile is an Exabyte-scale data migration device used to move extremely large amounts of data to AWS. Migrate up to 100PB in a 45-foot long ruggedized shipping container, pulled by a semi-trailer truck.

Feature comparison matrix

	AWS SNOWCONE	AWS SNOWBALL EDGE STORAGE OPTIMIZED	AWS SNOWBALL EDGE COMPUTE OPTIMIZED	AWS SNOWMOBILE
Usable HDD Storage	8 TB	80 TB	N/A	100 PB
Usable SSD Storage	14 TB	1 TB	28 TB	No
Usable vCPUs	4 vCPUs	40 vCPUs	104 vCPUs	N/A
Usable Memory	4 GB	80 GB	416 GB	N/A
Device Size	9in x 6in x 3in 227 mm x 148.6 mm x 82.65 mm	548 mm x 320 mm x 501 mm	548 mm x 320 mm x 501 mm	45 ft. shipping container
Device Weight	4.5 lbs. (2.1 kg)	49.7 lbs. (22.3 kg)	49.7 lbs. (22.3 kg)	N/A
Storage Clustering	No	Yes, 5-10 nodes	Yes, 5-10 nodes	N/A
256-bit Encryption	Yes	Yes	Yes	Yes
HIPAA Compliant	No	Yes, eligible	Yes, eligible	Yes, eligible

What Is AWS Snowcone?

AWS Snowcone is a portable, rugged, and secure device for edge computing and data transfer. You can use a Snowcone device to collect, process, and move data to the AWS Cloud, either offline by shipping the device to AWS, or online by using AWS DataSync.

It can be challenging to run applications in austere (non–data center) edge environments, or where there is a lack of consistent network connectivity. These

locations often lack the space, power, and cooling needed for data center IT equipment.

Snowcone is available in two flavors:

- **Snowcone** – Snowcone has two vCPUs, 4 GB of memory, and 8 TB of hard disk drive (HDD) based storage.
- **Snowcone SSD** – Snowcone SSD has two vCPUs, 4 GB of memory, and 14 TB of solid state drive (SSD) based storage.

With two CPUs and terabytes of storage, a Snowcone device can run edge computing workloads that use Amazon Elastic Compute Cloud (Amazon EC2) instances and store data securely.

Snowcone devices are small (8.94” x 5.85” x 3.25” / 227 mm x 148.6 mm x 82.65 mm), so they can be placed next to machinery in a factory to collect, format, and transport data back to AWS for storage and analysis. A Snowcone device weighs about 4.5 lbs. (2 kg), so you can carry one in a backpack, use it with battery-based operation, and use the Wi-Fi interface to gather sensor data.

Snowcone devices offer a file interface with Network File System (NFS) support. Snowcone devices support data transfer from on-premises Windows, Linux, and macOS servers and file-based applications through the NFS interface.

Like AWS Snowball, AWS Snowcone has multiple layers of security encryption capabilities. You can use either of these services to collect, process, and transfer data to AWS, and run edge computing workloads that use Amazon EC2 instances. Snowcone is designed for data migration needs up to dozens of terabytes. It can be used in space-constrained environments where Snowball Edge devices don't fit.

How it works

AWS Snowcone is a small, rugged, and secure device offering edge computing, data storage, and data transfer on-the-go, in austere environment with little or no connectivity.

Use cases

Accelerate fleet data collection and analysis
Easily collect terabytes of data daily from large vehicle fleets in a small form factor with ruggedized design, power options, and enhanced security.

Gather IoT data in extreme conditions
Deploy an edge computing and storage solution designed for environments with limited space, bandwidth, and harsh environmental conditions such as factories, mines, and oil fields.

Improve patient outcomes
Deliver superior care to patients in transit or in the field—and transmit crucial data in real time with built-in Wi-Fi and AWS DataSync.

Expedite content distribution
Collect and process content — including high-resolution images—and improve team performance in fast-paced, space-constrained environments.

What Is AWS Snowball Edge?

AWS Snowball Edge is a type of Snowball device with on-board storage and compute power for select AWS capabilities. Snowball Edge can do local processing and edge-computing workloads in addition to transferring data between your local environment and the AWS Cloud.

Each Snowball Edge device can transport data at speeds faster than the internet. This transport is done by shipping the data in the appliances through a regional carrier. The appliances are rugged, complete with E Ink shipping labels.

Snowball Edge devices have three options for device configurations—*Storage Optimized*, *Compute Optimized*, and *Compute Optimized with GPU*. When this guide refers to Snowball Edge devices, it's referring to all options of the device. When specific information applies only to one or more optional configurations of devices (such as how the Snowball Edge with GPU has an on-board GPU), it is called out specifically.

AWS Snowball Edge Features

Snowball Edge devices have the following features:

- Large amounts of storage capacity or compute functionality for devices. This depends on the options you choose when you create your job.
- Network adapters with transfer speeds of up to 100 Gbit/second.
- Encryption is enforced, protecting your data at rest and in physical transit.
- You can import or export data between your local environments and Amazon S3, and physically transport the data with one or more devices without using the internet.
- Snowball Edge devices are their own rugged box. The built-in E Ink display changes to show your shipping label when the device is ready to ship.
- Snowball Edge devices come with an on-board LCD display that can be used to manage network connections and get service status information.
- You can cluster Snowball Edge devices for local storage and compute jobs to achieve data durability across 5–10 devices and locally grow or shrink storage on demand.
- You can use the file interface to read and write data to an AWS Snowball Edge device through a file share or Network File System (NFS) mount point.
- You can write Python-language Lambda functions and associate them with Amazon S3 buckets when you create an AWS Snowball Edge device job. Each function triggers when a local Amazon S3 `PUT` object action is run on the associated bucket on the device.
- Snowball Edge devices have Amazon S3 and Amazon EC2 compatible endpoints available, enabling programmatic use cases.
- Snowball Edge devices support the new `sbe1`, `sbe-c`, and `sbe-g` instance types, which you can use to run compute instances on the device using Amazon Machine Images (AMIs).

How AWS Snowball Edge Works

AWS Snowball Edge devices are owned by AWS, and they reside at your on-premises location while they're in use.

There are three job types you can use with an AWS Snowball Edge device. Although the job types differ in their use cases, every job type has the same workflow for how you order, receive, and return devices. Regardless of the job type, every job follows a data erasure of the National Institute of Standards and Technology (NIST) 800-88 standard after the job completes.

The shared workflow

1. **Create the job** – Each job is created in the AWS Snow Family Management Console or programmatically through the job management API. The status for a job can be tracked in the console or through the API.
2. **A device is prepared for your job** – We prepare an AWS Snowball Edge device for your job, and the status of your job is now **Preparing Snowball**.
3. **A device is shipped to you by your region's carrier** – The carrier takes over from here, and the status of your job is now **In transit to you**. You can find your tracking number and a link to the tracking website on the console or with the job management API. For information about who your region's carrier is, see Shipping Considerations for AWS Snowball.
4. **Receive the device** – A few days later, your region's carrier delivers the AWS Snowball Edge device to the address that you provided when you created the job, and the status of your job changes to **Delivered to you**. When it arrives, you'll notice that it didn't arrive in a box, because the device is its own shipping container.
5. **Get your credentials and download the Snowball Edge client** – Get ready to start transferring data by getting your credentials, your job manifest, and the manifest's unlock code, and then downloading the Snowball Edge client.
    - The Snowball Edge client is the tool that you use to manage the flow of data from the device to your on-premises data destination.

      You can download and install the Snowball Edge client from the AWS Snowball resources page.

      You must download the Snowball Edge client from the AWS Snowball Edge Resources page and install on a powerful workstation that you own.
    - The manifest is used to authenticate your access to the device, and it is encrypted so that only the unlock code can decrypt it. You can get

the manifest from the console or with the job management API when the device is on-premises at your location.
- The unlock code is a 29-character code used to decrypt the manifest. You can get the unlock code from the console or with the job management API. We recommend that you keep the unlock code saved somewhere separate from the manifest to prevent unauthorized access to the device while it's at your facility.

6. **Position the hardware** – Move the device into your data center and open it following the instructions on the case. Connect the device to power and your local network.
7. **Power on the device** – Next, power on the device by pressing the power button above the LCD display. Wait a few minutes, and the **Ready** screen appears.
8. **Get the IP address for the device** – The LCD display has a **CONNECTION** tab on it. Tap this tab and get the IP address for the AWS Snowball Edge device.
9. **Use the Snowball Edge client to unlock the device** – When you use the Snowball Edge client to unlock the AWS Snowball Edge device, enter the IP address of the device, the path to your manifest, and the unlock code. The Snowball Edge client decrypts the manifest and uses it to authenticate your access to the device.
10. **Use the device** – The device is up and running. You can use it to transfer data or for local compute and storage. You can read and write data with the Amazon S3 interface or the Network File System (NFS) mount point.
11. **Prepare the device for its return trip** – After you're done with the device in your on-premises location and the file interface status is **Complete**, press the power button above the LCD display. It takes about 20 seconds or so for the device to power off. Unplug the device and its power cables into the cable nook on top of the device, and shut all three of the device's doors. The device is now ready to be returned.
12. **Your region's carrier returns the device to AWS** – When the carrier has the AWS Snowball Edge device, the status for the job becomes **In transit to AWS**.

    Note

    There are additional steps for export and cluster jobs.

AWS Snowball

Accelerate moving offline data or remote storage to the cloud.

With AWS Snowball (Snowball), customers can transfer hundreds of terabytes or petabytes of data between their on-premises data centers and Amazon Simple Storage Service (Amazon S3). PHI stored in AWS Snowball must be encrypted at-rest consistent with the Guidance. When creating an import job, customers must specify the ARN for the AWS KMS key to be used to protect data within the Snowball. In addition, during the creation of the import job, customers should choose a destination S3 bucket that meets the encryption standards set by the Guidance.

While Snowball does not currently support server-side encryption with AWS KMS-managed keys (SSE-KMS) or server-side encryption with customer provided keys (SSE-C), Snowball does support server-side encryption with Amazon S3-managed encryption keys (SSE-S3).

Alternatively, customers can use the encryption methodology of their choice to encrypt PHI before storing the data in AWS Snowball.

Currently, customers may use the standard AWS Snowball appliance or AWS Snowmobile as part of our BAA.

How it works

In the AWS Snow Family console, select your preferred device, either Snowball Edge Compute Optimized or Snowball Edge Storage Optimized. Create a job with an Amazon S3 bucket, select Amazon Simple Notification Service (Amazon SNS) for tracking, and configure options like Amazon EC2 AMIs and a GPU. AWS prepares and ships the device to you, and you receive it in approximately 4-6 days. Once the device arrives, power it up and use AWS OpsHub to unlock it. Connect to your LAN. Use AWS OpsHub to manage the device, transfer data, or launch EC2 instances. When done, shut down and return the device to AWS. The shipping label automatically appears on the E Ink screen. When the device arrives at the AWS Region, any data stored in your on-board bucket(s) is moved to your S3 bucket and verified in about the same time it took you to load the device. All data is then securely erased from the device, and it is sanitized of any customer information.

AWS Snowball
Migrate petabyte-scale data to AWS with Snowball. For jobs that require multiple devices, track the stage of your device with Snow's Large Data Migration Manager.

AWS Snowball with Tape Gateway

Migrate petabyte-scale data stored on physical tapes to AWS using AWS Snowball

Use cases

Migrate data at petabyte-scale

Move databases,backups, archives, healthcare records, analytics datasets, IoT sensor data and media content to the cloud - especially when network conditions are limited.

Process and analyze data locally

Run Amazon Machine Images (AMIs) within Amazon EC2 and deploy AWS Lambda code on Snowball Edge devices with machine learning (ML) or other applications.

Optimize manufacturing data

Collect and analyze on-site factory data to refine processes and improve safety, efficiency, and productivity.

AWS Snowmobile

AWS Snowmobile is operated by AWS as a managed service. As such, AWS will contact the customer to determine requirements for deployment and arrange for network connectivity as well as provide assistance moving data. Data stored on Snowmobile is encrypted using the same guidance provided for AWS Snowball.

How it works

AWS Snowmobile moves extremely large amounts of data to AWS. Transfer up to 100 PB per Snowmobile, a 45-foot-long ruggedized shipping container pulled by a semi-trailer truck.

Use cases

Migrate massive amounts of data
Quickly move huge volumes of data to the cloud, including video libraries, image repositories, or even a complete data center.

Customize data transfer operations to your location
Every physical site is different. AWS tailors delivery and migration requirements to meet the needs of your site.

Meet security requirements for data migration
Keep your physically transferred data secure with 24/7 video surveillance, tamper-resistant hardware, GPS tracking, data encryption, and optional security personnel.

AWS Snow Family FAQs

General

Q: What is the AWS Snow Family?

The AWS Snow Family is a collection of physical devices that help migrate large amounts of data into and out of the cloud without depending on networks. This helps you apply the wide variety of AWS services for analytics, file systems, and archives to your data. AWS Snowcone is the smallest member of the AWS Snow Family of edge computing and data transfer devices. Snowcone is portable, rugged, and secure. You can use Snowcone to collect, process, and move data to AWS, either offline by shipping the device, or online with AWS DataSync. AWS Snowball is a suitcase-sized data migration and edge computing device that comes in two device options: Compute Optimized and Storage Optimized. Snowball Edge Storage Optimized devices provide 40 vCPUs of compute capacity coupled with 80 terabytes of usable block or Amazon S3-compatible object storage. It is well-suited for local storage and large-scale data transfer. Snowball Edge Compute Optimized devices provide 52 vCPUs, 42 terabytes of usable block or object storage, and an optional GPU for use cases such as advanced machine learning and full motion video analysis in disconnected environments. AWS Snowmobile is a shipping container moved with a tractor-trailer. These services can assist with data migration, disaster recovery, data center shutdown, and remote data collection projects.

Q: What are some example use cases?

You can use AWS Snow Family services for data transfer and occasional pre-processing on location. Some large data transfer examples include cloud migration, disaster recovery, data center relocation, and/or remote data collection projects. These projects typically require you to migrate large amounts of data in the shortest, and most cost-effective, amount of time.

Some example use cases for Snowcone and Snowball Edge computing capabilties include IoT sensor stream capture, on-the-fly media transcoding, image compression, aggregating metrics, and control signaling and alarming.

Q: Why do I need Snow Family services?

It can take a long time to transfer large amounts of data over the wire, and some locations don't have any connectivity at all. The Snow Family helps expedite data transfers in a more secure and cost-effective way. Each service has a pre-set capacity level to make it easy to choose the right option for you. Snowcone and Snowball Edge further help bring computing applications closer to the data source to enhance analysis and deliver real-time results.

Q: Which Snow Family service is best for me?

The Snow Family comes in multiple capacities and form factors to fit most data migration and/or remote data collection projects. First, determine the amount of storage capacity you need. Then, evaluate your need for onboard computing capabilities.

Q: How quickly can I migrate data?

The Snowball services can typically transfer up to 100 TBs in about a week. Snowmobile can transfer data at a rate of up to 1TB/s, which means 100 PBs can be loaded in just a few weeks. In comparison, a dedicated T3 line at 50Mb/s takes years.

Q: How do I choose between Snow Family and other AWS data migration services?

AWS Storage Gateway and AWS Direct Connect services are good choices when network bandwidth limitations do not exist. For the most efficient means of data transfer, whether connected to a network or not, the Snow Family of services provides good choices as well. AWS offers a variety of tools to help you move data via networks, roads, and technology partners. See the pages below to find out which service is best for you.

Q: What AWS Regions are supported?

Snow Family services are available for use in specific AWS Regions.

Q: What safety and compliance guidelines should I be aware of?

Download the AWS Snow Family Safety and Compliance Guideline for best practices and compliance statements based on your region.

AWS Snowcone FAQ

General

Q: What is AWS Snowcone?

Snowcone is small rugged, edge compute and data storage product. You can use Snowcone to collect, process, and transfer data to AWS, either offline by shipping the device, or online with AWS DataSync. Running applications in austere (non-data center) edge environments or where there is lack of consistent network connectivity or low bandwidth can be challenging because these locations often lack the space, power, and cooling needed for data center IT equipment. With 2 vCPUs, 4 GB of memory, and 8 TB of usable storage (14 TB for Snowcone SSD), all Snowcone devices can run edge computing workloads that use Amazon EC2 instances, and store data securely. Snowcone's small size (8.94 inches long x 5.85 inches wide x 3.25 inches tall / 227 mm x 148.6 mm x 82.65 mm) enables you to set it next to machinery in a factory to collect, format, and transport data back to AWS for storage and analysis. Snowcone weighs 4.5 lbs. (2.1 kg), so you can carry a device in a backpack, use it with battery-based operation, and use the the wired network interface to collect data. Snowcone features a file interface with NFS support. Snowcone devices supports data transfer from on-premises Windows, Linux, and MacOS systems as well as file-based applications via the NFS interface.

Q: How does AWS Snowcone work?

You start by ordering one or more AWS Snowcone devices in the AWS Snow Family console based on how much data you need to transfer along with the compute performance required. The buckets, data, and Amazon EC2 AMIs you select are automatically configured, encrypted, and pre-installed on your devices. You will receive the Snowcone device in approximately 4-6 days. To receive multiple AWS Snowcone devices, you need to setup a job order in the AWS Snow Family console for each Snowcone device.

Once a device arrives, you connect it to your on-premises network and set the IP address either manually or automatically with DHCP. You need to download and install AWS OpsHub, a GUI-

based application for managing the Snowcone device, on any Windows, Linux, or MacOS client machine, such as a laptop. Then open AWS OpsHub and unlock the device. You will then be presented with a dashboard showing your device and its system metrics. You can then launch instances to deploy your edge applications or migrate your data to the device with just a few clicks in AWS OpsHub. You can transfer data online from and to your device using AWS DataSync.

When your compute or data transfer job is complete and the device is ready to be returned, the E Ink shipping label on the device will automatically update the return address, ensuring that the Snowcone device is delivered to the correct AWS facility. Once the device ships, you can receive tracking status via messages sent by Amazon Simple Notification Service (Amazon SNS), generated texts and emails, or directly from the Console.

Q: Who should use AWS Snowcone?

You can use Snowcone for: 1) edge computing applications, to collect data, process the data to gain immediate insight, and then transfer the data online to AWS; 2) to transfer data that is continuously generated by sensors or machines online to AWS in a factory or at other edge locations; 3) to distribute media, scientific, or other content from AWS Storage services to your partners and customers; 4) to aggregate content by transferring media, scientific, or other content from your edge locations to AWS; and, 5) one-time data migration scenarios where your data is ready to be transferred, where AWS Snowcone offers a simple, quick, and low-cost way to transfer up to 8 TB of data into AWS by shipping the device back to AWS.

For mobile deployments, AWS Snowcone can run on specified battery power. For a light workload at 25% CPU usage, the device can run on a 65W battery for up to approximately 6 hours. You can use the wired network interface on Snowcone to collect data. AWS Snowcone is low power, portable, lightweight, and vibration resistant, allowing it to be used in a wide variety of remote and austere locations.

Q: When can I delete the data on my disk(s) after I've copied the data onto Snowcone and shipped the Snowcone back to AWS?

Wait to confirm that the Snowcone has been received by AWS and your data has successfully been transferred into appropriate S3 buckets prior to you deleting any data on your disk(s). While AWS verifies the integrity of files copied to Snowcone during the S3 transfer, it is your responsibility to verify the integrity of data before deleting it from your disk(s). AWS is not liable for any lost or corrupted data during copy or transit.

Specifications

Q: What are the dimensions of AWS Snowcone?

AWS Snowcone and Snowcone SSD weighs 4.5 lbs. (2.1 kg). It is 8.94 inches long, 5.85 inches wide, and 3.25 inches tall (227 mm x 148.6 mm x 82.65 mm).

Q: What are the hardware specifications of AWS Snowcone?

AWS Snowcone provides 8 TB of available storage while Snowcone SSD provides 14TB. Both run specific Amazon EC2 instances with 2 available CPUs and 4 GB of available memory to support your applications and AWS IoT Greengrass functions.

Q: How do I power the Snowcone device?

Snowcone does not include a power supply, as it ships with the smallest possible form factor. You have the option to run Snowcone via plug-in power source or battery. Here are the details to guide you when ordering a power supply:

USB-C power adapter

Use a USB-C power adapter to power the Snowcone device for plugged-in power or for stationary (non-mobile) operating environments. The list of USB-C power adapters tested for operation with Snowcone is provided below. To power your Snowcone device, you can purchase either one of the USB-C power adapters listed below or any USB-C power adapter that is rated for 45W+ and your operating environment temperature.

- Apple 61W USB-C Power Adapter
- Lenovo USB-C 65W Standard AC Adapter
- Dell 130-Watt Type-C AC power adapter

USB-C battery

Use a USB-C battery to power the Snowcone device in mobile or portable operating environments. The list of USB-C batteries tested for use with Snowcone is provided below. To power your Snowcone device, you can purchase either one of the USB-C batteries listed below or any USB-C battery that is rated for 45W+.

- Dell Power Bank Plus - USB C, 65Wh
- Omni 20c+ 100W USB-C

Q: What networking does AWS Snowcone support?

For wired networking, AWS Snowcone and Snowcone SSD provides 2 ports that auto-negotiate for 1 Gb or 10 Gb Ethernet networks.

Q: What are the ruggedization specifications of AWS Snowcone?

AWS Snowcone and Snowcone SSD is designed to meet stringent standards for ruggedization, including ISTA-3A, ASTM D4169, and MIL-STD-810G for free-fall shock, operational vibration, and more. It is designed to tolerate falls up to up to 3.8 feet (1.15 meters). It also meets the IP65 International Protection Marking IEC standard, meaning it is both dust-tight – allowing no dust inside the enclosure when sealed – and water resistant, including protection from water jets on all sides. The device has a wide operating temperature range from freezing (0°C/32°F) to desert-like conditions (38°C/100°F). When in storage or being shipped, Snowcone withstands even harsher temperatures (-32°C/-25.6°F to 63°C/145.4°F).

Performance

Q: What is Snowcone's data copy performance?

AWS Snowcone provides 8 TB of usable storage capacity via the NFS protocol while Snowcone SSD provides 14TB of usable storage. When client workstations are copying data to Snowcone, the following best practices are recommended.

Snowcone performs best when large files are transferred. Snowcone is not optimized for transferring many smaller files with a size of 5 MB or below. This limitation is due to the overhead involved in processing the metadata and encryption. If you intend to copy many files smaller than 5 MB, you should batch the small files using tar, zip, or similar utility before

copying to a Snowcone device. Snowcone performs best when it is used for write workloads (copying data into Snowcone) or read workloads (copying data from Snowcone), and not both. Performance cannot be guaranteed in a mixed read-write scenario.

Using a Linux or macOS workstation is the preferred method to copy data to a Snowcone. This is due to the protocol optimizations on these platforms. Customers with other platforms will likely experience lower performance. You will experience improvement through future service updates. If you want to use Snowcone for mixed read-write workload use cases, the best option is to spin up a Linux based NFS server on a Snowcone-based EC2 instance. You can have up to a 128 GB root file system instance to keep the data and simplify deployment.

AWS Snowcone devices are not recommended for the following scenarios:

1. Snowcone is a secure data transfer device and not a Network Attached Storage (NAS) device. The NFS share is intended to be a mechanism to move files to or from AWS. Continuous read/write workloads do not perform well on the Snowcone NFS share.
2. Snowcone cannot be used as a primary storage for database, file server, or similar applications due to the performance characteristics.
3. Simultaneous file transfers, or multiple file copy streams have the effect of randomizing data and thus slowing the overall transfer speed. Copying files using a single client in a serial manner is recommended to minimize randomness.

Additional note: Snowcone will occasionally perform internal data cleanup tasks after significant I/O operations. During such moments, Snowcone disk performance may seem to be reduced. It is recommended to leave Snowcone with the power on after significant I/O operations are performed, so the cleanup task can finish and normal disk performance is restored.

AWS OpsHub for device management and monitoring

Q: How does AWS OpsHub help me with managing and monitoring my AWS Snowcone devices?

AWS OpsHub is an application for managing the AWS Snow Family devices, including AWS Snowcone and Snowcone SSD. The AWS OpsHub graphical user interface (GUI) makes it easy to setup and manage your Snowcone devices so you can quickly run your edge compute workloads and migrate data to the cloud. With just a few clicks, you can use AWS OpsHub to unlock and configure Snowcone, drag-and-drop data, launch applications, or monitor device metrics. Download AWS OpsHub here and install it on Windows or MacOS client machines, such as a laptop. There is no cost to use OpsHub. For more information, refer to the AWS OpsHub documentation.

Q: How do I use AWS OpsHub for AWS Snowcone?

Once you download AWS OpsHub and install it on a client machine, AWS OpsHub can connect to the Snowcone device on the same network. Then you open OpsHub and unlock the device. You will then be presented with a dashboard showing your device and its system metrics. You can then begin deploying your edge applications or migrating your data to the device with just a few clicks.

Q: How do I use AWS OpsHub to monitor AWS Snowcone?

AWS OpsHub makes monitoring AWS Snowcone easy by providing you with a central dashboard to view all your Snow Family devices and their system metrics.

Q: How do I use AWS OpsHub for edge applications?

AWS OpsHub enables you to run your edge applications in minutes by allowing you to launch instances, configure network settings, and provision storage volumes with a few clicks.

Q: How do I use AWS OpsHub to copy data to AWS Snowcone?

AWS OpsHub makes data transfers to AWS Snowcone simple by allowing you to drag-and-drop files or folders onto the device. It enables you to easily view what is stored on the device.
Q: When do I use AWS OpsHub versus the AWS Management Console?

You use AWS OpsHub at your edge location to manage and monitor your AWS Snowcone device and the AWS services that run on Snowcone. AWS OpsHub is an application that runs on a local client machine, such as a laptop that is connected to your local network. In contrast, you use the AWS Management Console to manage and operate the AWS services running in the cloud. The AWS Management Console is a web-based application that operates when you have a connection to the internet.

Edge Computing Capabilities

Q: When should I use the EC2 compatible instances on AWS Snowcone?

You should use the EC2-compatible instances when you have an application running on the edge that is managed and deployed as a virtual machine (an Amazon Machine Image, or AMI). AWS Snowcone supports the SNC1 instance type with three instances including snc1.micro (1 CPU and 1 GB RAM), snc1.small (1 CPU and 2 GB RAM), and snc1.medium (2 CPU and 4 GB RAM). The support for EC2-compatible instances on Snowcone enables you to build and test your application on EC2.

Q: How do I get started with local computing on AWS Snowcone?

You can enable and provision EC2 AMIs during AWS Snowcone job creation using either the AWS Console, AWS Snowball SDK, or AWS CLI.

Q: How do I manage the instances on AWS Snowcone?

Use AWS OpsHub to manage your instances on AWS Snowcone. Download and install AWS OpsHub. Once you have unlocked the device using OpsHub, navigate to the Amazon EC2 page. From there, click “Create instance”. This will create an Amazon EC2 instance based on the Amazon Machine Image that you had preloaded onto the device when you created the job. You can then connect to the instances and run your edge application. OpsHub also provides single-click buttons to start, stop, terminate, and reboot your EC2 instances.

Q: What operating systems can I run using this feature?

EC2 instances hosted on AWS Snowcone provides default support for a variety of free-to-use operating systems (OS) like Ubuntu and CentOS. They will appear as AMIs that can be loaded onto Snowcone without any modification. To run other OSes you will need to import them using Amazon EC2 VM Import/Export (VMIE). You are responsible for installing any required licenses.

Q: Does Snowcone support encrypted AMIs?

Currently, Snowcone does not support encrypted AMIs. Support for encrypted AMIs are planned with future updates to our system.

Q: How do I ensure that my AMIs are compatible to run on EC2-compatible instances on AWS Snowcone?

AMIs that run on the T2 instance types including t2.micro, t2.small, and t2.medium in AWS are compatible with the SNC1 instance types of snc1.micro, snc1.small, and snc1.medium available on AWS Snowcone. We recommend that you first test your applications in the comparable T2 instance type to ensure they can run on Snowcone. If VirtIO disk and network virtualization drivers are not already installed, you must install them before creating an AMI from your instance.

Q: How is software licensing handled with compute instances on AWS Snowcone?

You can opt to use the open-source CentOS or Ubuntu Linux AMIs the service provides, or you can bring your own OS image. Beyond the supported open-source AMIs noted above, you are responsible for licensing any software that you run on your instance. Specifically, for Windows operating systems, you can bring your existing license to the running instances on the device, by installing the licensed OS in your AMIs in EC2. Once your VM is uploaded using VM Import/Export and saved as an AMI, you can specify to have it installed on your device when you place your Snowcone order.

Q: How do I get started with the latest AWS IoT Greengrass on Snow devices?

AWS IoT Greengrass is an IoT edge runtime (open source starting with version 2.0) and a cloud service that helps you build, deploy and manage IoT applications on your devices. AWS Snow devices running AWS IoT Greengrass can operate as an IoT hub, data aggregation point, application monitor, or a lightweight analytics engine.

To get started with AWS IoT Greengrass on an AWS Snow device, follow the steps listed below:

1. When you are ready to place your job order on the AWS Snow Family console, you can opt-in to install the AWS IoT Greengrass AMI, which uses Amazon Linux 2 (AL2) AMI for the AWS Snow Family on the Snow device of your choice.
2. Once you receive the device, you can use AWS OpsHub for Snow Family to unlock the device with the credentials provided after the job is created.
3. After the device is powered on and unlocked, you can launch the AL2 AMI for AWS Snow Family on the applicable Snow device and remotely log in to it using your existing SSH keys or by creating new SSH keys.
4. Now you are ready to install the latest version AWS IoT Greengrass on the Snow device following the instructions here.
5. Once the installation is complete, you will be able to manage the AWS Snow Family device and deploy IoT workloads from the AWS IoT console.

Online Data Transfer between AWS Snowcone and AWS Storage services with AWS DataSync

Q: What is AWS DataSync?

AWS DataSync is an online data transfer service that simplifies, automates, and accelerates copying large amounts of data to and from AWS storage services over the internet or AWS Direct Connect. The DataSync agent comes pre-installed on your Snowcone device and can

transfer data between the device and Amazon S3 buckets, Amazon EFS file systems, or Amazon FSx for Windows File Server file systems. DataSync automatically handles moving files and objects, scheduling data transfers, monitoring the progress of transfers, encryption, verification of data transfers, and notifying customers of any issues.

Q: How do I transfer data online between AWS Snowcone and AWS Storage services?

To transfer data online to AWS, connect the AWS Snowcone device to the external network and use AWS OpsHub or the CLI to launch the DataSync agent AMI that comes pre-installed on the device. Download OpsHub here. Activate the AWS DataSync agent using the AWS Management Console or CLI, and set up your online data transfer task between AWS Snowcone's NFS store, and Amazon S3, Amazon EFS, or Amazon FSx for Windows File Server.

Q: When do I use online data transfer with AWS DataSync versus shipping the AWS Snowcone device back to AWS for offline data transfer?

You can use AWS DataSync for online transfers from and to your AWS Snowcone device, such as: 1) edge computing applications, to collect data, process the data to gain immediate insight, and then transfer the data online to AWS; 2) to transfer data that is generated by sensors or machines online to AWS in a factory or at other edge locations; 3) to distribute media, scientific, or other content online from AWS Storage services to your partners and customers; and, 4) to aggregate content by transferring media, scientific, or other content online from your edge locations to AWS.

For one-time edge compute or data transfer workflows or for Snowcone workflows in edge locations without a Wide Area Network (WAN) link or inadequate WAN bandwidth, you should ship the AWS Snowcone device back to AWS to complete the data transfer.

Offline Data Transfer between Snowcone and Amazon S3

Q: How do I import my data offline to Amazon S3 with AWS Snowcone?

You connect the AWS Snowcone device to your on-premises network and then use AWS OpsHub to quickly unlock the device. Download OpsHub here. You can copy data from on-premises storage devices to AWS Snowcone through the NFS file interface. When you have copied the data to Snowcone, the E Ink shipping label is designed to ensure the device is automatically sent to the correct AWS facility. You can track the Snowcone device by using Amazon SNS generated text messages or emails, along with using the Console.

Q: How much data can I transfer offline using AWS Snowcone?

You can transfer up to 8 TB with a single AWS Snowcone device and can transfer larger data sets with multiple devices, either in parallel, or sequentially. For example, you can transfer 24 TB of data with 3 Snowcone devices. For larger data transfer jobs, you can use the Snowball Edge Storage Optimized device. You can transfer up to 80 TB with a single Snowball Edge Storage Optimized device and can transfer larger data sets with multiple devices, either in parallel, or sequentially.

Networking

Q: Can I use more than one physical network interface on a Snow device?

Yes. You can create a direct network interface that can be associated with any of the physical network ports on your Snow device. You can also create multiple direct network interfaces and associate them with different physical network ports on your Snow device. Multiple direct network interfaces can be associated with a single EC2 instance as well.

Q: Does Snowcone support multicast streams, routing, or load balancing?

Yes. Direct network interfaces provide instances with layer 2 network access without any intermediary translation or filtering, enabling multicast, routing, or load balancing use cases. Direct network interfaces also support optional VLAN tags and allow you to optionally customize the MAC address of the network interface.

Regional Availability

Q: In what Regions is AWS Snowcone available?

AWS Snowcone is currently available in the US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), Canada (Central), South America (Sao Paulo), Europe (Ireland), Europe (Frankfurt), Europe (London), Europe (Paris), Asia Pacific (Sydney), Asia Pacific (Tokyo), Asia Pacific (Singapore), and Asia Pacific (Mumbai) Regions.

Q: Can an AWS Snowcone be shipped to an alternate AWS Region?

No. AWS Snowcone devices are designed to be requested and used with a single AWS Region. It may not be requested from one region and returned to another.

Security

Q: Does AWS Snowcone encrypt my data?

All data on AWS Snowcone device is always automatically encrypted using 256-bit keys that you manage by using the AWS Key Management Service (KMS). The encryption keys are never stored on the device. This helps ensure that your data stays secure during device transit.

Q: How does AWS Snowcone physically secure my data?

The AWS Snowcone device has a Trusted Platform Module (TPM) that provides a hardware root of trust. The TPM also provides interfaces to the trusted software stack during the measurements and verification of the boot environment integrity after the power is switched on, and before the AWS Snowcone device is ready to be used. AWS also uses additional tamper-indicating inspection processes after each AWS Snowcone device is received back to the AWS Region. This helps to ensure the integrity of the AWS Snowcone device, and with the Snowcone service's encryption features, it helps preserve the confidentiality of your data.

Q: Is there a way to easily track my AWS Snowcone device?

AWS Snowcone and Snowcone SSD use an innovative, E Ink shipping label designed to ensure the device is automatically sent to the correct AWS facility when you have completed your compute or data transfer job. You can track the Snowcone device by using Amazon SNS generated text messages or emails, and via the AWS Console.

Workflow Integration Tools

Q: Does the AWS Snowcone support API access?

Yes. The AWS Snow Family Job Management API provides programmatic access to the job creation and management features for Snowcone. It is a simple, standards-based REST web service interface, designed to work with any internet development environment.

Q: What can I do with the AWS Snow Family Job Management API?

The AWS Snowball Job Management API allows partners and customers to build custom integrations to manage the process of requesting Snowcone devices and communicating job status. The API provides a simple web service interface that you can use to create, list, update, and cancel jobs from anywhere on the web. Using this web service, developers can easily build applications that manage Snowcone jobs.

Q: Can I get a history of AWS Snowball API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. To receive a history of Snowball API calls made on your account for AWS Snowcone, you simply turn on AWS CloudTrail in the AWS Management Console. Further details on Snowball API logging by CloudTrail can be accessed here.

AWS Snowball FAQs

General

Q: What is AWS Snowball?

AWS Snowball is a service that provides secure, rugged devices, so you can bring AWS computing and storage capabilities to your edge environments, and transfer data into and out of AWS. Those rugged devices are commonly referred to as AWS Snowball or AWS Snowball Edge devices. Previously, AWS Snowball referred specifically to an early hardware version of these devices, however that model has been replaced by updated hardware. Now the AWS Snowball service operates with Snowball Edge devices, which include on-board computing capabilities as well as storage.

Q: What is AWS Snowball Edge?

AWS Snowball Edge is an edge computing and data transfer device provided by the AWS Snowball service. It has on-board storage and compute power that provides select AWS services for use in edge locations. Snowball Edge comes in two options, Storage Optimized and Compute Optimized, to support local data processing and collection in disconnected environments such as ships, windmills, and remote factories.

Q: What happened with the original 50 TB and 80 TB AWS Snowball devices?

The original Snowball devices were transitioned out of service and Snowball Edge Storage Optimized are now the primary devices used for data transfer.

Q: Can I still order the original Snowball 50 TB and 80 TB devices?

No. For data transfer needs now, please select the Snowball Edge Storage Optimized devices.

Q: How does Snowball Edge work?

You start by requesting one or more Snowball Edge Compute Optimized or Snowball Edge Storage Optimized devices in the AWS Management Console based on how much data you need to transfer and the compute needed for local processing. The buckets, data, Amazon EC2 AMIs, and Lambda functions you select are automatically configured, encrypted, and preinstalled on your devices before they are shipped to you. Once a device arrives, you connect it to your local network and set the IP address either manually or automatically with DHCP. Then use the Snowball Edge client software, job manifest, and unlock code to verify the integrity of the Snowball Edge device or cluster, and unlock it for use. The manifest and unlock code are uniquely generated and crypto-logically bound to your account and the Snowball Edge shipped

to you, and cannot be used with any other devices. Data copied to Snowball Edge is automatically encrypted and stored in the buckets you specify.

All logistics and shipping is done by Amazon, so when copying is complete and the device is ready to be returned, the E Ink shipping label will automatically update the return address, ensuring that the Snowball Edge device is delivered to the correct AWS facility. Once the device ships, you can receive tracking status via messages sent by Amazon Simple Notification Service (Amazon SNS), generated texts and emails, or directly from the console.

All of the management for your Snowball Edge resources can be performed in the AWS management console and these operations require system engineers.

Q: What is the difference between Snowball Edge and Snowball?

AWS Snowball now refers to the service overall, and Snowball Edge are the current types of devices that the service uses – sometimes referred to generically as AWS Snowball devices. Originally, early Snowball hardware designs were for data transport only. Snowball Edge has the additional capability to run computing locally, even when there is no network connection available.

Q: What is the difference between the Snowball Edge Storage Optimized and Snowball Edge Compute Optimized options?

Snowball Edge Storage Optimized is the optimal choice if you need to securely and quickly transfer dozens of terabytes to petabytes of data to AWS. It is also a good fit for running general purpose analysis such as IoT data aggregation and transformation. It provides up to 80 TB of usable HDD storage, 40 vCPUs, 1 TB of SATA SSD storage, and up to 40 Gb network connectivity to address large scale data transfer and pre-processing use cases. We recommend using Snowball Edge Compute Optimized for use cases that require access to powerful compute and high-speed storage for data processing before transferring it into AWS. It features 104 vCPUs, 28 TB of NVMe SSD, and up to 100 Gb networking to run applications such as high-resolution video processing, advanced IoT data analytics, and real-time optimization of machine learning models in environments with limited connectivity.

Q: Who should use Snowball Edge?

Consider Snowball Edge if you need to run computing in rugged, austere, mobile, or disconnected (or intermittently connected) environments. Also consider it for large-scale data transfers and migrations when bandwidth is not available for use of a high-speed online transfer service, such as AWS DataSync.

Snowball Edge Storage Optimized is the optimal data transfer choice if you need to securely and quickly transfer terabytes to petabytes of data to AWS. You can use Snowball Edge Storage Optimized if you have a large backlog of data to transfer or if you frequently collect data that needs to be transferred to AWS and your storage is in an area where high-bandwidth internet connections are not available or cost-prohibitive.

You can also use Snowball Edge to run edge computing workloads, such as performing local analysis of data on a Snowball Edge cluster and writing it to the S3-compatible endpoint. You can streamline it into existing workflows leveraging built-in capabilities such as the NFS file interface and migrate files to the device while maintaining file metadata.

Snowball Edge can operate in remote locations or harsh operating environments, such as factory floors, oil and gas rigs, mining sites, hospitals, and on moving vehicles. Snowball Edge is pre-configured and does not have to be connected to the internet, so processing and data collection can take place within isolated operating environments. Snowball Edge allows you to

run the same software at the edge and access select AWS capabilities as you would with full connectivity to AWS.

Q. Can I use Snowball Edge to migrate data from one AWS Region to another AWS Region?

No. Snowball Edge is intended to serve as a data transport solution for moving high volumes of data into and out of a designated AWS Region. For use cases that require data transfer between AWS Regions, we recommend using S3 Cross-Region Replication as an alternative.

Q: How much data can I transfer using Snowball Edge?

You can transfer virtually any amount of data with Snowball Edge, from a few terabytes to many petabytes. You can transfer up to approximately 80 TB with a single Snowball Edge Storage Optimized device and can transfer even larger data sets with multiple devices, either in parallel, or sequentially.

Q: How long does it take to transfer my data?

Data transfer speed is affected by a number of factors including local network speed, file size, and the speed at which data can be read from your local servers. The end-to-end time to transfer up to 80 TB of data into AWS with Snowball Edge is approximately one week, including the usual shipping and handling time in AWS data centers.

Q: How long can I have a Snowball Edge for a specific job?

For security purposes, jobs using an AWS Snowball Edge device must be completed within 360 days of being prepared. If you need to keep one or more devices for longer than 360 days, contact AWS Support. Otherwise, after 360 days, the device becomes locked, can no longer be accessed, and must be returned. If the AWS Snowball Edge device becomes locked during an import job, we can still transfer the existing data on the device into Amazon S3.

Q: What are the specifications of the Snowball Edge devices?

Please see the AWS Snowball Features page for feature details and the Snowball Edge documentation page for a complete list of hardware specs, including network connections, thermal and power requirements, decibel output, and dimensions.

Q: What network interfaces does Snowball Edge support?

Snowball Edge Storage Optimized for data transfer devices have two 10G RJ45 ports, one 10/25G SFP28 port, and one 40G/100G QSFP28 port.

Snowball Edge Storage Optimized for edge compute devices have one 10G RJ45 port, one 10/25G SFP28 port, and one 40G QSFP+ port.

The Snowball Edge Compute Optimized devices (including the GPU option) have two 10G RJ45 ports, one 10/25G SFP28 port, and one 40G/100G QSFP28 port.

Q: What is the Snowball Edge default shipping option? Can I choose expedited shipping?

As a default, Snowball Edge uses two-day shipping by UPS. You can choose expedited shipping if your jobs are time-sensitive.

Edge Computing Capabilities

Q: Does Snowball Edge support EC2 instances?

Yes. The Snowball Edge Storage Optimized option supports SBE1 instance.

The Snowball Edge Compute Optimized option features more powerful and larger instances, SBE-C for compute-intensive applications.

The Snowball Edge Compute Optimized device with an optional GPU, can use SBE-G instances to accelerate your application's performance.

The support for EC2-compatible instances on Snowball Edge devices enables you to build and test on EC2, then operate your AMI on a Snowball Edge to address workloads that sit in remote or disconnected locations.

Q: Does Snowball Edge support Lambda functions?

Yes, Lambda functions are hosted and can be executed on Snowball Edge in response to data storage events.

Q: How do Lambda functions work on Snowball Edge?

Lambda functions are hosted locally on Snowball Edge. As data is written to your appliance, Lambda functions can be triggered to act on that data. In the same way as they act in AWS, Lambda functions can call other services, update objects, or make other changes.

Q: How should I choose between Amazon EC2 compute instances and AWS Lambda functions for my compute needs?

AWS Lambda is a good choice for new applications that want to take advantage of the serverless computing model in AWS and want to run the same applications on the device. Amazon EC2 instances are a good choice when you have existing applications that you would like to run on the device for data pre-processing or when refactoring your existing applications to the serverless model isn't desirable.

Q: How can I use the GPU with AWS Snowball Edge's SBE instances?

The GPU option on AWS Snowball Edge Compute Optimized comes with SBE-G instances that can take advantage of the onboard GPU for accelerating the application performance. After receiving the device, select the option to use the SBE-G instance in order to use the on-board GPU with your application.

Q: When should I use the EC2 compatible instances on AWS Snowball Edge?

You should use the EC2 compatible instances when you have an application running on the edge that is managed and deployed as a virtual machine (an Amazon Machine Image, or AMI).

Q: Can multiple Snowball Edge devices be clustered together?

Yes, multiple Snowball Edge Storage Optimized or Compute Optimized devices can be clustered into a larger durable storage pool with a single S3-compatible endpoint. For example, if you have 6 Storage Optimized devices, they can be configured to be a single cluster that exposes a single S3 compatible endpoint with 400 TB of storage. Alternatively, they can be used individually without clustering, each hosting a separate S3 compatible endpoint with 80 TB of usable storage. A durable cluster cannot be created using a mix of Storage Optimized and Compute Optimized devices.

Q: When would I consider clustering Snowball Edge devices together?

With a Snowball Edge cluster, you increase local storage durability and scalability. Clustering Snowballs creates durable, scalable, S3 compatible local storage. Data can be shipped to AWS by swapping Snowballs in and out of the cluster seamlessly. Snowball Edge clusters allow you to scale your local storage capacity up or down depending on your requirements by adding or removing appliances, eliminating the need to buy expensive hardware.

Q: How do I get started with local computing on Snowball Edge?

You can enable and provision Amazon EC2 AMIs or Lambda functions during AWS Snowball Edge job creation using either the AWS Console, AWS Snowball SDK, or AWS CLI.

Q: Can I use existing Amazon EC2 APIs to start, stop, and manage instances on the device?

Yes. AWS Snowball Edge provides an Amazon EC2-compatible endpoint that can be used to start, stop, and manage your instances on AWS Snowball Edge. This endpoint is compatible with the AWS CLI and AWS SDK.

Q: What Amazon EC2 features does AWS Snowball Edge support?

The Amazon EC2 endpoint running on AWS Snowball Edge, provides a set of EC2 features that customers would find most useful for edge computing scenarios. This includes APIs to run, terminate, and describe your installed AMIs and running instances. Snowball Edge also supports block storage for EC2 images, which is managed using a set of the Amazon EBS API commands.

Q: Can I use an existing Amazon EBS volume with AWS Snowball Edge?

No. At this time, you cannot use an existing EBS volume with AWS Snowball Edge, however, Snowball Edge does offer block storage volumes, which are managed with an EBS-compatible API.

Q: What steps do I need to take to run Amazon EC2 instances on AWS Snowball Edge?

To run instances, provide the AMI IDs during job creation and the images come pre-installed when the device is shipped to you.

Q: Can I convert my images from other hypervisors to AMIs and vice versa?

Yes. You can import or export your KVM/VMware images to AMIs using the EC2 VM Import/Export service. Refer to the VM Import/Export documentation for more details.

This is necessary in order to run licensed software, including operating systems other than those which the AWS Snowball service provides.

Q: What operating systems can I run using this feature?

Amazon EC2 on Snowball Edge provides default support for a variety of free-to-use operating systems (OS) like Ubuntu and CentOS. They will appear as AMI's that can be loaded onto Snowball Edge without any modification. To run other OSes that require licenses on Snowball Edge EC2 instances, you must provide your own license, and then export the AMI using Amazon EC2 VM Import/Export (VMIE).

Q: What kind of workloads can I run on SBE1 and SBE-C instances?

SBE1 instances feature up to 40 vCPUs, ephemeral instance storage for root volumes, and 32GB of memory, and are designed to support edge applications, such as IoT sensor data

collection, image compression, data collection, and machine learning. SBE1 instances can also use Snowball Edge SATA SSD and HDD block storage for persistent volumes.

SBE-C instances feature up to 104 vCPUs, ephemeral instance storage for root volumes, and 418GB of memory for running a wide variety of compute-intensive applications, such as advanced machine learning, full motion video analysis, LAMP stacks, and EC2-hosted containers, in environments with little or no internet connectivity. SBE-C instances can also use Snowball Edge NVMe SSD and HDD block storage for persistent volumes.

Q: How do I ensure that my AMIs are compatible to run on EC2-compatible instances on AWS Snowball Edge?

AMIs that run on the C5 instance type in AWS are compatible with SBE1 instances available on AWS Snowball Edge Storage Optimized in the vast majority of cases. We recommend that you first test your applications in the C5 instance type to ensure they can be run on the Snowball Edge Storage Optimized device.

AMIs that run on the M5a instance type in AWS are compatible with SBE-C instances available on AWS Snowball Edge Compute Optimized in the vast majority of cases. For SBE-C instances running on the AWS Snowball Edge Compute Optimized device, we recommend you test your applications on the M5a instance type.

For SBE-G instance types for the Snowball Edge Compute Optimized with the GPU option, we recommend you first test your applications against the EC2 P3 instance types.

Q: Can I install more than one instance on a device?

Yes. You can run multiple instances on a device as long as the total resources used across all instances are within the limits for your Snowball Edge device.

Q: How do I use SBE1, SBE-C, and SBE-G instances on an AWS Snowball Edge cluster?

All the EC2 compatible instances can run on each node of an AWS Snowball Edge cluster. When you provision an AWS Snowball Edge cluster using the AWS Console, you can provide details for instances to run on each node of the cluster, for example, the AMI you want to run and the instance type and size you want to use. Nodes can use the same or different AMIs across each node in a cluster.

Q: How do I launch an instance manually?

Each AMI has an AMI ID associated with it. You can use run-instance command to start the instance by providing this ID. Running this command returns an instance-id value that can be used to manage this instance.

Q: How do I manage the instances on AWS Snowball Edge?

You can check the status of all the images that are installed on the device using the describe-images command. To see the active instances of instances running on the device, you can use the describe-instance-status command.

Q: How do I terminate an existing instance?

You can terminate a running instance using the terminate-instance command.

Q: How are my AMIs protected while in transit?

Snowball Edge encrypts all data, including AMIs, with 256-bit encryption. You manage your encryption keys by using the AWS Key Management Service (KMS). Your keys are never stored on the device and you need both the keys and an unlock code to use the device on-premises. In addition to using a tamper-evident enclosure, Snowball Edge uses industry-standard Trusted Platform Modules (TPM) designed to detect any unauthorized modifications to the hardware, firmware, or software. AWS visually and cryptographically inspects every device for any signs of tampering.

Q: Can I add AMIs to my Snowball Edge device after it has been deployed?

Yes. You can import virtual machine (VM) images as AMIs to your Snow device while it is onsite.

Q: How is software licensing handled with compute instances on AWS Snowball Edge?

You are responsible for licensing any software that you run on your instance. Specifically, for Windows operating systems, you can bring your existing license to the running instances on the device, by installing the licensed OS in your AMI in EC2, and then using VM Import/Export to load the AMI to your Snowball Edge device.

Q: How do I get started with the latest AWS IoT Greengrass on Snow devices?

AWS IoT Greengrass is an IoT edge runtime (open source starting with version 2.0) and a cloud service that helps you build, deploy and manage IoT applications on your devices. AWS Snow devices running AWS IoT Greengrass can operate as an IoT hub, data aggregation point, application monitor, or a lightweight analytics engine.

To get started with AWS IoT Greengrass on an AWS Snow device follow the steps listed below:

1. When you are ready to place your job order on the AWS Snow Family console, you can opt-in to install the AWS IoT Greengrass AMI, which uses Amazon Linux 2 (AL2) AMI for the AWS Snow Family on the Snow device of your choice.
2. Once you receive the device, you can use AWS OpsHub for Snow Family to unlock the device with the credentials provided after the job is created.
3. After the device is powered on and unlocked, you can launch the AL2 AMI for AWS Snow Family on the applicable Snow device and remotely log in to it using your existing SSH keys or by creating new SSH keys.
4. Now you are ready to install the latest version AWS IoT Greengrass on the Snow device following the instructions here.
5. Once the installation is complete, you will be able to manage the AWS Snow Family device and deploy IoT workloads from the AWS IoT console.

Block Storage for Amazon EC2 on Snowball Edge

Q: What is block storage on AWS Snowball Edge?

You can run block storage on both Snowball Edge Compute Optimized and Snowball Edge Storage Optimized devices. You attach block storage volumes to EC2 instances using a subset of Amazon EBS capabilities that enable you to configure and manage volumes for EC2 instances on Snowball Edge devices.

Q: Why should I use AWS Snowball Edge block storage?

Snowball Edge block storage enables you to have multiple persistent block storage volumes – in addition to your root volume – for your Amazon EC2 based applications on the device. This

can provide higher performance and more storage capacity for EC2 applications on Snowball Edge than you can achieve with only a root volume. You can now attach multiple volumes to your EC2 instances. Volumes that are attached to EC2 instances on Snowball Edge persist independently from the life of the instance, enabling you to deploy multiple applications on Snowball Edge, and to start and stop the applications as needed.

Q: What are the types of block storage volumes I can use, and how much capacity can each volume type use?

Snowball Edge block storage provides performance-optimized SSD volumes (sbp1), and capacity-optimized HDD volumes (sbg1), to meet IOPS and throughput requirements for a wide-variety of data processing and data collection applications. Block storage volumes have a maximum size of 10 TB per volume, and you can attach up to 10 volumes to any EC2 instance on Snowball Edge.

On Snowball Edge Compute Optimized, you can use up to 7 TB of NVMe SSD for sbp1 volumes, which are good for latency-sensitive applications, such as machine learning. On Snowball Edge Storage Optimized, you can use up to 1 TB of SATA SSD for sbp1 volumes, which are good for pre-processing data.

On both the Snowball Edge Storage Optimized and Compute Optimized devices, you can use capacity-optimized volumes, sbg1, to store data on HDDs. This volume type is appropriate for data collection and less IOPS-intensive applications. It has a maximum volume size of 10 TB, and you can attach multiple volumes to any instance. You can use up to 86 TB of HDD storage for sbg1 on Snowball Edge Storage Optimized. You can use up to 40 TB of HDD storage for sbg1 on Snowball Edge Compute Optimized.

Q: How do I get started with block storage on Snowball Edge?

By default, all Snowball Edge devices are now shipped with the block storage feature. Once you unlock the device you can use AWS CLI or SDK to create volumes and attach them to an Amazon EC2 instance. You can attach multiple volumes to each EC2 instance, however, a single volume can only be attached to a single instance at any time.

Q: How is Snowball Edge block storage different from Amazon EBS?

Snowball Edge block storage has different performance, availability, and durability characteristics than Amazon EBS volumes. Also, it provides only a subset of Amazon EBS capabilities. For example, snapshot functionality is not currently supported on Snowball Edge block storage.

Q: Which Amazon EBS APIs does SBE block storage support?

To interact with block storage on SBE, you can use create, delete, attach, detach, and describe volumes EBS APIs.

Q: Which Amazon Machine Images can I use on Snowball Edge to utilize block storage?

Any Amazon Machine Image (AMI) running on Snowball Edge can access up to 10 block storage volumes at once. Generic AMIs provided by AWS and custom AMIs can access any block storage volume. There are no special requirements to make the block storage volumes work. However, certain operating systems perform better with specific drivers.

Q: Can the volumes on one device be accessible to Amazon EC2 instances running on another device?

Volumes created on a single Snowball Edge are only accessible to the EC2 instances running on that device.

Q: How can we monitor storage capacity used by various volumes?

You can use describe-device command from the Snowball client to monitor how much block storage is been used on the device. When you create a volume, all of the storage capacity requested is allocated to it based on the available capacity.

Q: Can I transfer data stored on block storage to Amazon EBS volumes in the cloud?

Not directly, no. Data on block storage volumes on Snowball Edge is deleted when the device returns to AWS. If you wish to preserve data in block storage volumes, you must copy the data into the Amazon S3 compatible storage on Snowball Edge. This data will then be copied into your S3 bucket when the device returns to AWS.

Q: Can I operate object and block storage on the same device?

Yes, you can use the Amazon S3-compatible object storage, and the Snowball Edge block storage on the same device. The object storage and block storage used for sbg1 volumes share the same HDD capacity. The underlying storage features work together so that an increase in I/O demand for block or object storage does not impede the availability and performance of the other.

Q: Do I need to configure volumes or any storage resources when ordering my Snowball Edge from the AWS Console?

No, you add volumes to your Amazon EC2 instances after you have received the device.

Q: Do I need to allocate storage resources on the device between block and object storage?

No. You can dynamically add or remove volumes and objects based on your application needs.

Q: Are the volumes encrypted by default?

Snowball Edge is designed with security in mind for the most sensitive data. All data written into block volumes is encrypted by keys provided by you through AWS Key Management Service (KMS). All volumes are encrypted using the same keys selected during Snowball Edge job creation. The keys are not permanently stored on the device and are erased after loss of power.

Q: What are best practices to achieve optimum performance with Snowball Edge block storage?

Additional volumes attached using the block storage offer up to 10 times higher performance compared to the root volumes. We recommended you use relatively smaller root volumes, and create additional block storage volumes for storing data for your Amazon EC2 applications. Please see Snowball Edge's technical documentation for performance best practices, and recommended drivers.

Regional Availability

Q: In what Regions are Snowball Edge available?

Check the Regional Service Availability pages for the latest information.

Q: Can a Snowball Edge be shipped to an alternate AWS Region?

No. Snowball Edge devices are designed to be requested and used within a single AWS Region. The device may not be requested from one Region and returned to another. Snowball Edge devices used for imports or exports from an AWS Region in the EU may be used with any of the other EU countries. Check the Regional Service Availability pages for the latest information.

Security

Q: Does Snowball Edge encrypt my data?

Snowball Edge encrypts all data with 256-bit encryption. You manage your encryption keys by using the AWS Key Management Service (AWS KMS). Your keys are never stored on the device and all memory is erased when it is disconnected and returned to AWS.

Q: How does Snowball Edge physically secure my data?

In addition to using a tamper-resistant enclosure, Snowball Edge uses industry-standard Trusted Platform Modules (TPM) designed to detect any unauthorized modifications to the hardware, firmware, or software. AWS visually and cryptographically inspects every device for any signs of tampering and to verify that no changes were detected by the TPM.

Q: How does Snowball Edge help digitally secure my data?

Snowball Edge is designed with security in mind for the most sensitive data. All data is encrypted by keys provided by you through AWS Key Management Service (KMS). The keys are not permanently stored on the device and are erased after loss of power. Applications and Lambda functions run in a physically isolated environment and do not have access to storage. Lastly, after your data has been transferred to AWS, your data is erased from the device using standards defined by National Institute of Standards and Technology. Snowball Edge devices are hardened against attack and all configuration files are encrypted and signed with keys that are never present on the device.

Q: Is there a way to easily track my data transfer jobs?

Snowball Edge uses an innovative, E Ink shipping label designed to ensure the device is automatically sent to the correct AWS facility. When you have completed your data transfer job, you can track it by using Amazon SNS generated text messages or emails, and the console.

Q: Can I get a history of Snowball API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. To receive a history of Snowball API calls made on your account, you simply turn on CloudTrail in the AWS Management Console; The following API calls in Snowball are not recorded and delivered: DescribeAddress (in response), CreateAddress (in request), DescribeAddresses (in response).

Import Data with Snowball Edge

Q: How do I transfer my data to the Snowball Edge?

After you have connected and activated the Snowball Edge, you can transfer data from local sources to the device through the S3-compatible endpoint or the NFS file interface, both available on the device. You can also use the Snowball client to copy data.

Q: When can I delete the data on my disk(s) after I've copied the data onto Snowball Edge and shipped the Snowball Edge back to AWS?

Wait to confirm that the Snowball Edge has been received by AWS and your data has successfully been transferred into appropriate S3 buckets prior to you deleting any data on your disk(s). While AWS verifies the integrity of files copied to Snowball Edge during the S3 transfer, it is your responsibility to verify the integrity of data before deleting it from your disk(s). AWS is not liable for any lost or corrupted data during copy or transit.

Q: What do I do when the data has been transferred to the Snowball Edge?

When the data transfer job is complete, the E Ink display on the Snowball Edge automatically updates the return shipping label to indicate the correct AWS facility to ship to. Just drop off the Snowball Edge at the nearest UPS and you're all set. You can track the status of your transfer job through Amazon SNS generated text messages or emails, or directly in the AWS Management Console.

Q. Is it possible to stop data ingestion from an AWS Snowball device once returned to AWS?

Yes. You can stop the data ingestion to an Amazon Simple Storage Service (Amazon S3) bucket by cancelling the job in the AWS Snow Management Console or by contacting AWS support.

Q. Is it possible to stop data from being deleted on a returned AWS Snowball device after the data import is completed?

No. AWS uses an automated workflow process to securely delete the data ingest and clean the returned device with a complete erasure of the Snowball device according to NIST 800-88 standards. Additionally, it is not possible to return the same device back to you after the data import is complete.

Export Data with Snowball Edge

Q: What does it cost to export my data?

In addition to the Export job fees detailed on our pricing page, you will also be charged all fees incurred to retrieve your data from Amazon S3.

Q: How quickly can I access my exported data?

We typically start exporting your data within 24 hours of receiving your request, and exporting data can take as long as a week. Once the job is complete and the device is ready, we ship it to you using the shipping options you selected when you created the job.

Q. Do I get a checksum or any kind of receipt on what was loaded into Amazon S3?

Yes. AWS saves an import log report to your bucket. This report contains per file information including the date and time of the upload, the Amazon S3 key, MD5 checksum, and number of bytes.

Q. I created an export job of my Amazon S3 bucket and it contains keys which are in an Amazon S3 Glacier storage class. Will the keys be exported ?

The Snowball export job from Amazon S3 workflow does not have access to the objects stored in the Amazon S3 Glacier or Amazon S3 Glacier Deep Archive storage classes. You must first restore these objects from the Glacier or Glacier Deep archive for a minimum of 10 days or until the Snow export job completes to ensure that these restored objects are successfully copied to the Snowball device.

Large Data Migration Manager

Q: Why should I use Large Data Migration Manager?

Large Data Migration Manager helps you plan, and monitor your large data migration from 500TB minimum to petabytes of data. First, Large Data Migration Manager enables you to create a plan for your migration projects that use multiple AWS Snow Family devices to complete your petabyte scale data migration or data movement from the rugged, mobile edge. Creating a plan helps you and your partners onboard to Snow and align on project goals such as data size to be migrated and project duration. Once a plan is in place, Large Data Migration Manager provides a central location in AWS Snow Family management console for you to stay updated with the progress of all your Snow jobs (number of outstanding jobs, current data ingested etc.), and view estimated schedules for placing the next job orders. Finally, you can control the project plan as you monitor the migration and can extend or end the migration when you deem appropriate.

Prior to the Large Data Migration Manager launch, you had to track all of this information yourself and spend time coordinating with your partners and AWS for tracking data ingestion progress and placing job orders. Large Data Migration Manager saves you time and effort by keeping track of all project details and allows you to focus on your overall goal of data migration.

Q: How do I get started with using Large Data Migration Manager?

You start by creating a data migration or data movement project plan in the AWS Management Console. To create a plan, you are prompted for your import job type specifics that includes plan name, service access roles and notification preference. Once a plan is created, you need to create a site where the Snow devices will be shipped. Site information details include name and shipping address for each site, data size amount, number of concurrent Snow jobs, Snow job type, Snow device, fill rate (as per the last monitored data), project start and end dates. After you create your site, you can review your automatically created Snow job ordering schedule, which helps you know when to order your Snow jobs. You can either clone from prior existing jobs or add a job that was already created to the site.

Q: How do I update my project plan using Large Data Migration Manager?

You can update your project plan in one of two ways: (1) you modify your plan information by updating data size amount, number of concurrent Snow jobs, or (2) Snow's Data Migration Manager calculates the average Snow job duration (from order creation to completion) and average fill rate on a per site basis to automatically adjust your plan. Data Migration Manager then uses these plan updates to adjust your ordering schedule to inform you of your project status and if additional Snow jobs are required.

Q: How do I monitor the Snow jobs using Large Data Migration Manager?

You monitor your Snow jobs using the Data Migration Manager dashboard or viewing your project plan summary. Using the Data Migration Manager dashboard, you can quickly monitor project status and identify issues at a glance. With this new capability, you can track your overall migration or data movement transfer progress, time remaining, Snow job status, site status, average Snow job fill rate, average Snow job duration, and upcoming order schedule across your plan and site.

Q: Do I have to pay money to use Large Data Migration Manager?

No. Large Data Migration Manager is available to customers using AWS Snow Family.

Q: In what regions is Large Data Migration Manager available?

Large Data Migration Manager is available in all commercial regions where Snowball Edge devices are available.

AWS OpsHub

Q: What is AWS OpsHub for Snow Family?

AWS OpsHub is an application that you can download from the Snowball resources page. It offers a graphical user interface for managing the AWS Snow Family devices. AWS OpsHub makes it easy to setup and manage AWS Snowball devices enabling you to rapidly deploy edge computing workloads and simplify data migration to the cloud. With just a few clicks in AWS OpsHub, you can unlock and configure devices, drag-and-drop data to devices, launch and manage EC2 instances on devices, or monitor device metrics. AWS OpsHub is available globally at no extra charge.

Q: How does AWS OpsHub for Snow Family work?

AWS OpsHub is an application that you can download and install on any Windows or Mac client machine, such as a laptop. Once you have installed AWS OpsHub and have your AWS Snow Family device on site, open AWS OpsHub and unlock the device. You will then be presented with a dashboard showing your device and its system metrics. You can then begin deploying your edge applications or migrating your data to the device with just a few clicks.

Q: Can I use AWS OpsHub with a Snow Family device that I ordered before AWS OpsHub launched?

Yes. However, the task automation features are available for only Snow Family devices ordered after AWS OpsHub launched on April 16, 2020. All other functionality will be available for all devices, including those ordered before AWS OpsHub launched.

Q: When do I use AWS OpsHub compared to the AWS Management Console?

You use AWS OpsHub to manage and operate your AWS Snow Family devices and the AWS services that run on them. AWS OpsHub is an application that runs on a local client machine, such as a laptop, and can operate in disconnected or connected environments. In contrast, you use the AWS Management Console to manage and operate the AWS services running in the cloud. The AWS Management Console is a web-based application that operates when you have a connection to the internet.

Q: How do I keep my AWS OpsHub software up to date?

AWS OpsHub will automatically check for AWS OpsHub software updates when the client machine that AWS OpsHub is running on is connected to the internet. When there is a software update, you will be notified on the application and will be given the option to download and update the latest software. Additionally, you can visit the Snowball resources page and check for the latest version of AWS OpsHub.

Q: Does AWS OpsHub validate and encrypt the data I transfer to the AWS Snow Family devices?

Yes. When you copy data to AWS Snow Family devices using AWS OpsHub, checksums are used to ensure that the data you copy to the device is the same as the original. Also, all data written to AWS Snow Family devices is encrypted by default.

Tape Data Migration

Q: Why should I use AWS Snowball to migrate my tape data?

When you use a Snowball Edge Storage Optimized device with Tape Gateway, you can quickly and securely migrate petabyte-scale physical tape data to S3 Glacier Flexible Retrieval or S3 Glacier Deep Archive—reducing your data storage costs without changing your tape-based backup workflows.

Q: How do I get started with Snowball to migrate my tape data?

To get started, in the AWS Snow Family console, order a Snowball Edge Storage Optimized device with Tape Gateway. When you receive your device, unlock it, and connect to your local network. Then start Tape Gateway, which looks like a physical tape library. Connect to AWS and copy data from physical tapes to virtual tapes on Tape Gateway using your existing backup application.

After you complete your data copy, ship the Snowball Edge device back to AWS. Your data will be stored in either S3 Glacier Flexible Retrieval or S3 Glacier Deep Archive storage class. You can view your virtual tapes stored on AWS through the AWS Storage Gateway console and access data on them through a Tape Gateway that runs on premises as a virtual machine or hardware appliance or on an Amazon Elastic Compute Cloud (Amazon EC2) instance.

Q: Can I export virtual tapes stored on AWS to my on-premises data center using Snowball?

No, you cannot export your virtual tapes stored on AWS to your data center using Snowball. You can use a Snowball Edge Storage Optimized device with Tape Gateway to import data into AWS. To retrieve virtual tapes stored on AWS, you can use a Tape Gateway that runs on premises as a virtual machine or hardware appliance, or on an Amazon EC2 instance.

Q: When can I delete the data on my physical tapes after I've copied the data to Snowball and shipped the device back to AWS?

Before you delete any data on your physical tapes, wait for confirmation that your Snowball Edge Optimized device has been received by AWS and your data has been successfully transferred as virtual tapes to S3 Glacier Flexible Retrieval or S3 Glacier Deep Archive. To access your virtual tapes, navigate to your AWS Storage Gateway console. While AWS verifies the integrity of data copied to your Snowball Edge device during the migration, it is your responsibility to verify the integrity of data before deleting it from your physical tapes. AWS is not liable for any lost or corrupted data during copy or transit.

Q: Where are my virtual tapes stored on AWS?

You can choose to store your virtual tapes in either S3 Glacier Flexible Retrieval or S3 Glacier Deep Archive storage class in the AWS Region you specified when you ordered a Snowball Edge Storage Optimized device in the AWS Snow Family console.

Q: How do I access my data on virtual tapes?

You can read-from and write-to virtual tapes on Snowball until you export virtual tapes from your backup application. After you export virtual tapes, your data stored on Snowball is inaccessible. After you ship a Snowball Edge Storage Optimized device back to AWS and the transfer of your physical tapes to virtual tapes is complete, you can access data stored on virtual tapes through a Tape Gateway that runs on premises as a virtual machine or a hardware appliance or on an Amazon EC2 instance on AWS.

Q: Is my network connection to AWS used for data migration?

No, your network connection to AWS is not used for data migration. Data from your backup application is copied to your Snowball Edge Storage Optimized device and not synchronized with AWS over the network.

Q: Where can I learn more about Tape Gateway in AWS Storage Gateway?

Tape Gateway is a tape storage interface in Storage Gateway.

Workflow Integration Tools

Q: Does the Snowball Edge support API access?

Yes. The Snowball Job Management API provides programmatic access to the job creation and management features of a Snowball or Snowball Edge. It is a simple, standards-based REST web service interface, designed to work with any Internet development environment.

Q: What can I do with the Snowball Job Management API?

The AWS Snowball Job Management API allows partners and customers to build custom integrations to manage the process of requesting Snowballs and communicating job status. The API provides a simple web service interface that you can use to create, list, update, and cancel jobs from anywhere on the web. Using this web service, developers can easily build applications that manage Snowball jobs.

Q: What is the S3 Adapter?

The S3 SDK Adapter for Snowball provides an S3-compatible interface for reading and writing data on a Snowball or Snowball Edge.

Q: What can I do with the S3 Adapter?

The S3 Adapter allows customers to help applications write data from file and non-file sources to S3 buckets on the Snowball or Snowball Edge device. It also includes interfaces to copy data with the same encryption as is available through the Snowball client.

Q: Why would I use the S3 Adapter rather than the Snowball Client?

The Snowball Client is a turnkey tool that makes it easier to copy file-based data to Snowball. Customers who prefer a tighter integration can use the S3 Adapter to easily extend their existing applications and workflows to seamlessly integrate with Snowball.

Q: How is my data secured when I use the S3 Adapter?

The S3 Adapter writes data using the same advanced encryption mechanism that the Snowball Client provides.

Q: Which programming languages does the Snowball S3 Adapter support?

The S3 Adapter communicates over REST which is language-agnostic.

AWS Snowmobile FAQs

General

Q: What is AWS Snowmobile?

AWS Snowmobile is the first exabyte-scale data migration service that allows you to move very large datasets from on-premises to AWS. Each Snowmobile is a secured data truck with up to 100PB storage capacity that can be dispatched to your site and connected directly to your network backbone to perform high-speed data migration. You can quickly migrate an exabyte of data with ten Snowmobiles in parallel from a single location or multiple data centers. Snowmobile is offered by AWS as a managed service.

Q: How does Snowmobile work?

After you have placed your inquiry for a Snowmobile, AWS personnel will contact you to determine requirements for deploying a Snowmobile and schedule the job, and will drive the required Snowmobile equipment to your site. Once on site, they will connect it to your local network so that you can use your high-speed local connection to quickly transfer data from your local storage appliances or servers to the Snowmobile. After the data transfer is complete, the Snowmobile will be returned to your designated AWS region where your data will be uploaded into the AWS storage services you have selected, such as S3 or Glacier. Finally, AWS will work with you to validate that your data has been successfully uploaded.

Q: Who should use a Snowmobile?

Snowmobile enables customers to quickly migrate exabyte-scale datasets from on-premises to AWS in a more secure, fast, and low-cost manner. Use cases include migrating 100's of petabytes of data (such as video libraries, genomic sequences, seismic data, satellite images), and financial records to run big data analytics on AWS, or shutting down legacy data centers and moving all local data in exabytes to AWS. Before Snowmobile, migrating data at such scale would typically take years which was too slow for many customers. With Snowmobile, you can now request multiple data trucks each with up to 100PB capacity to be dispatched on-site, connected to your local high speed network backbone, and transfer your exabyte-scale datasets to AWS in as quickly as a few weeks, plus transport time.

Q: What are the specifications of a Snowmobile?

Each Snowmobile comes with up to 100PB of storage capacity housed in a 45-foot long High Cube shipping container that measures 8 foot wide, 9.6 foot tall and has a curb weight of approximately 68,000 pounds. The ruggedized shipping container is tamper-resistant, water-resistant, temperature controlled, and GPS-tracked.

Q: How should I choose between Snowmobile and Snowball?

To migrate large datasets of 10PB or more in a single location, you should use Snowmobile. For datasets less than 10PB or distributed in multiple locations, you should use Snowball. In addition, you should evaluate the amount of available bandwidth in your network backbone. If you have a high speed backbone with hundreds of Gb/s of spare throughput, then you can use Snowmobile to migrate the large datasets all at once. If you have limited bandwidth on your backbone, you should consider using multiple Snowballs to migrate the data incrementally.

Q: How much data can I transfer to Snowmobile?

Each Snowmobile has a total capacity of up to 100 petabytes and multiple Snowmobiles can be used in parallel to transfer exabytes of data.
Q: Are there site requirements to use a Snowmobile?

The Snowmobile needs physical access to your data center to allow for network connectivity. It comes with a removable connector rack with up to two kilometers of networking cable that can directly connect to the network backbone in your data center. The Snowmobile can be parked in a covered area at your data center, or in an uncovered area that is adjacent to your data center, and close enough to run the networking cable. The parking area needs to hold a standard 45-foot High Cube trailer with a minimum of 6'-0" (1.83m) of peripheral clearance. Snowmobile can operate at ambient temperatures up to 85F (29.4C) before an auxiliary chiller unit is required. AWS can provide the auxiliary chiller if needed based on the site survey findings.
Q: How is a Snowmobile powered?

A fully powered Snowmobile requires ~350KW. Snowmobile can be connected to available utility power sources at your location if sufficient capacity is available. Otherwise, AWS can dispatch a separate generator set along with the Snowmobile if your site permits such generator use. This generator set takes a similar amount of space as the Snowmobile which is parking for a vehicle approximately the same size as a 45-foot container trailer.

Getting started

Q: What is a Snowmobile job?

A Snowmobile job encapsulates the end-to-end data migration process using a Snowmobile. There are five main steps:

1. Site Survey, where AWS personnel will work with you to understand your migration objectives, data center environment, and network configurations in order to help you determine a migration plan;
2. Site Preparation, where you (the customer) will identify and make available local resources such as parking space and power source for the Snowmobile, local security, network address, ports, and available rack positions to connect the Snowmobile with the local network backbone;
3. Dispatch and Setup, where AWS personnel will dispatch a Snowmobile to your site and configure it for you so it can be accessed securely as a network storage target;
4. Data Migration, where you will copy data from any number of sources within your data center to the Snowmobile, and
5. Return and Upload, where the Snowmobile is returned to an AWS region that you have designated where your data will be uploaded into the AWS storage services you have selected.

Q: How do I connect my data center to Snowmobile?

Each Snowmobile comes with a removable high-speed connector rack on wheels with two kilometers of ruggedized networking cable. The connector rack can be rolled to a location inside your data center and connected directly to your network backbone. This way, the Snowmobile will operate as a network storage target inside your network for you to perform high-speed data transfer.
Q: How do I copy my data to a Snowmobile?

Once the Snowmobile is connected to your data center, it will appear as a network storage target. You can copy data from local storage devices to the Snowmobile using the same tools and in the same manner as data copied to any network attached storage device with an NFS interface.
Q: Can I connect to Snowmobile via a NFS endpoint?

Yes. The Snowmobile will appear as a standard NFS mount on your network that you can connect to via your existing tools and applications.
Q: How do I get started with Snowmobile?

Using Snowmobile

Q: How long does it take to transfer my data to a Snowmobile?

The Snowmobile is designed to transfer data at a rate up to 1 Tb/s, which means you could fill a 100PB Snowmobile in less than 10 days. The actual transfer speed may vary depending on the available local network capacity at your site and the speed of your on-premises storage devices. AWS will provide a way for you to test your local network throughput and the copy speed from your data sources to properly size the Snowmobile job, before dispatching the Snowmobile.
Q: What type of connections does Snowmobile provide?

The Snowmobile comes with a removable connector cabinet that needs to be mounted on one of your data center racks where it can be connected directly to your high-speed network backbone. The connector rack provides multiple 40Gb/s interfaces that can transfer up to 1 Tb/s in aggregate.
Q: After my data has been imported to AWS, what happens to the copy on Snowmobile?

When the data import has been processed and verified, AWS performs a software erasure of the Snowmobile that follows the National Institute of Standards and Technology (NIST) guidelines for media sanitization (NIST 800-88).
Q: How do I verify that my data has been successfully copied to Snowmobile?

At the time your data is copied into the Snowmobile, a set of logs will be generated with checksums for each file transferred. These logs are available to you for verification. The logs are also used when data is imported from the Snowmobile to AWS to verify that all data has been transferred successfully.
Q: Do I need to keep a local copy of my data while a copy is shipped back to AWS on a Snowmobile?

Yes. You should always keep your source copy until AWS has worked with you to verify that the Snowmobile copy has been successfully uploaded to AWS.
Q: Can I export data from AWS with Snowmobile?

Snowmobile does not support data export. It is designed to let you quickly, easily, and more securely migrate exabytes of data to AWS. When you need to export data from AWS, you can use AWS Snowball Edge to quickly export up to 100TB per appliance and run multiple export jobs in parallel as necessary.

Security

Q: How is Snowmobile designed to keep data secure digitally?

Your data is encrypted with keys you provided before it is written to the Snowmobile. All data is encrypted with 256-bit encryption. You can manage your encryption keys with the AWS Key

Management Service (KMS). Your keys are never permanently stored on the Snowmobile, and are erased as soon as power is removed from the Snowmobile.
Q: How is Snowmobile designed to keep data secure physically?

In addition to digital encryption, the Snowmobile is only operated by AWS personnel and physical access to the data container is controlled via secure access hardware controls. All data storage equipment is separated from the network access ports used to load or remove data. This way, physical access to the data container is not needed to operate the container after it has been set up. In addition, Snowmobile is protected by 24/7 video surveillance and alarm monitoring, GPS tracking, and may be escorted by a security vehicle during transit.
Q: Can I still choose to encrypt my data before transferring to Snowmobile?

Yes. You can always encrypt your data before transferring it to the Snowmobile.

Region availability

Q: What AWS Regions are supported?

Snowmobile can be made available for use with AWS services in specific AWS regions. To discuss data transport needs specific for your region please follow up with AWS Sales, or see the Regional Service Availability pages for more information.

Amazon Elastic Container Registry (Amazon ECR)

Easily store, share, and deploy your container software anywhere.

How it works

Amazon ECR is a fully managed container registry offering high-performance hosting, so you can reliably deploy application images and artifacts anywhere.

Use cases

Manage software vulnerabilities
Meet your image compliance security requirements using the tightly integrated Amazon Inspector vulnerability management service to automate vulnerability assessment scanning and remediation ticket routing.

Streamline your deployment workloads
Publish containerized applications with a single command and easily integrate your self-managed environments.

Manage image lifecycle policies
Automatically preserve the most recent images and archive ones you don't need. Use rules and tagging to access images quickly.

What is Amazon Elastic Container Registry?

Amazon Elastic Container Registry (Amazon ECR) is an AWS managed container image registry service that is secure, scalable, and reliable. Amazon ECR supports private repositories with resource-based permissions using AWS IAM. This is so that specified users or Amazon EC2 instances can access your container repositories and images. You can use your preferred CLI to push, pull, and manage Docker images, Open Container Initiative (OCI) images, and OCI compatible artifacts.

Note

Amazon ECR supports public container image repositories as well. For more information, see What is Amazon ECR Public in the *Amazon ECR Public User Guide*.

The AWS container services team maintains a public roadmap on GitHub. It contains information about what the teams are working on and allows all AWS customers the ability to give direct feedback. For more information, see AWS Containers Roadmap.

Components of Amazon ECR

Amazon ECR contains the following components:

Registry

An Amazon ECR private registry is provided to each AWS account; you can create one or more repositories in your registry and store images in them. For more information, see Amazon ECR private registry.

Authorization token

Your client must authenticate to Amazon ECR registries as an AWS user before it can push and pull images. For more information, see Private registry authentication.

Repository

An Amazon ECR repository contains your Docker images, Open Container Initiative (OCI) images, and OCI compatible artifacts. For more information, see Amazon ECR private repositories.

Repository policy

You can control access to your repositories and the images within them with repository policies. For more information, see Private repository policies.

Image

You can push and pull container images to your repositories. You can use these images locally on your development system, or you can use them in Amazon ECS task definitions and Amazon EKS pod specifications. For more information, see Using Amazon ECR images with Amazon ECS and Using Amazon ECR Images with Amazon EKS.

Features of Amazon ECR

Amazon ECR provides the following features:

- Lifecycle policies help with managing the lifecycle of the images in your repositories. You define rules that result in the cleaning up of unused images. You can test rules before applying them to your repository. For more information, see Lifecycle policies.
- Image scanning helps in identifying software vulnerabilities in your container images. Each repository can be configured to **scan on push**. This ensures that each new image pushed to the repository is scanned. You can then retrieve the results of the image scan. For more information, see Image scanning.
- Cross-Region and cross-account replication makes it easier for you to have your images where you need them. This is configured as a registry setting and is on a per-Region basis. For more information, see Private registry settings.
- Pull through cache rules provide a way to cache repositories in remote public registries in your private Amazon ECR registry. Using a pull through cache rule, Amazon ECR will periodically reach out to the remote registry to ensure the cached image in your Amazon ECR private registry is up to date. For more information, see Using pull through cache rules.

Amazon Elastic Container Registry FAQs

General

Q: What is Amazon Elastic Container Registry (Amazon ECR)?

Amazon ECR is a fully managed container registry that makes it easy for developers to share and deploy container images and artifacts. Amazon ECR is integrated with Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), and AWS Lambda, simplifying your development to production workflow. Amazon ECR eliminates the need to operate your own container repositories or worry about scaling the underlying infrastructure. Amazon ECR hosts your images in a highly available and scalable architecture, allowing you to deploy containers for your applications reliably. Integration with AWS Identity and Access Management (IAM) provides resource-level control of each repository that lets you share images across your organization or with anyone in the world.

Q: Why should I use Amazon ECR?
Amazon ECR eliminates the need to operate and scale the infrastructure required to power your container registry. Amazon ECR uses Amazon Simple Storage Service (S3) for storage to make your container images highly available and accessible, allowing you to deploy new containers for your applications reliably. Amazon ECR transfers your container images over HTTPS and automatically encrypts your images at rest. You can configure policies to manage permissions for each repository and restrict access to IAM users, roles, or other AWS accounts. Amazon ECR integrates with Amazon ECS, Amazon EKS, AWS Fargate, AWS Lambda, and the Docker CLI, allowing you to simplify your development and production workflows. You can easily push your container images to Amazon ECR using the Docker CLI from your development machine, and Amazon container orchestrators or compute can pull them directly for production deployments.

Q: What is the pricing for Amazon ECR?
With Amazon ECR, there are no upfront fees or commitments. You pay only for the amount of data you store in your public or private repositories, and data transferred to the internet.

Q: Is Amazon ECR a global service?
Amazon ECR is a Regional service and is designed to give you flexibility in how images are deployed. You have the ability to push/pull images to the same AWS Region where your Docker cluster runs for the best performance. You can also access Amazon ECR anywhere that Docker runs, such as desktops and on-premises environments. Pulling images between Regions or out to the internet will have additional latency and data transfer costs.

Q: Can Amazon ECR host public container images?
Yes. Amazon ECR has a highly available container registry and website that makes it easy for you to share or search for public container software. Anyone with or without an AWS account can use the Amazon ECR public gallery to search for and download commonly used container images such as operating systems, AWS published images, and files, such as Helm charts, for Kubernetes.

Q: What is the difference between Amazon ECR public and private repositories?
A private repository does not offer content search capabilities and requires Amazon IAM-based authentication using AWS account credentials before allowing images to be pulled. A public repository has descriptive content and allows anyone anywhere to pull images without needing an AWS account or using IAM credentials. Public repository images are also available in the Amazon ECR public gallery.

Q: What compliance capabilities can I enable on Amazon ECR?
You can use AWS CloudTrail on Amazon ECR to provide a history of all API actions such as who pulled an image and when tags were moved between images. Administrators can also find which EC2 instances pulled which images.

Using Amazon ECR

Q: How do I get started using Amazon ECR?
The best way to get started with Amazon ECR is to use the Docker CLI to push and pull your first image. Visit our Getting Started page for more information.

Q: Can I access Amazon ECR inside a VPC?
Yes. You can set up AWS PrivateLink endpoints to allow your instances to pull images from your private repositories without traversing through the public internet.

Q: What's the best way to manage my repositories and images?
Amazon ECR provides a command line interface and APIs to create, monitor, and delete repositories and set repository permissions. You can perform the same actions in the Amazon ECR console, which can be accessed via the "Repositories" section of the Amazon ECR console. Amazon ECR also integrates with the Docker CLI, allowing you to push, pull, and tag images on your development machine.

Q: How do I publicly share an image using Amazon ECR?
You publish an image to the Amazon ECR public gallery by signing into your AWS account and pushing to a public repository you create. You are assigned a unique alias per account to use in image URLs that identifies all public images that you publish.

Q: Can I use a custom alias for my public images?
Yes. You can request a custom alias such as your organization or project name, unless it's a reserved alias. Names that identify AWS services are reserved. Names that identify AWS Marketplace sellers may also be reserved. We will review and approve your custom alias request within a few days unless your alias request violates the AWS Acceptable Use Policy or other AWS policies.

Q: How do I pull a public image from Amazon ECR?
You pull using the familiar 'docker pull' command with the URL of the image. You can easily search for this URL by finding images using a publisher alias, image name, or image description using the Amazon ECR public gallery. Image URLs are in the format public.ecr.aws/<alias>/<image>:<tag>, for example public.ecr.aws/eks/aws-alb-ingress-controller:v1.1.5

Q: Does Amazon ECR replicate images across AWS Regions?
Yes. Amazon ECR is designed to give you flexibility in where you store and how you deploy your images. You can create deployment pipelines that build images, push them to Amazon ECR in one Region, and Amazon ECR can automatically replicate them to other Regions and accounts for deployment to multi-Region clusters.

Q: Can I use Amazon ECR within local and on-premises environments?
Yes. You can access Amazon ECR anywhere that Docker runs such as desktops and on-premises environments.

Q: Does the Amazon ECR public gallery provide AWS-published images?
Yes. Services such as Amazon EKS, Amazon SageMaker and AWS Lambda publish their official public use container images and artifacts to Amazon ECR.

Q: Does Amazon ECR work with Amazon ECS?
Yes. Amazon ECR is integrated with Amazon ECS, allowing you to easily store, run, and manage container images for applications running on Amazon ECS. All you need to do is specify the Amazon ECR repository in your task definition and Amazon ECS will retrieve the appropriate images for your applications.

Q: Does Amazon ECR work with AWS Elastic Beanstalk?
Yes. AWS Elastic Beanstalk supports Amazon ECR for both single and multi-container Docker environments, allowing you to easily deploy container images stored in Amazon ECR with AWS Elastic Beanstalk. All you need to do is specify the Amazon ECR repository in your Dockerrun.aws.json configuration and attach the AmazonEC2ContainerRegistryReadOnly policy to your container instance role.

Q: What version of Docker Engine does Amazon ECR support?
Amazon ECR currently supports Docker Engine 1.7.0 and up.

Q: What version of the Docker Registry API does Amazon ECR support?
Amazon ECR supports the Docker Registry V2 API specification.

Q: Will Amazon ECR automatically build images from a Dockerfile?
No. However, Amazon ECR integrates with a number of popular CI/CD solutions to provide this capability. See the Amazon ECR Partners page for more information.

Q: Does Amazon ECR support federated access?
Yes. Amazon ECR is integrated with AWS Identity and Access Management (IAM), which supports identity federation for delegated access to the AWS Management Console or AWS APIs.

Q: What version of the Docker Image Manifest specification does Amazon ECR support?
Amazon ECR supports the Docker Image Manifest V2, Schema 2 format. In order to maintain backwards compatibility with Schema 1 images, Amazon ECR will continue to accept images uploaded in the Schema 1 format. Additionally, Amazon ECR can down-translate from a Schema 2 to a Schema 1 image when pulling with an older version of Docker Engine (1.9 and below).

Q: Does Amazon ECR support the Open Container Initiative (OCI) format?
Yes. Amazon ECR is compatible with the Open Container Initiative (OCI) image specification, letting you push and pull OCI images and artifacts. Amazon ECR can also translate between Docker Image Manifest V2, Schema 2 images and OCI images on pull.

Security

Q: How does Amazon ECR help ensure that container images are secure?
Amazon ECR automatically encrypts images at rest using Amazon S3 server-side encryption or AWS KMS encryption and transfers your container images over HTTPS. You can configure policies to manage permissions and control access to your images using AWS Identity and Access Management (IAM) users and roles without having to manage credentials directly on your EC2 instances.

Q: How can I use AWS Identity and Access Management (IAM) for permissions?
You can use IAM resource-based policies to control and monitor who and what (e.g., EC2 instances) can access your container images, as well as how, when, and where they can access them. To get started, use the AWS Management Console to create resource-based policies for your repositories. Alternatively, you can use sample policies and attach them to your repositories via the Amazon ECR CLI.

Q: Can I share my images across AWS accounts?
Yes. Here is an example of how to create and set a policy for cross-account image sharing.

Q: Does Amazon ECR scan container images for vulnerabilities?
You can enable Amazon ECR to automatically scan your container images for a broad range of operating system vulnerabilities. You can also scan images using an API command, and

Amazon ECR will notify you over API and in the console when a scan completes. For enhanced image scanning, you can turn on Amazon Inspector.

What Is Amazon Elastic Container Registry Public?

Amazon Elastic Container Registry Public is a managed AWS container image registry service that is secure, scalable, and reliable. Amazon ECR supports public image repositories with resource-based permissions using AWS IAM so that specific users can access your public repositories to push images. Developers can use their preferred CLI to push and manage Docker images, Open Container Initiative (OCI) images, and OCI compatible artifacts. Your images are publicly available to pull, either anonymously or using an Amazon ECR Public authentication token.

Note

Amazon ECR supports private container image repositories as well. For more information, see What is Amazon ECR in the *Amazon Elastic Container Registry User Guide*.

The AWS container services team maintains a public roadmap on GitHub. It contains information about what the teams are working on and allows all AWS customers the ability to give direct feedback. For more information, see AWS Containers Roadmap.

Components of Amazon ECR Public

Amazon ECR Public contains the following components:

Amazon ECR Public Gallery

The Amazon ECR Public Gallery is the public portal that lists all public repositories hosted on Amazon ECR Public. Visit the Amazon ECR Public Gallery at https://gallery.ecr.aws. For more information, see Using the Amazon ECR Public Gallery.

Registry

A public registry is provided to each AWS account; you can create public image repositories in your public registry and store images in them. For more information, see Amazon ECR public registries.

Authorization token

Your client must authenticate to a public registry as an AWS user before it can push images to a public repository. For image pulls, Amazon ECR Public accepts both anonymous pulls and pulls using an authentication token. For more information, see Registry authentication.

Repository

An Amazon ECR image repository contains your Docker images, Open Container Initiative (OCI) images, and OCI compatible artifacts. For more information, see Amazon ECR public repositories.

Repository policy

You can control access to your repositories and the images within them with repository policies. For more information, see Public repository policies.

Image

You can push and pull container images to your repositories. You can use these images locally on your development system, or you can use them in Amazon ECS task definitions and Amazon EKS pod specifications.

Amazon Elastic Container Service (Amazon ECS)

Run highly secure, reliable, and scalable containers.

How it works

Amazon ECS is a fully managed container orchestration service that makes it easy for you to deploy, manage, and scale containerized applications.

Use cases

Deploy in a hybrid environment
Build container-based applications on-premises or in the cloud with Amazon ECS Anywhere and enjoy consistent tooling, management, workload scheduling, and monitoring across environments.

Support batch processing
Plan, schedule, and execute batch computing workloads across the full range of AWS services, including Amazon Elastic Compute Cloud (EC2), Fargate, and Amazon EC2 Spot Instances.

Scale web applications
Automatically scale and run web applications in multiple Availability Zones with the performance, scale, reliability, and availability of AWS.

What is Amazon Elastic Container Service?

Amazon Elastic Container Service (Amazon ECS) is a highly scalable and fast container management service. You can use it to run, stop, and manage containers on a cluster. With Amazon ECS, your containers are defined in a task definition that you use to run an individual task or task within a service. In this context, a service is a configuration that you can use to run and maintain a specified number of tasks simultaneously in a cluster. You can run your tasks and services on a serverless infrastructure that's managed by AWS Fargate. Alternatively, for more control over your infrastructure, you can run your tasks and services on a cluster of Amazon EC2 instances that you manage.

Amazon ECS provides the following features:

- A serverless option with AWS Fargate. With AWS Fargate, you don't need to manage servers, handle capacity planning, or isolate container workloads for security. Fargate handles the infrastructure management aspects of your workload for you. You can schedule the placement of your containers across your cluster based on your resource needs, isolation policies, and availability requirements.
- Integration with AWS Identity and Access Management (IAM). You can assign granular permissions for each of your containers. This allows for a high level of isolation when building your applications. In other words, you can launch your containers with the security and compliance levels that you've come to expect from AWS.
- AWS managed container orchestration. As a fully managed service, Amazon ECS comes with AWS configuration and operational best practices built-in. This also means that you don't need to manage control plane, nodes, or add-ons. It's integrated with both AWS and third-party tools, such as Amazon Elastic Container Registry and Docker. This integration makes it easier for teams to focus on building the applications, not the environment.

- Continuous integration and continuous deployment (CI/CD). This is a common process for microservice architectures that are based on Docker containers. You can create a CI/CD pipeline that takes the following actions:
    - Monitors changes to a source code repository
    - Builds a new Docker image from that source
    - Pushes the image to an image repository such as Amazon ECR or Docker Hub
    - Updates your Amazon ECS services to use the new image in your application
- Support for service discovery. This is a key component of most distributed systems and service-oriented architectures. With service discovery, your microservice components are automatically discovered as they're created and terminated on a given infrastructure.
- Support for sending your container instance log information to CloudWatch Logs. After you send this information to Amazon CloudWatch, you can view the logs from your container instances in one convenient location. This prevents your container logs from taking up disk space on your container instances.

The AWS container services team maintains a public roadmap on GitHub. The roadmap contains information about what the teams are working on and enables AWS customers to provide direct feedback. For more information, see AWS Containers Roadmap on the GitHub website.

Launch types

There are two models that you can use to run your containers:

- Fargate launch type - This is a serverless pay-as-you-go option. You can run containers without needing to manage your infrastructure.
- EC2 launch type - Configure and deploy EC2 instances in your cluster to run your containers.

The Fargate launch type is suitable for the following workloads:

- Large workloads that need to be optimized for low overhead
- Small workloads that have occasional burst
- Tiny workloads
- Batch workloads

The EC2 launch type is suitable for the following workloads:

- Workloads that require consistently high CPU core and memory usage
- Large workloads that need to be optimized for price
- Your applications need to access persistent storage
- You must directly manage your infrastructure

Access Amazon ECS

You can create, access, and manage your Amazon ECS resources using any of the following interfaces:

- **AWS Management Console** — Provides a web interface that you can use to access your Amazon ECS resources.
- **AWS Command Line Interface (AWS CLI)** — Provides commands for a broad set of AWS services, including Amazon ECS. It's supported on Windows, Mac, and Linux. For more information, see AWS Command Line Interface.
- **AWS SDKs** — Provides language-specific APIs and takes care of many of the connection details. These include calculating signatures, handling request retries, and error handling. For more information, see AWS SDKs.
- **AWS Copilot** — Provides an open-source tool for developers to build, release, and operate production ready containerized applications on Amazon ECS. For more information, see AWS Copilot on the GitHub website.
- **Amazon ECS CLI** — Provides a command line interface for you to run your applications on Amazon ECS and AWS Fargate using the Docker Compose file format. You can quickly provision resources, push and pull images using Amazon Elastic Container Registry, and monitor running applications on Amazon ECS or Fargate. You can also test containers that are running locally along with containers in the Cloud within the CLI. For more information, see Amazon ECS CLI on the GitHub website.
- **AWS CDK** — Provides an open-source software development framework that you can use to model and provision your cloud application resources using familiar programming languages. The AWS CDK provisions your resources in a safe, repeatable manner through AWS CloudFormation. For more information, see Getting started with Amazon ECS using the AWS CDK.

Amazon Elastic Container Service FAQs

General

Q: What is Amazon Elastic Container Service?

Amazon Elastic Container Service (ECS) is a highly scalable, high performance container management service that supports Docker containers and allows you to easily run applications on a managed cluster of Amazon Elastic Compute Cloud (Amazon EC2) instances. Amazon ECS eliminates the need for you to install, operate, and scale your own cluster management infrastructure. With simple API calls, you can launch and stop container-enabled applications, query the complete state of your cluster, and access many familiar features like security groups, Elastic Load Balancing, Amazon Elastic Block Store (EBS) volumes, and Identity Access Management (IAM). roles. You can use Amazon ECS to schedule container placement across your cluster based on your resource needs and availability requirements. You can also integrate your own scheduler or third-party schedulers to meet business or application specific requirements.

Q: Why should I use Amazon ECS?

Amazon ECS makes it easy to use containers as a building block for your applications by eliminating the need for you to install, operate, and scale your own cluster management infrastructure. Amazon ECS lets you schedule long-running applications, services, and batch processes using Docker containers. Amazon ECS maintains application availability and allows you to scale your containers up or down to meet your application's capacity requirements. Amazon ECS is integrated with familiar features like Elastic Load Balancing, EBS volumes, Amazon Virtual Private Cloud (VPC), and IAM. Simple APIs let you integrate and use your own schedulers or connect Amazon ECS into your existing software delivery process.

Q: What is the pricing for Amazon ECS?

There is no additional charge for Amazon ECS. You pay for AWS resources (e.g. Amazon EC2 instances or EBS volumes) you create to store and run your application. You only pay for what you use, as you use it; there are no minimum fees and no upfront commitments.

Q: How is Amazon ECS different from AWS Elastic Beanstalk?

AWS Elastic Beanstalk is an application management platform that helps customers easily deploy and scale web applications and services. It keeps the building block provisioning (e.g., EC2, Amazon RDS, Elastic Load Balancing, AWS Auto Scaling, and Amazon CloudWatch), application deployment, and health monitoring abstracted from the user so they can focus on writing code. You simply specify which container images to deploy, the CPU and memory requirements, the port mappings, and the container links.

Elastic Beanstalk will automatically handle all the details such as provisioning an Amazon ECS cluster, balancing load, auto-scaling, monitoring, and container placement across your cluster. Elastic Beanstalk is ideal if you want to leverage the benefits of containers withthe simplicity of deploying applications from development to production by uploading a container image. You can work with Amazon ECS directly if you want more fine-grained control for custom application architectures.

Q: How is Amazon ECS different from AWS Lambda?

Amazon ECS is a highly scalable Docker container management service that allows you to run and manage distributed applications that run in Docker containers. AWS Lambda is an event-driven task compute service that runs your code in response to “events” such as changes in data, website clicks, or messages from other AWS services without you having to manage any compute infrastructure.

Using Amazon ECS

Q: How do I get started using Amazon ECS?

Visit our Getting Started page for more information on how to start using Amazon ECS.

Q: Does Amazon ECS support any other container types?

No. Docker is the only container platform supported by Amazon ECS at this time.

Q: I want to launch containers. Why do I have to launch tasks?

Docker encourages you to split your applications up into their individual components, and Amazon ECSis optimized for this pattern. Tasks allow you to define a set of containers you would like to place together (or part of the same placement decision), their properties, and how they may be linked. Tasks include all the information Amazon ECS needs to make the placement decision. To launch a single container, your task Definition should only include one container definition.

Q: Does Amazon ECS support applications and services?

Yes. The Amazon ECS Service scheduler can manage long-running applications and services. The service scheduler helps you maintain application availability and allows you to scale your containers up or down to meet your application's capacity requirements. The service scheduler allows you to distribute traffic across your containers using Elastic Load Balancing (ELB). Amazon ECS will automatically register and deregister your containers from the associated load balancer.

The service scheduler will also automatically recover containers that become unhealthy (i.e. fail ELB health checks) or stop running to ensure you have the desired number of healthy containers supporting your application.

You can scale your application up and down by changing the number of containers you want the service to run. You can update your application by changing its definition or using a new image. The scheduler will automatically start new containers using the new definition and stop containers running the previous version (waiting for the ELB connections to drain if ELB is used).

Q: Does Amazon ECS support dynamic port mapping?

Yes. It is possible to associate a service on Amazon ECS to an Application Load Balancer (ALB) for the ELB service. The ALB supports a target group containing a set of instance ports. You can specify a dynamic port in the ECS task definition which gives the container an unused port when it is scheduled on the EC2 instance. The ECS scheduler will automatically add the task to the Application Load Balancer’s target group using this port.

Q: Does Amazon ECS support batch jobs?

Yes. You can use Amazon ECS Run task to run one or more tasks once. Run task starts the task on an instance that meets the task’s requirements including CPU, memory, and ports.

Q: Can I use my own scheduler with Amazon ECS?

ECS provides Blox, a collection of open-source projects for container management and orchestration. Blox makes it easy to consume events from Amazon ECS, store the cluster state locally, and query the local data store through APIs. Blox also includes a daemon scheduler that

can be used as a reference for how to use the cluster state server. See the Blox GitHub page to learn more.

Q: Can I use my own Amazon Machine Image (AMI)?

Yes. You can use any AMI that meets the Amazon ECS AMI specification. We recommend starting from the Amazon ECS-enabled Amazon Linux AMI. Partner AMIs compatible with Amazon ECS are also available. You can review the Amazon ECS AMI specification in the documentation.

Q: How can I configure my container instances to pull from Amazon Elastic Container Registry?

Amazon ECR is integrated with Amazon ECS allowing you to easily store, run, and manage container images for applications running on Amazon ECS. All you need to do is specify the Amazon ECR repository in your task Definition and attach the AmazonEC2ContainerServiceforEC2Role to your instances. Then Amazon ECS will retrieve the appropriate images for your applications.

Q: How does AWS Fargate work with Amazon ECS?

With Fargate, the concept of server provisioning, cluster management, and orchestration completely goes away. Amazon ECS uses containers provisioned by Fargate to automatically scale, load balance, and manage scheduling of your containers for availability, providing an easier way to build and operate containerized applications.

Q: How should I choose between using AWS Fargate with Amazon ECS or just using ECS?

Amazon ECS supports Fargate technology and customers canchoose AWS Fargate to launch their containers without having to provision or manage Amazon EC2 instances. AWS Fargate is the easiest way to launch and run containers on AWS. Customersr equiring greater control of their EC2 instances (to support compliance and governance requirements or broader customization options) can choose to use ECS without Fargate to launch EC2 instances.

Security and Compliance

Q: How does Amazon ECS isolate containers belonging to different customers?

Amazon ECS schedules containers for execution on customer-controlled Amazon EC2 instances or with AWS Fargate and builds on the same isolation controls and compliance settings available for EC2 customers. Your compute instances are located in a Virtual Private Cloud (VPC) with an IP range that you specify. You decide which instances are exposed to the Internet and which remain private.

- Your EC2 instances use an IAM role to access the ECS service.
- Your ECS tasks use an IAM role to access services and resources.
- Security Groups and networks ACLs allow you to control inbound and outbound network access to and from your instances.
- You can connect your existing IT infrastructure to resources in your VPC using industry-standard encrypted IPsec VPN connections.
- You can provision your EC2 resources as Dedicated Instances. Dedicated Instances are Amazon EC2 Instances that run on hardware dedicated to a single customer for additional isolation.

Q: Can I apply additional security configuration and isolation frameworks to my container instances?

Yes. As an Amazon EC2 customer, you have root access to the operating system (OS) of your container instances. You can take ownership of the OS security settings, as well as configure additional software components for security capabilities such as monitoring, patch management, log management, and host intrusion detection.

Q: Can I operate container instances with different security settings or segregate different tasks across different environments?

Yes. You can configure your different container instances using the tooling of your choice. Amazon ECS allows you to control the placement of tasks in different container instances through the construct of clusters and targeted launches.

Q: Does Amazon ECS support retrieving Docker images from a private or internal source?

Yes. Customers can configure their container instances to access a private Docker image registry within a VPC or a registry that's accessible outside a VPC such as Amazon Elastic Container Registry (ECR).

Q: How do I configure IAM roles for ECS tasks?

You first need to create an IAM role for your task, using the 'Amazon EC2 Container Service Task Role' service role and attaching a policy with the required permissions. When you create a new task definition or a task definition revision, you can then specify a role by selecting it from the 'Task Role' drop-down or using the 'taskRoleArn' filed in the JSON format.

Q: With which compliance programs does Amazon ECS conform?

Amazon ECS meets the standards for PCI DSS Level 1, ISO 9001, ISO 27001, ISO 27017, ISO 27018, SOC 1, SOC 2, SOC 3, and HIPAA eligibility.

For more information, visit our compliance pages.

Q: Can I use Amazon ECS for Protected Health Information (PHI) and other HIPAA-regulated workloads?

Yes. Amazon ECS is HIPAA-eligible. If you have an executed Business Associate Addendum (BAA) with AWS, you can use Amazon ECS to process encrypted Protected Health Information (PHI) using Docker containers deployed onto the AWS Fargate launch-type or Amazon EC2 compute instances.

For more information, please visit our page on HIPAA compliance. If you plan to process, store, or transmit PHI and do not have an executed BAA from AWS, please contact us for more information.

Q: Can I use Amazon ECS for US Government-regulated workloads or processing sensitive Controlled Unclassified Information (CUI)?

Yes. By using the AWS GovCloud (US) region, containers and clusters managed by Amazon ECS can meet the requirements to sensitive data and regulated workloads with your containers.

For more information, visit our page on AWS GovCloud.

Amazon Elastic Kubernetes Service (Amazon EKS)

The most trusted way to start, run, and scale Kubernetes.

How it works

Amazon EKS is a managed Kubernetes service to run Kubernetes in the AWS cloud and on-premises data centers. In the cloud, Amazon EKS automatically manages the availability and scalability of the Kubernetes control plane nodes responsible for scheduling containers, managing application availability, storing cluster data, and other key tasks. With Amazon EKS, you can take advantage of all the performance, scale, reliability, and availability of AWS infrastructure, as well as integrations with AWS networking and security services. On-premises, EKS provides a consistent, fully-supported Kubernetes solution with integrated tooling and simple deployment to AWS Outposts, virtual machines, or bare metal servers.

Run Amazon EKS in the cloud

Run Amazon EKS on AWS Outposts

Run Amazon EKS Anywhere

Deploy applications with your own tools

Use cases

Deploy across hybrid environments

Manage your Kubernetes clusters and applications in hybrid environments and run Kubernetes in your data centers.

Model machine learning (ML) workflows
Efficiently run distributed training jobs using the latest Amazon Elastic Compute Cloud (EC2) GPU-powered instances, including Inferentia, and deploy training and inferences using Kubeflow.

Build and run web applications
Create applications that automatically scale up and down and run in a highly available configuration across multiple Availability Zones (AZs) with out-of-the-box networking and security integrations.

What is Amazon EKS?

Amazon Elastic Kubernetes Service (Amazon EKS) is a managed service that you can use to run Kubernetes on AWS without needing to install, operate, and maintain your own Kubernetes control plane or nodes. Kubernetes is an open-source system for automating the deployment, scaling, and management of containerized applications. Amazon EKS:

- Runs and scales the Kubernetes control plane across multiple AWS Availability Zones to ensure high availability.
- Automatically scales control plane instances based on load, detects and replaces unhealthy control plane instances, and it provides automated version updates and patching for them.
- Is integrated with many AWS services to provide scalability and security for your applications, including the following capabilities:
    - Amazon ECR for container images
    - Elastic Load Balancing for load distribution
    - IAM for authentication
    - Amazon VPC for isolation
- Runs up-to-date versions of the open-source Kubernetes software, so you can use all of the existing plugins and tooling from the Kubernetes community. Applications that are running on Amazon EKS are fully compatible with applications running on any standard Kubernetes environment, no matter whether they're running in on-premises data centers or public clouds. This means that you can easily migrate any standard Kubernetes application to Amazon EKS without any code modification.

Amazon EKS control plane architecture

Amazon EKS runs a single tenant Kubernetes control plane for each cluster. The control plane infrastructure isn't shared across clusters or AWS accounts. The control plane consists of at least two API server instances and three etcd instances that run across three Availability Zones within an AWS Region. Amazon EKS:

- Actively monitors the load on control plane instances and automatically scales them to ensure high performance.
- Automatically detects and replaces unhealthy control plane instances, restarting them across the Availability Zones within the AWS Region as needed.
- Leverages the architecture of AWS Regions in order to maintain high availability. Because of this, Amazon EKS is able to offer an SLA for API server endpoint availability.

Amazon EKS uses Amazon VPC network policies to restrict traffic between control plane components to within a single cluster. Control plane components for a cluster can't view or receive communication from other clusters or other AWS accounts, except as authorized with Kubernetes RBAC policies. This secure and highly available configuration makes Amazon EKS reliable and recommended for production workloads.

How does Amazon EKS work?

Amazon Elastic Kubernetes Service

Amazon Elastic Kubernetes Service (Amazon EKS) is a fully-managed, certified Kubernetes conformant service that simplifies the process of building, securing, operating, and maintaining Kubernetes clusters on AWS. Amazon EKS integrates with core AWS services such as CloudWatch, Auto Scaling Groups, and IAM to provide a seamless experience for monitoring, scaling and load balancing your containerized applications.

Amazon EKS also integrates with AWS App Mesh and provides a Kubernetes-native experience to consume service mesh features and bring rich observability, traffic controls and security features to applications. Amazon EKS provides a scalable, highly-available control plane for Kubernetes workloads. When running applications on Amazon EKS, as with Amazon ECS, you can choose to provide the underlying compute power for your containers with EC2 instances or with AWS Fargate.

Table: Amazon EKS deployment features

Capability	**Description**
Provision	Amazon EKS provisions certain resources to support containerized applications: • Load Balancers, if needed. • Compute Resources ("workers"). Amazon EKS supports Windows and Linux. • Application Container Instances ("pods"). Refer to Getting Started with Amazon EKS for more details on Amazon EKS cluster provisioning.
Configure	Amazon EKS supports customization of the compute resources ("workers") if using EC2 instances to supply compute power. EKS also supports customization of the runtime conditions of the application containers ("pods"). Refer to Worker Nodes and Fargate Pod Configuration documentation for more details.
Deploy	Amazon EKS supports the same deployment strategies as Kubernetes, see Writing a Kubernetes Deployment Spec -> Strategy for more details.
Scale	Amazon EKS scales workers with Kubernetes Cluster Autoscaler, and pods with Kubernetes Horizontal Pod Autoscaler and Kubernetes Vertical Pod Autoscaler.
Monitor	The Amazon EKS control plane logs provide audit and diagnostic information directly to CloudWatch Logs. The

Capability	Description
	Amazon EKS control plane also integrates with AWS CloudTrail to record actions taken in Amazon EKS. Refer to Logging and Monitoring Amazon EKS for more details.

Amazon EKS allows organizations to leverage open source Kubernetes tools and plugins, and can be a good choice for organizations migrating to AWS with existing Kubernetes environments. The following diagram illustrates Amazon EKS being used to manage a general containerized application.

Figure: Amazon EKS use case

Amazon Elastic Kubernetes Service FAQs

General

Q: What is Amazon Elastic Kubernetes Service (Amazon EKS)?

A: Amazon EKS is a managed service that makes it easy for you to run Kubernetes on AWS without installing and operating your own Kubernetes control plane or worker nodes.

Q: What is Kubernetes?

A: Kubernetes is an open-source container orchestration system allowing you to deploy and manage containerized applications at scale. Kubernetes arranges containers into logical

groupings for management and discoverability, then launches them onto clusters of Amazon Elastic Compute Cloud (Amazon EC2) instances. Using Kubernetes, you can run containerized applications including microservices, batch processing workers, and platforms as a service (PaaS) using the same toolset on premises and in the cloud.

Q: Why should I use Amazon EKS?

A: Amazon EKS provisions and scales the Kubernetes control plane, including the application programming interface (API) servers and backend persistence layer, across multiple AWS Availability Zones (AZs) for high availability and fault tolerance. Amazon EKS automatically detects and replaces unhealthy control plane nodes and patches the control plane. You can run EKS using AWS Fargate, which provides serverless compute for containers. Fargate removes the need to provision and manage servers, lets you specify and pay for resources per application, and improves security through application isolation by design.

Amazon EKS is integrated with many AWS services to provide scalability and security for your applications. These services include Elastic Load Balancing for load distribution, AWS Identity and Access Management (IAM) for authentication, Amazon Virtual Private Cloud (VPC) for isolation, and AWS CloudTrail for logging.

Q: How does Amazon EKS work?

A: Amazon EKS works by provisioning (starting) and managing the Kubernetes control plane and worker nodes for you. At a high level, Kubernetes consists of two major components: a cluster of 'worker nodes' running your containers, and the control plane managing when and where containers are started on your cluster while monitoring their status.

Without Amazon EKS, you have to run both the Kubernetes control plane and the cluster of worker nodes yourself. With Amazon EKS, you provision your worker nodes using a single command in the EKS console, command-line interface (CLI), or API. AWS handles provisioning, scaling, and managing the Kubernetes control plane in a highly available and secure configuration. This removes a significant operational burden and allows you to focus on building applications instead of managing AWS infrastructure.

Q: Which operating systems does Amazon EKS support?

A: Amazon EKS supports Kubernetes-compatible Linux x86, ARM, and Windows Server operating system distributions. Amazon EKS provides optimized AMIs for Amazon Linux 2 and Windows Server 2019. EKS- optimized AMIs for other Linux distributions, such as Ubuntu, are available from their respective vendors.

Q: I have a feature request, who do I tell?

A: Please let us know what we can add or do better by opening a feature request on the AWS Container Services Public Roadmap

Integrations

Q: Does Amazon EKS work with my existing Kubernetes applications and tools?

A: Amazon EKS runs the open-source Kubernetes software, so you can use all the existing plug-ins and tooling from the Kubernetes community. Applications running on Amazon EKS are fully compatible with applications running on any standard Kubernetes environment, whether running in on-premises data centers or public clouds. This means that you can easily migrate any standard Kubernetes application to Amazon EKS without any code modifications.

Q: Does Amazon EKS work with AWS Fargate?

A: Yes. You can run Kubernetes applications as serverless containers using AWS Fargate and Amazon EKS.

Q: What are Amazon EKS add-ons?

A: EKS Add-Ons let you enable and manage Kubernetes operational software, which provides capabilities like observability, scaling, networking, and AWS cloud resource integrations for your EKS clusters. At launch, EKS add-ons supports controlling the launch and version of the AWS VPC CNI plugin through the EKS API.

Q: Why should I use Amazon EKS add-ons?

A: Amazon EKS add-ons provides one-click installation and management of Kubernetes operational software. Go from cluster creation to running applications in a single command, while easily keeping the operational software required for your cluster up to date. This ensures your Kubernetes clusters are secure and stable and reduces the amount of work needed to start and manage production-ready Kubernetes clusters on AWS.

Kubernetes versions and updates

Q: Which Kubernetes versions does Amazon EKS support?

A: See the Amazon EKS documentation for currently supported Kubernetes versions. Amazon EKS will continue to add support for additional Kubernetes versions in the future.

Q: Can I update my Kubernetes cluster to a new version?

A: Yes. Amazon EKS performs managed, in-place cluster upgrades for both Kubernetes and Amazon EKS platform versions. This simplifies cluster operations and lets you take advantage of the latest Kubernetes features, as well as the updates to Amazon EKS configuration and security patches.

There are two types of updates you can apply to your Amazon EKS cluster: Kubernetes version updates and Amazon EKS platform version updates. As new Kubernetes versions are released and validated for use with Amazon EKS, we will support three stable Kubernetes versions as part of the update process at any given time.

Q: What is an EKS platform version?

A: Amazon EKS platform versions represent the capabilities of the cluster control plane, such as which Kubernetes API server flags are enabled, as well as the current Kubernetes patch version. Each Kubernetes minor version has one or more associated Amazon EKS platform versions. The platform versions for different Kubernetes minor versions are independent.

When a new Kubernetes minor version is available in Amazon EKS (for example, 1.13), the initial Amazon EKS platform version for that Kubernetes minor version starts at eks.1. However, Amazon EKS releases new platform versions periodically to enable new Kubernetes control plane settings and to provide security fixes.

Q: Why would I want manual control over Kubernetes version updates?

A: New versions of Kubernetes introduce significant change to the Kubernetes API, which can change application behavior. Manual control over Kubernetes cluster versioning lets you test applications against new versions of Kubernetes before upgrading production clusters. Amazon EKS offers the ability to choose when you introduce changes to your EKS cluster.

Q: How do I update my worker nodes?

A: AWS publishes EKS-optimized Amazon Machine Images (AMIs) that include the necessary worker node binaries (Docker and Kubelet). This AMI is updated regularly and includes the most up-to-date version of these components. You can update your EKS managed nodes to the latest versions of the EKS-optimized AMIs with a single command in the EKS console, API, or CLI.

If you are building your own custom AMIs to use for EKS worker nodes, AWS also publishes Packer scripts that document our build steps, allowing you to identify the binaries included in each version of the AMI.

AWS Fargate

Serverless compute for containers.

How it works

AWS Fargate is a serverless, pay-as-you-go compute engine that lets you focus on building applications without managing servers. AWS Fargate is compatible with both Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS).

Use cases

Web apps, APIs, and microservices
Build and deploy your applications, APIs, and microservices architectures with the speed and immutability of containers. Fargate removes the need to own, run, and manage the lifecycle of a compute infrastructure so that you can focus on what matters most: your applications.

Run and scale container workloads

Use Fargate with Amazon ECS or Amazon EKS to easily run and scale your containerized data processing workloads. Fargate also enables you to migrate and run your Amazon ECS Windows containers without refactoring or rearchitecting your legacy applications.

Support AI and ML training applications
Create an AI and ML development environment that is flexible and portable. With Fargate, achieve the scalability you need to boost server capacity without over-provisioning—to train, test, and deploy your machine learning (ML) models.

Optimize Costs
With AWS Fargate there are no upfront expenses, pay for only the resources used. Further optimize with Compute Savings Plans and Fargate Spot, then use Graviton2 powered Fargate for up to 40% price performance improvements.

AWS Fargate

This is about using Amazon EKS to run Kubernetes pods on AWS Fargate.

AWS Fargate is a technology that provides on-demand, right-sized compute capacity for containers. With AWS Fargate, you don't have to provision, configure, or scale groups of virtual machines on your own to run containers. You also don't need to choose server types, decide when to scale your node groups, or optimize cluster packing. You can control which pods start on Fargate and how they run with Fargate profiles. Fargate profiles are defined as part of your Amazon EKS cluster.

Amazon EKS integrates Kubernetes with AWS Fargate by using controllers that are built by AWS using the upstream, extensible model provided by Kubernetes. These controllers run as part of the Amazon EKS managed Kubernetes control plane and are responsible for scheduling native Kubernetes pods onto Fargate. The Fargate controllers include a new scheduler that runs alongside the default Kubernetes scheduler in addition to several mutating and validating admission controllers. When you start a pod that meets the criteria for running on Fargate, the Fargate controllers that are running in the cluster recognize, update, and schedule the pod onto Fargate.

This topic describes the different components of pods that run on Fargate, and calls out special considerations for using Fargate with Amazon EKS.

AWS Fargate considerations

Here are some things to consider about using Fargate on Amazon EKS.

- AWS Fargate with Amazon EKS is available in all Amazon EKS Regions except AWS GovCloud (US-East) and AWS GovCloud (US-West).
- Each pod that runs on Fargate has its own isolation boundary. They don't share the underlying kernel, CPU resources, memory resources, or elastic network interface with another pod.
- Network Load Balancers and Application Load Balancers (ALBs) can be used with Fargate with IP targets only. For more information, see Create a network load balancer and Application load balancing on Amazon EKS.
- Fargate exposed services only run on target type IP mode, and not on node IP mode. The recommended way to check the connectivity from a service running on a managed node and a service running on Fargate is to connect via service name.
- Pods must match a Fargate profile at the time that they're scheduled to run on Fargate. Pods that don't match a Fargate profile might be stuck as `Pending`. If a matching Fargate profile exists, you can delete pending pods that you have created to reschedule them onto Fargate.
- Daemonsets aren't supported on Fargate. If your application requires a daemon, reconfigure that daemon to run as a sidecar container in your pods.
- Privileged containers aren't supported on Fargate.
- Pods running on Fargate can't specify `HostPort` or `HostNetwork` in the pod manifest.
- The default `nofile` and `nproc` soft limit is 1024 and the hard limit is 65535 for Fargate pods.
- GPUs aren't currently available on Fargate.
- Pods that run on Fargate are only supported on private subnets (with NAT gateway access to AWS services, but not a direct route to an Internet Gateway), so your cluster's VPC must have private subnets available. For clusters without outbound internet access, see Private cluster requirements.
- You can use the Vertical Pod Autoscaler to set the initial correct size of CPU and memory for your Fargate pods, and then use the Horizontal Pod Autoscaler to scale those pods. If you want the Vertical Pod Autoscaler to automatically re-deploy pods to Fargate with larger CPU and memory combinations, set the mode for the Vertical Pod Autoscaler to either `Auto` or `Recreate` to ensure correct functionality. For more information, see the Vertical Pod Autoscaler documentation on GitHub.
- DNS resolution and DNS hostnames must be enabled for your VPC. For more information, see Viewing and updating DNS support for your VPC.

- Amazon EKS Fargate adds defense-in-depth for Kubernetes applications by isolating each Pod within a Virtual Machine (VM). This VM boundary prevents access to host-based resources used by other Pods in the event of a container escape, which is a common method of attacking containerized applications and gain access to resources outside of the container.

  Using Amazon EKS doesn't change your responsibilities under the shared responsibility model. You should carefully consider the configuration of cluster security and governance controls. The safest way to isolate an application is always to run it in a separate cluster.
- Fargate profiles support specifying subnets from VPC secondary CIDR blocks. You might want to specify a secondary CIDR block. This is because there's a limited number of IP addresses available in a subnet. As a result, there's also a limited number of pods that can be created in the cluster. By using different subnets for pods, you can increase the number of available IP addresses. For more information, see Adding IPv4 CIDR blocks to a VPC.
- The Amazon EC2 instance metadata service (IMDS) isn't available to pods that are deployed to Fargate nodes. If you have pods that are deployed to Fargate that need IAM credentials, assign them to your pods using IAM roles for service accounts. If your pods need access to other information available through IMDS, then you must hard code this information into your pod spec. This includes the AWS Region or Availability Zone that a pod is deployed to.
- You can't deploy Fargate pods to AWS Outposts, AWS Wavelength or AWS Local Zones.
- Amazon EKS must periodically patch AWS Fargate pods to keep them secure. We attempt the updates in a way that reduces impact, but there are times when pods must be deleted if they aren’t successfully evicted. There are some actions you can take to minimize disruption. For more information, see Fargate pod patching.
- The Amazon VPC CNI plugin for Amazon EKS is installed on Fargate nodes. You can't use Alternate compatible CNI plugins with Fargate nodes.

AWS Fargate FAQs

General

Overview

Q: What is AWS Fargate?

AWS Fargate is a serverless compute engine for containers that works with both Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS). AWS Fargate makes it easy to focus on building your applications. Fargate eliminates the need to provision and manage servers, lets you specify and pay for resources per application, and improves security through application isolation by design.

Q: Why should I use AWS Fargate?

AWS Fargate enables you to focus on your applications. You define your application content, networking, storage, and scaling requirements. There is no provisioning, patching, cluster capacity management, or infrastructure management required.

Q: What use cases does AWS Fargate support?

AWS Fargate supports all of the common container use cases including microservices architecture applications, batch processing, machine learning applications, and migrating on-premises applications to the cloud.

Q: What is the pricing of AWS Fargate?

With AWS Fargate, you pay only for the amount of vCPU, memory, and storage resources consumed by your containerized applications.

vCPU and memory resources are calculated from the time your container images are pulled until the Amazon ECS task or EKS pod terminates, rounded up to the nearest second. A minimum charge of 1 minute applies. 20 GB of ephemeral storage is available for all Fargate Tasks and Pods by default—you only pay for any additional storage that you configure.

Fargate supports Spot and Compute Savings Plan pricing options just like with Amazon EC2 instances. You can find additional details on the pricing page.

Q: How does AWS Fargate work with Amazon ECS and Amazon EKS?

Amazon Elastic Container Service (ECS) is a highly scalable, high performance container management service that supports Docker containers and Amazon Elastic Kubernetes Service (EKS) is a fully managed Kubernetes service. Both ECS and EKS use containers provisioned by Fargate to automatically scale, load balance, and optimize container availability through managed scheduling, providing an easier way to build and operate containerized applications.

Q: How should I choose when to use AWS Fargate?

Choose AWS Fargate for its isolation model and security. You should also select Fargate if you want to launch containers without having to provision or manage EC2 instances. If you require greater control of your EC2 instances or broader customization options, then use ECS or EKS without Fargate. Use EC2 for GPU workloads, which are not supported on Fargate today.

Q: Can I run my Arm-based applications on AWS Fargate?

Yes. AWS Fargate allows you to run your Arm-based applications by using Arm-compatible container images or multi-architecture container images in Amazon Elastic Container Registry (Amazon ECR). You can simply specify the CPU Architecture as ARM64 in your Amazon ECS Task Definition to target AWS Fargate powered by Arm-based AWS Graviton2 Processors.

Q: Why should I use AWS Fargate powered by Graviton2 processors?

AWS Graviton2 processors are custom built by Amazon Web Services using 64-bit Arm Neoverse cores to deliver the best price performance for your cloud workloads. AWS Fargate powered by AWS Graviton2 processors delivers up to 40% improved price/performance at 20% lower cost over comparable Intel x86-based Fargate for a variety of workloads such as application servers, web services, high-performance computing, and media processing. You get the same serverless benefits of AWS Fargate while optimizing performance and cost for running your containerized workloads.

Q: Can I run my Amazon ECS Windows containers on AWS Fargate?

Yes. AWS Fargate offers a serverless approach for running your Windows containers. It removes the need to provision and manage servers and lets you specify and pay for resources per application. Fargate provides task-level isolation and handles the necessary patching and updating to help provide a secure compute environment.

Can I use my existing Microsoft Windows License with AWS Fargate?

Since Fargate is a serverless compute engine, customers do not need to manage the underlying compute instances running in Fargate. Therefore, Fargate will manage the Windows OS licenses for you and the cost of doing so is built into the Fargate pricing.

Q: Which Windows Server versions are supported with AWS Fargate?

Fargate supports Windows Server 2019 Long-Term Servicing Channel (LTSC) release on Fargate Windows Platform Version 1.0.0 or later.

Fargate vCPU-based Service Quotas

Q: What is changing?

AWS Fargate is transitioning service quotas from the current Amazon ECS task and Amazon EKS pod count based concurrent quotas to vCPU-based quotas for On-Demand and Spot usage. The new vCPU-based quotas will replace the existing tasks and pods count-based quotas. With vCPU-based quotas, we are simplifying the service quotas experience as your accounts' usage against these quotas is now measured using vCPUs, the primary resource provisioned by your applications.

Q: How do vCPU-based quotas benefit me?

With vCPU-based quotas, Fargate uses the number of vCPUs provisioned by a task or pod as the quota unit. You can now more easily forecast, manage, and request quotas based on the vCPUs provisioned by your applications. Currently, you manage quotas on Fargate using task and pod count, undifferentiated by vCPUs your applications need. For example, an account with service quota of 250 tasks can launch up to 250 0.25 vCPU or 250 4vCPU tasks. With the new vCPU-based service quotas, a quota of 1,000 vCPUs allows you to concurrently launch up to 4,000 0.25 vCPU or up to 250 4 vCPUs tasks. With vCPU-based quotas, On-Demand tasks or pods and Spot tasks usage against the vCPU quotas are measured in terms of the number of vCPUs configured for your running tasks or pods.

Q: When can I start using vCPU-based quotas?

Fargate provides you the option to opt-in to vCPU quotas starting September 8, 2022. By opting in, you give yourself valuable time to make modifications to your limit management tools and minimize the risk of impact to your systems. Starting October 10, 2022, Fargate will automatically begin switching over accounts to use the new vCPU quotas in a phased manner.

You will still have the option to opt-out of vCPU quotas until end of October 2022. Starting November 1, 2022, Fargate will switch all remaining accounts to vCPU quotas, regardless of opt out status, and task and pod count-based quotas will no longer be supported.

Q: How do I opt in and out of vCPU-based quotas?

If you use Amazon ECS with Fargate, you can easily and quickly opt in and opt out of vCPU-based quotas by changing your ECS account setting using the CLI as documented here. If you use Amazon EKS with Fargate, you can file a request with a request in the AWS Support Center console. You opt in or out of the vCPU-based quotas for each of your AWS accounts. Once your request to opt-in to vCPU quotas is processed, your task and pod count's applied limit will be marked zero on the Service Quotas Console, only your vCPU-based quotas will be displayed. You should now start managing your Service Quotas using vCPU-based quotas.

Q: Are vCPU-based quotas regional?

Yes. Like task and pod count-based quotas, vCPU-based quotas for an AWS account are on a per-region basis.

Q: How can I view my current task and pod count-based quotas and new vCPU based-quotas?

You can find your current task and pod count quotas on the Service Quotas Console and by using the Service Quota API. Starting September 8, 2022, you will be able to view both current task and pod count-based quotas and new vCPU-based quotas on Service Quotas Console.

Q: Will I be able to view actual usage against these new quotas?

Yes. You can track and inspect your vCPU usage against these quotas in Service Quotas. Service Quotas also enables customers to use CloudWatch for configuring new alarms to warn customers of approaching their vCPU-based quotas.

Q: Will the migration to vCPU quotas affect running tasks and pods?

No, opting in to and out of vCPU-based quotas during this transition period will not affect any running tasks or pods.

Q: What if I run into issues with vCPU-based quotas?

If you run into issues with vCPU-based quotas, you can opt back out of vCPU quotas and remediate your systems. However, your account will automatically be transitioned back to vCPU quotas beginning November 2022. It is important for you to test your systems with vCPU quotas before November 2022.

Q: What are the changes I should be aware of with the migration to vCPU-based quotas?

If you integrate with the current quotas' limit exceeded error, we recommend testing your systems before the transition period ends. For instance, with vCPU quotas, Fargate will return a new error message when exceeding your new vCPU quotas. This new error message for On-Demand quotas is: "You've reached the limit on the number of vCPUs you can run concurrently" and for Spot quotas is: "You've reached the limit on the number of vCPUs you can run as spot tasks concurrently". We recommend reviewing your system for changes if you have integration with Service Quotas, Service Quota APIs, or templates. With Amazon CloudWatch metrics integration in Service Quotas, you can monitor Fargate usage against the new vCPU-based quotas by configuring new alarms to warn about approaching quotas.

Q: How can I request a quota increase for vCPU-based quotas?

You continue to request limit increases using the Service Quotas console. To request a limit increase, select “Request Limit Increase” in Service Quota console and state your requirement in vCPUs. If you continue to use task and pod count-based quotas, you can request a limit increase against the existing task and pod count quotas.

Q: Can I still launch the same number of tasks and pods?

Yes, vCPU-based quotas allow you to launch at least the same number of task or pods as you do today with task and pod count-based quotas. If your account already has an approved quota increase, you will continue to be able to launch at least the same number of tasks or pods. Like today, new AWS accounts may start with lower quotas than the default, and these quotas can increase over time. Read our documentation for more details.

Q: What happens to my quotas if I opt out of vCPU quotas during the transition period?

If you decide to opt out during the transition period, your quotas will revert to task and pod count-based limit values you had before you opted in. Note that Fargate will however automatically switch your accounts to vCPU quotas beginning November 1, 2022.

Q: What will happen if I take no action?

Your accounts will automatically begin to use vCPU-based quotas starting October 10, 2022 as we migrate your accounts to vCPU-based quotas in a phased manner. By testing and opting in earlier, you give yourself valuable time to make modifications to your limit management tools and minimize the risk of impact to your systems.

Q: Will these new quotas have an impact on my monthly bill?

No. Fargate’s pricing remains the same regardless of task and pod count-based, or vCPU-based quotas.

Security and Compliance

Q: With which compliance programs does AWS Fargate conform?

AWS Fargate meets the standards for PCI DSS Level 1, ISO 9001, ISO 27001, ISO 27017, ISO 27018, SOC 1, SOC 2, SOC 3, and HIPAA eligibility.

For more information, visit our compliance pages.

Q: Can I use AWS Fargate for Protected Health Information (PHI) and other HIPAA regulated workloads?

Yes. AWS Fargate is HIPAA-eligible. If you have an executed Business Associate Addendum (BAA) with AWS, you can process encrypted Protected Health Information (PHI) using Docker containers deployed onto Fargate.

For more information, please visit our page on HIPAA compliance. If you plan to process, store, or transmit PHI and do not have an executed BAA from AWS, please contact us for more information.

Q: Can I use AWS Fargate for US Government-regulated workloads or processing sensitive Controlled Unclassified Information (CUI)?

Yes. Fargate is available in AWS GovCloud (US) Regions. AWS GovCloud (US) is Amazon's isolated cloud infrastructure and services designed to address specific regulatory and compliance requirements of US Government agencies, as well as contractors, educational institutions, and other US customers that run sensitive workloads in the cloud. For a full list of AWS Regions where Fargate is available, please visit our Region table.

For more information, visit our page on AWS GovCloud.

Q: Which Windows Server versions are supported with AWS Fargate?

Fargate supports Windows Server 2019 Long-Term Servicing Channel (LTSC) release on Fargate Windows Platform Version 1.0.0 or later.

AWS App Runner

Deploy containerized web applications and APIs at scale.

How it works

AWS App Runner is a fully managed container application service that lets you build, deploy, and run containerized web applications and API services without prior infrastructure or container experience.

Use cases

Build frontend and backend applications
Use App Runner to build and run API services, backend web services, websites, and more.

Run thousands of microservices simultaneously
Flexibly scale each component of your application, decreasing risks to application resiliency and fostering increased agility and innovation.

Improve your time to market
Use AWS best practices and technologies to deploy and run containerized web applications at scale, drastically decreasing time to market.

What is AWS App Runner?

AWS App Runner is an AWS service that provides a fast, simple, and cost-effective way to deploy from source code or a container image directly to a scalable and secure web application in the AWS Cloud. You don't need to learn new technologies, decide which compute service to use, or know how to provision and configure AWS resources.

App Runner connects directly to your code or image repository. It provides an automatic integration and delivery pipeline with fully managed operations, high performance, scalability, and security.

Who is App Runner for?

If you're a *developer*, you can use App Runner to simplify the process of deploying a new version of your code or image repository.

For *operations teams*, App Runner enables automatic deployments each time a commit is pushed to the code repository or a new container image version is pushed to the image repository.

Accessing App Runner

You can define and configure your App Runner service deployments using any one of the following interfaces:

- **App Runner console** – Provides a web interface for managing your App Runner services.
- **App Runner API** – Provides a RESTful API for performing App Runner actions. For more information, see AWS App Runner API Reference.
- **AWS Command Line Interface (AWS CLI)** – Provides commands for a broad set of AWS services, including Amazon VPC, and is supported on Windows, macOS, and Linux. For more information, see AWS Command Line Interface.
- **AWS SDKs** – Provides language-specific APIs and takes care of many of the connection details, such as calculating signatures, handling request retries, and error handling. For more information, see AWS SDKs.

AWS App Runner FAQs

General

Q: Why should I use AWS App Runner?
A: AWS App Runner is the easiest way to run your web application (including API services, backend web services, and websites) on AWS. With App Runner, there is no infrastructure or container orchestration required. You can go from an existing container image, container registry, source code repository, or existing CI/CD workflow to a fully running containerized web application on AWS in minutes.

App Runner seamlessly integrates with your development workflow to provide the right level of automation to deploy your code or container image without the need for you to understand, provision, scale, or manage any AWS compute, networking, and routing resources. You get the simplicity to run thousands of applications that automatically scale based on your traffic needs. Your applications also run on infrastructure that is maintained and operated by AWS, providing security and compliance best practices such as automated security patches and encryption.

Q: How do I get started with AWS App Runner?
A: You can deploy an AWS App Runner application on AWS by going to the App Runner console or using the AWS CLI and creating an App Runner application. When creating the App Runner application, you can provide a container image, connect your container registry, or connect your source code repository, and enter any optional build and start commands. App Runner will automatically create the service with a corresponding secure URL.

Q: Can I still use AWS App Runner if I don't use containers?
A: Yes. AWS App Runner supports automatically building a container image on curated App Runner platforms that contain supported runtimes and frameworks. When you associate your existing source code repository and optionally provide App Runner with your runtime build and start commands, App Runner automatically containerizes your web application and provides a running web application.

Q: Can I migrate from AWS App Runner to Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), or another AWS service if I need more flexibility?
A: Yes. You can use the same container image deployed to App Runner on Amazon ECS or Amazon EKS. This gives you the flexibility to move to different services as your needs change.

You can deploy your code or containers directly to the new AWS service you choose using the tools and onboarding workflow provided by that service.

Q: What use cases does App Runner with Amazon Virtual Private Cloud (Amazon VPC) support enable?
A: Amazon VPC support on App Runner enables your service to access database engines like Amazon Aurora, MySQL, PostgreSQL, and MariaDB in Amazon Relational Database Service (RDS) instances that run in a VPC. This support also enables your service to access Redis caches hosted on Amazon ElastiCache nodes in a VPC; push messages to a message broker such as Amazon Managed Streaming for Apache Kafka or Amazon MQ in a VPC; and talk to a backend service running on AWS Fargate, supported in Amazon Elastic Container Service and Amazon Elastic Kubernetes Service, or Amazon Elastic Compute Cloud in a VPC. Finally, you can also enable your service to talk to an on premise database that can be connected via an AWS Direct Connect network connection set up in a VPC.

Developers

Q: What type of applications can I run on AWS App Runner?
A: AWS App Runner supports full stack development, including both frontend and backend web applications that use HTTP and HTTPS protocols. These applications include API services, backend web services, and websites. App Runner supports container images as well as runtimes and web frameworks including Node.js and Python.

Q: What type of deployment options do I have with AWS App Runner?
A: AWS App Runner supports multiple deployment options including the ability to immediately deploy a container image using the App Runner console or AWS CLI. If you have an existing CI/CD workflow that uses AWS CodePipeline, Jenkins, Travis CI, CircleCI, or another CI/CD toolchain, you can easily add App Runner as your deployment target using the App Runner API or AWS CLI. If you want App Runner to automatically provide continuous deployment for you, you can easily connect to your existing container registry or source code repository and App Runner will automatically provide a continuous deployment pipeline for you.

With App Runner, you can have a unique application for each of your container images or source code branches with separate environment variables, deployment types (such as a development application or production application), and build and start commands. Once deployed, you get all the benefits of running your web application on App Runner including default security, automatic scaling, and monitoring.

Q: How does AWS App Runner scale applications to varying demand?
A: AWS App Runner monitors the number of concurrent requests sent to your application and automatically adds additional instances based on request volume. If your application receives no incoming requests, App Runner will scale the containers down to a provisioned instance, a CPU-throttled instance ready to serve incoming requests within milliseconds. You can also optionally configure the number of concurrent requests sent to an instance in the autoscaling settings of your application.

Q: Can I use my own domain name with web applications I'm running on AWS App Runner?
A: Simply add your custom domain to your AWS App Runner application using the App Runner console or AWS CLI. After you add your custom domain name, App Runner provides instructions to help you update your DNS records with your DNS provider. App Runner supports custom sub domains (www.example.com), custom root domains (example.com), and wildcard domain (*.example.com).

Q: What if my application needs access to storage, database, or caching services?
A: AWS App Runner does not restrict your ability to connect an application to other storage, database, or application integration services. Customers can easily configure their applications

or containers with the necessary code and connection instructions, and their application is able to communicate with these external services over the network securely.

Q: How can I view the logs of my application running on AWS App Runner?
A: AWS App Runner fully integrates with Amazon CloudWatch Logs and provides you with runtime logs and deployment logs aggregated from the output streams of all of the system components, web frameworks, runtimes, build and deployment commands, and application/web servers. App Runner aggregates these into a single comprehensive channel available through the App Runner console, CloudWatch console, and AWS CLI.

Q: Can I run web applications on AWS App Runner and persistent applications on AWS Fargate with an orchestrator?
A: Yes. If you need to run other applications such as content management systems that need a persistent file system or machine learning jobs, you can use AWS Fargate with an orchestrator that is capable of multiple resources such as graphics acceleration or persistent volumes. If you are using the Copilot CLI, you can continue to use the tool as it has support for both App Runner and ECS/Fargate. You can also use Amazon CloudWatch as your single pane of glass to monitor applications running across App Runner, Amazon ECS tasks running on Fargate, and Amazon EKS pods running on Fargate.

Q: Is AWS App Runner supported by Copilot?
A: AWS Copilot is a command line interface (CLI) that enables customers to quickly launch and easily manage containerized applications on AWS. You can use Copilot to quickly get up and running on AWS App Runner. You can also use Copilot as your standard CLI to operate across AppRunner and/or ECS/Fargate.

Q: How do I enable VPC support for App Runner applications?
A: While creating an App Runner application, you can pass VPC ID, subnets, and security groups. This information will be used to create network interfaces that allow communication to a VPC. If you pass multiple subnets, App Runner will create multiple network interfaces—one for each subnet. For better availability, we recommend you specify at least two subnets.

Amazon Simple Storage Service (S3)

Object storage built to retrieve any amount of data from anywhere.

How it works

Amazon Simple Storage Service (Amazon S3) is an object storage service offering industry-leading scalability, data availability, security, and performance. Customers of all sizes and industries can store and protect any amount of data for virtually any use case, such as data lakes, cloud-native applications, and mobile apps. With cost-effective storage classes and easy-to-use management features, you can optimize costs, organize data, and configure fine-tuned access controls to meet specific business, organizational, and compliance requirements.

Use cases

Build a data lake
Run big data analytics, artificial intelligence (AI), machine learning (ML), and high performance

computing (HPC) applications to unlock data insights.

Back up and restore critical data
Meet Recovery Time Objectives (RTO), Recovery Point Objectives (RPO), and compliance requirements with S3's robust replication features.

Archive data at the lowest cost
Move data archives to the Amazon S3 Glacier storage classes to lower costs, eliminate operational complexities, and gain new insights.

Run cloud-native applications
Build fast, powerful mobile and web-based cloud-native apps that scale automatically in a highly available configuration.

What is Amazon S3?

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. Customers of all sizes and industries can use Amazon S3 to store and protect any amount of data for a range of use cases, such as data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides management features so that you can optimize, organize, and configure access to your data to meet your specific business, organizational, and compliance requirements.

Features of Amazon S3

Storage classes

Amazon S3 offers a range of storage classes designed for different use cases. For example, you can store mission-critical production data in S3 Standard for frequent access, save costs by storing infrequently accessed data in S3 Standard-IA or S3 One Zone-IA, and archive data at the lowest costs in S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive.

You can store data with changing or unknown access patterns in S3 Intelligent-Tiering, which optimizes storage costs by automatically moving your data between four access tiers when your access patterns change. These four access tiers include two low-latency access tiers optimized for frequent and infrequent access, and two opt-in archive access tiers designed for asynchronous access for rarely accessed data.

For more information, see Using Amazon S3 storage classes. For more information about S3 Glacier Flexible Retrieval, see the *Amazon S3 Glacier Developer Guide*.

Storage management

Amazon S3 has storage management features that you can use to manage costs, meet regulatory requirements, reduce latency, and save multiple distinct copies of your data for compliance requirements.

- S3 Lifecycle – Configure a lifecycle policy to manage your objects and store them cost effectively throughout their lifecycle. You can transition objects to other S3 storage classes or expire objects that reach the end of their lifetimes.
- S3 Object Lock – Prevent Amazon S3 objects from being deleted or overwritten for a fixed amount of time or indefinitely. You can use Object Lock to help meet regulatory requirements that require *write-once-read-many (WORM)* storage or to simply add another layer of protection against object changes and deletions.
- S3 Replication – Replicate objects and their respective metadata and object tags to one or more destination buckets in the same or different AWS Regions for reduced latency, compliance, security, and other use cases.
- S3 Batch Operations – Manage billions of objects at scale with a single S3 API request or a few clicks in the Amazon S3 console. You can use Batch Operations to perform operations such as **Copy**, **Invoke AWS Lambda function**, and **Restore** on millions or billions of objects.

Access management

Amazon S3 provides features for auditing and managing access to your buckets and objects. By default, S3 buckets and the objects in them are private. You have access only to the S3 resources that you create. To grant granular resource permissions that support your specific use case or to audit the permissions of your Amazon S3 resources, you can use the following features.

- S3 Block Public Access – Block public access to S3 buckets and objects. By default, Block Public Access settings are turned on at the account and bucket level.
- AWS Identity and Access Management (IAM) – Create IAM users for your AWS account to manage access to your Amazon S3 resources. For example,

you can use IAM with Amazon S3 to control the type of access a user or group of users has to an S3 bucket that your AWS account owns.

- Bucket policies – Use IAM-based policy language to configure resource-based permissions for your S3 buckets and the objects in them.
- Amazon S3 access points – Configure named network endpoints with dedicated access policies to manage data access at scale for shared datasets in Amazon S3.
- Access control lists (ACLs) – Grant read and write permissions for individual buckets and objects to authorized users. As a general rule, we recommend using S3 resource-based policies (bucket policies and access point policies) or IAM policies for access control instead of ACLs. ACLs are an access control mechanism that predates resource-based policies and IAM. For more information about when you'd use ACLs instead of resource-based policies or IAM policies, see Access policy guidelines.
- S3 Object Ownership – Disable ACLs and take ownership of every object in your bucket, simplifying access management for data stored in Amazon S3. You, as the bucket owner, automatically own and have full control over every object in your bucket, and access control for your data is based on policies.
- Access Analyzer for S3 – Evaluate and monitor your S3 bucket access policies, ensuring that the policies provide only the intended access to your S3 resources.

Data processing

To transform data and trigger workflows to automate a variety of other processing activities at scale, you can use the following features.

- S3 Object Lambda – Add your own code to S3 GET, HEAD, and LIST requests to modify and process data as it is returned to an application. Filter rows, dynamically resize images, redact confidential data, and much more.
- Event notifications – Trigger workflows that use Amazon Simple Notification Service (Amazon SNS), Amazon Simple Queue Service (Amazon SQS), and AWS Lambda when a change is made to your S3 resources.

Storage logging and monitoring

Amazon S3 provides logging and monitoring tools that you can use to monitor and control how your Amazon S3 resources are being used. For more information, see Monitoring tools.

Automated monitoring tools

- Amazon CloudWatch metrics for Amazon S3 – Track the operational health of your S3 resources and configure billing alerts when estimated charges reach a user-defined threshold.
- AWS CloudTrail – Record actions taken by a user, a role, or an AWS service in Amazon S3. CloudTrail logs provide you with detailed API tracking for S3 bucket-level and object-level operations.

Manual monitoring tools

- Server access logging – Get detailed records for the requests that are made to a bucket. You can use server access logs for many use cases, such as conducting security and access audits, learning about your customer base, and understanding your Amazon S3 bill.
- AWS Trusted Advisor – Evaluate your account by using AWS best practice checks to identify ways to optimize your AWS infrastructure, improve security and performance, reduce costs, and monitor service quotas. You can then follow the recommendations to optimize your services and resources.

Analytics and insights

Amazon S3 offers features to help you gain visibility into your storage usage, which empowers you to better understand, analyze, and optimize your storage at scale.

- Amazon S3 Storage Lens – Understand, analyze, and optimize your storage. S3 Storage Lens provides 29+ usage and activity metrics and interactive dashboards to aggregate data for your entire organization, specific accounts, AWS Regions, buckets, or prefixes.
- Storage Class Analysis – Analyze storage access patterns to decide when it's time to move data to a more cost-effective storage class.
- S3 Inventory with Inventory reports – Audit and report on objects and their corresponding metadata and configure other Amazon S3 features to take action in Inventory reports. For example, you can report on the replication and encryption status of your objects. For a list of all the metadata available for each object in Inventory reports, see Amazon S3 Inventory list.

Strong consistency

Amazon S3 provides strong read-after-write consistency for PUT and DELETE requests of objects in your Amazon S3 bucket in all AWS Regions. This behavior

applies to both writes of new objects as well as PUT requests that overwrite existing objects and DELETE requests. In addition, read operations on Amazon S3 Select, Amazon S3 access control lists (ACLs), Amazon S3 Object Tags, and object metadata (for example, the HEAD object) are strongly consistent. For more information, see Amazon S3 data consistency model.

How Amazon S3 works

Amazon S3 is an object storage service that stores data as objects within buckets. An *object* is a file and any metadata that describes the file. A *bucket* is a container for objects.

To store your data in Amazon S3, you first create a bucket and specify a bucket name and AWS Region. Then, you upload your data to that bucket as objects in Amazon S3. Each object has a *key* (or *key name*), which is the unique identifier for the object within the bucket.

S3 provides features that you can configure to support your specific use case. For example, you can use S3 Versioning to keep multiple versions of an object in the same bucket, which allows you to restore objects that are accidentally deleted or overwritten.

Buckets and the objects in them are private and can be accessed only if you explicitly grant access permissions. You can use bucket policies, AWS Identity and Access Management (IAM) policies, access control lists (ACLs), and S3 Access Points to manage access.

Buckets

A bucket is a container for objects stored in Amazon S3. You can store any number of objects in a bucket and can have up to 100 buckets in your account. To request an increase, visit the Service Quotas Console.

Every object is contained in a bucket. For example, if the object named photos/puppy.jpg is stored in the DOC-EXAMPLE-BUCKET bucket in the US West (Oregon) Region, then it is addressable using the URL https://DOC-EXAMPLE-BUCKET.s3.us-west-2.amazonaws.com/photos/puppy.jpg. For more information, see Accessing a Bucket.

When you create a bucket, you enter a bucket name and choose the AWS Region where the bucket will reside. After you create a bucket, you cannot change the

name of the bucket or its Region. Bucket names must follow the bucket naming rules. You can also configure a bucket to use S3 Versioning or other storage management features.

Buckets also:

- Organize the Amazon S3 namespace at the highest level.
- Identify the account responsible for storage and data transfer charges.
- Provide access control options, such as bucket policies, access control lists (ACLs), and S3 Access Points, that you can use to manage access to your Amazon S3 resources.
- Serve as the unit of aggregation for usage reporting.

For more information about buckets, see Buckets overview.

Objects

Objects are the fundamental entities stored in Amazon S3. Objects consist of object data and metadata. The metadata is a set of name-value pairs that describe the object. These pairs include some default metadata, such as the date last modified, and standard HTTP metadata, such as `Content-Type`. You can also specify custom metadata at the time that the object is stored.

An object is uniquely identified within a bucket by a key (name) and a version ID (if S3 Versioning is enabled on the bucket). For more information about objects, see Amazon S3 objects overview.

Keys

An *object key* (or *key name*) is the unique identifier for an object within a bucket. Every object in a bucket has exactly one key. The combination of a bucket, object key, and optionally, version ID (if S3 Versioning is enabled for the bucket) uniquely identify each object. So you can think of Amazon S3 as a basic data map between "bucket + key + version" and the object itself.

Every object in Amazon S3 can be uniquely addressed through the combination of the web service endpoint, bucket name, key, and optionally, a version. For example, in the URL `https://DOC-EXAMPLE-BUCKET.s3.us-west-2.amazonaws.com/photos/puppy.jpg`, `DOC-EXAMPLE-BUCKET` is the name of the bucket and `photos/puppy.jpg` is the key.

For more information about object keys, see Creating object key names.

S3 Versioning

You can use S3 Versioning to keep multiple variants of an object in the same bucket. With S3 Versioning, you can preserve, retrieve, and restore every version of every object stored in your buckets. You can easily recover from both unintended user actions and application failures.

For more information, see Using versioning in S3 buckets.

Version ID

When you enable S3 Versioning in a bucket, Amazon S3 generates a unique version ID for each object added to the bucket. Objects that already existed in the bucket at the time that you enable versioning have a version ID of null. If you modify these (or any other) objects with other operations, such as CopyObject and PutObject, the new objects get a unique version ID.

For more information, see Using versioning in S3 buckets.

Bucket policy

A bucket policy is a resource-based AWS Identity and Access Management (IAM) policy that you can use to grant access permissions to your bucket and the objects in it. Only the bucket owner can associate a policy with a bucket. The permissions attached to the bucket apply to all of the objects in the bucket that are owned by the bucket owner. Bucket policies are limited to 20 KB in size.

Bucket policies use JSON-based access policy language that is standard across AWS. You can use bucket policies to add or deny permissions for the objects in a bucket. Bucket policies allow or deny requests based on the elements in the policy, including the requester, S3 actions, resources, and aspects or conditions of the request (for example, the IP address used to make the request). For example, you can create a bucket policy that grants cross-account permissions to upload objects to an S3 bucket while ensuring that the bucket owner has full control of the uploaded objects. For more information, see Bucket policy examples.

In your bucket policy, you can use wildcard characters on Amazon Resource Names (ARNs) and other values to grant permissions to a subset of objects. For

example, you can control access to groups of objects that begin with a common prefix or end with a given extension, such as .html.

S3 Access Points

Amazon S3 Access Points are named network endpoints with dedicated access policies that describe how data can be accessed using that endpoint. Access Points are attached to buckets that you can use to perform S3 object operations, such as GetObject and PutObject. Access Points simplify managing data access at scale for shared datasets in Amazon S3.

Each access point has its own access point policy. You can configure Block Public Access settings for each access point. To restrict Amazon S3 data access to a private network, you can also configure any access point to accept requests only from a virtual private cloud (VPC).

For more information, see Managing data access with Amazon S3 access points.

Access control lists (ACLs)

You can use ACLs to grant read and write permissions to authorized users for individual buckets and objects. Each bucket and object has an ACL attached to it as a subresource. The ACL defines which AWS accounts or groups are granted access and the type of access. ACLs are an access control mechanism that predates IAM. For more information about ACLs, see Access control list (ACL) overview.

By default, when another AWS account uploads an object to your S3 bucket, that account (the object writer) owns the object, has access to it, and can grant other users access to it through ACLs. You can use Object Ownership to change this default behavior so that ACLs are disabled and you, as the bucket owner, automatically own every object in your bucket. As a result, access control for your data is based on policies, such as IAM policies, S3 bucket policies, virtual private cloud (VPC) endpoint policies, and AWS Organizations service control policies (SCPs).

A majority of modern use cases in Amazon S3 no longer require the use of ACLs, and we recommend that you disable ACLs except in unusual circumstances where you need to control access for each object individually. With Object Ownership, you can disable ACLs and rely on policies for access control. When you disable ACLs,

you can easily maintain a bucket with objects uploaded by different AWS accounts. You, as the bucket owner, own all the objects in the bucket and can manage access to them using policies. For more information, see Controlling ownership of objects and disabling ACLs for your bucket.

Regions

You can choose the geographical AWS Region where Amazon S3 stores the buckets that you create. You might choose a Region to optimize latency, minimize costs, or address regulatory requirements. Objects stored in an AWS Region never leave the Region unless you explicitly transfer or replicate them to another Region. For example, objects stored in the Europe (Ireland) Region never leave it.

Note

You can access Amazon S3 and its features only in the AWS Regions that are enabled for your account. For more information about enabling a Region to create and manage AWS resources, see Managing AWS Regions in the *AWS General Reference*.

For a list of Amazon S3 Regions and endpoints, see Regions and endpoints in the *AWS General Reference*.

Amazon S3 data consistency model

Amazon S3 provides strong read-after-write consistency for PUT and DELETE requests of objects in your Amazon S3 bucket in all AWS Regions. This behavior applies to both writes to new objects as well as PUT requests that overwrite existing objects and DELETE requests. In addition, read operations on Amazon S3 Select, Amazon S3 access controls lists (ACLs), Amazon S3 Object Tags, and object metadata (for example, the HEAD object) are strongly consistent.

Updates to a single key are atomic. For example, if you make a PUT request to an existing key from one thread and perform a GET request on the same key from a second thread concurrently, you will get either the old data or the new data, but never partial or corrupt data.

Amazon S3 achieves high availability by replicating data across multiple servers within AWS data centers. If a PUT request is successful, your data is safely stored. Any read (GET or LIST request) that is initiated following the receipt of a successful

PUT response will return the data written by the PUT request. Here are examples of this behavior:

- A process writes a new object to Amazon S3 and immediately lists keys within its bucket. The new object appears in the list.
- A process replaces an existing object and immediately tries to read it. Amazon S3 returns the new data.
- A process deletes an existing object and immediately tries to read it. Amazon S3 does not return any data because the object has been deleted.
- A process deletes an existing object and immediately lists keys within its bucket. The object does not appear in the listing.

Note

- Amazon S3 does not support object locking for concurrent writers. If two PUT requests are simultaneously made to the same key, the request with the latest timestamp wins. If this is an issue, you must build an object-locking mechanism into your application.
- Updates are key-based. There is no way to make atomic updates across keys. For example, you cannot make the update of one key dependent on the update of another key unless you design this functionality into your application.

Bucket configurations have an eventual consistency model. Specifically, this means that:

- If you delete a bucket and immediately list all buckets, the deleted bucket might still appear in the list.
- If you enable versioning on a bucket for the first time, it might take a short amount of time for the change to be fully propagated. We recommend that you wait for 15 minutes after enabling versioning before issuing write operations (PUT or DELETE requests) on objects in the bucket.

Concurrent applications

This section provides examples of behavior to be expected from Amazon S3 when multiple clients are writing to the same items.

In this example, both W1 (write 1) and W2 (write 2) finish before the start of R1 (read 1) and R2 (read 2). Because S3 is strongly consistent, R1 and R2 both return `color = ruby`.

In the next example, W2 does not finish before the start of R1. Therefore, R1 might return color = ruby or color = garnet. However, because W1 and W2 finish before the start of R2, R2 returns color = garnet.

In the last example, W2 begins before W1 has received an acknowledgement. Therefore, these writes are considered concurrent. Amazon S3 internally uses last-writer-wins semantics to determine which write takes precedence. However, the order in which Amazon S3 receives the requests and the order in which applications receive acknowledgements cannot be predicted because of various factors, such as network latency. For example, W2 might be initiated by an Amazon EC2 instance in the same Region, while W1 might be initiated by a host that is farther away. The best way to determine the final value is to perform a read after both writes have been acknowledged.

Related services

After you load your data into Amazon S3, you can use it with other AWS services. The following are the services that you might use most frequently:

- **Amazon Elastic Compute Cloud (Amazon EC2)** – Provides secure and scalable computing capacity in the AWS Cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage.
- **Amazon EMR** – Helps businesses, researchers, data analysts, and developers easily and cost-effectively process vast amounts of data. Amazon EMR uses a hosted Hadoop framework running on the web-scale infrastructure of Amazon EC2 and Amazon S3.
- **AWS Snow Family** – Helps customers that need to run operations in austere, non-data center environments, and in locations where there's a lack of consistent network connectivity. You can use AWS Snow Family devices to locally and cost-effectively access the storage and compute power of the AWS Cloud in places where an internet connection might not be an option.
- **AWS Transfer Family** – Provides fully managed support for file transfers directly into and out of Amazon S3 or Amazon Elastic File System (Amazon EFS) using Secure Shell (SSH) File Transfer Protocol (SFTP), File Transfer Protocol over SSL (FTPS), and File Transfer Protocol (FTP).

Accessing Amazon S3

You can work with Amazon S3 in any of the following ways:

AWS Management Console

The console is a web-based user interface for managing Amazon S3 and AWS resources. If you've signed up for an AWS account, you can access the Amazon S3 console by signing into the AWS Management Console and choosing **S3** from the AWS Management Console home page.

AWS Command Line Interface

You can use the AWS command line tools to issue commands or build scripts at your system's command line to perform AWS (including S3) tasks.

The AWS Command Line Interface (AWS CLI) provides commands for a broad set of AWS services. The AWS CLI is supported on Windows, macOS, and Linux. To get started, see the *AWS Command Line Interface User Guide*. For more information about the commands for Amazon S3, see s3api and s3control in the *AWS CLI Command Reference*.

AWS SDKs

AWS provides SDKs (software development kits) that consist of libraries and sample code for various programming languages and platforms (Java, Python, Ruby, .NET, iOS, Android, and so on). The AWS SDKs provide a convenient way to create programmatic access to S3 and AWS. Amazon S3 is a REST service. You can send requests to Amazon S3 using the AWS SDK libraries. which wrap the underlying Amazon S3 REST API and simplify your programming tasks. For example, the SDKs take care of tasks such as calculating signatures, cryptographically signing requests, managing errors, and retrying requests automatically. For information about the AWS SDKs, including how to download and install them, see Tools for AWS.

Every interaction with Amazon S3 is either authenticated or anonymous. If you are using the AWS SDKs, the libraries compute the signature for authentication from the keys that you provide. For more information about how to make requests to Amazon S3, see Making requests.

Amazon S3 REST API

The architecture of Amazon S3 is designed to be programming language-neutral, using AWS-supported interfaces to store and retrieve objects. You can access S3

and AWS programmatically by using the Amazon S3 REST API. The REST API is an HTTP interface to Amazon S3. With the REST API, you use standard HTTP requests to create, fetch, and delete buckets and objects.

To use the REST API, you can use any toolkit that supports HTTP. You can even use a browser to fetch objects, as long as they are anonymously readable.

The REST API uses standard HTTP headers and status codes, so that standard browsers and toolkits work as expected. In some areas, we have added functionality to HTTP (for example, we added headers to support access control). In these cases, we have done our best to add the new functionality in a way that matches the style of standard HTTP usage.

If you make direct REST API calls in your application, you must write the code to compute the signature and add it to the request. For more information about how to make requests to Amazon S3, see Making requests.

Note

SOAP API support over HTTP is deprecated, but it is still available over HTTPS. Newer Amazon S3 features are not supported for SOAP. We recommend that you use either the REST API or the AWS SDKs.

PCI DSS compliance

Amazon S3 supports the processing, storage, and transmission of credit card data by a merchant or service provider, and has been validated as being compliant with Payment Card Industry (PCI) Data Security Standard (DSS). For more information about PCI DSS, including how to request a copy of the AWS PCI Compliance Package, see PCI DSS Level 1.

Amazon Simple Storage Service FAQs

General S3 FAQs

Q: What is Amazon S3?

Amazon S3 is object storage built to store and retrieve any amount of data from anywhere. It's a simple storage service that offers industry leading durability, availability, performance, security, and virtually unlimited scalability at very low costs.

Q: What can I do with Amazon S3?

Amazon S3 provides a simple web service interface that you can use to store and retrieve any amount of data, at any time, from anywhere. Using this service, you can easily build applications that make use of cloud native storage. Since Amazon S3 is highly scalable and you only pay for what you use, you can start small and grow your application as you wish, with no compromise on performance or reliability.

Amazon S3 is also designed to be highly flexible. Store any type and amount of data that you want, read the same piece of data a million times or only for emergency disaster recovery, build a simple FTP application or a sophisticated web application such as the Amazon.com retail web site. Amazon S3 frees developers to focus on innovation instead of figuring out how to store their data.

Q: How can I get started using Amazon S3?

To sign up for Amazon S3, click this link. You must have an Amazon Web Services account to access this service. If you do not already have an account, you will be prompted to create one when you begin the Amazon S3 sign-up process. After signing up, please refer to the Amazon S3 documentation and sample code in the Resource Center to begin using Amazon S3.

Q: What can developers do with Amazon S3 that they could not do with an on-premises solution?

Amazon S3 lets developers leverage Amazon's own benefits of massive scale with no up-front investment or performance compromises. Developers are now free to innovate knowing that no matter how successful their businesses become, it will be inexpensive and simple to ensure their data is quickly accessible, always available, and secure.

Q: What kind of data can I store in Amazon S3?

You can store virtually any kind of data in any format. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: How much data can I store in Amazon S3?

The total volume of data and number of objects you can store are unlimited. Individual Amazon S3 objects can range in size from a minimum of 0 bytes to a maximum of 5 TB. The largest object that can be uploaded in a single PUT is 5 GB. For objects larger than 100 MB, customers should consider using the Multipart Upload capability.

Q: Can I have a bucket that has different objects in different storage classes?

Yes, you can have an S3 bucket that has different objects stored in S3 Standard, S3 Intelligent-Tiering, S3 Standard-IA, S3 One Zone-IA, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive.

Q: What does Amazon do with my data in Amazon S3?

Amazon will store your data and track its associated usage for billing purposes. Amazon will not otherwise access your data for any purpose outside of the Amazon S3 offering, except when required to do so by law. Please refer to the Amazon Web Services Licensing Agreement for details.

Q: Does Amazon store its own data in Amazon S3?

Yes. Developers within Amazon use Amazon S3 for a wide variety of projects. Many of these projects use Amazon S3 as their authoritative data store and rely on it for business-critical operations.

Q: How is Amazon S3 data organized?

Amazon S3 is a simple key-based object store. When you store data, you assign a unique object key that can later be used to retrieve the data. Keys can be any string, and they can be constructed to mimic hierarchical attributes. Alternatively, you can use S3 Object Tagging to organize your data across all of your S3 buckets and/or prefixes.

Q: How do I interface with Amazon S3?

Amazon S3 provides a simple, standards-based REST web services interface that is designed to work with any internet-development toolkit. The operations are intentionally made simple to make it easy to add new distribution protocols and functional layers.

Q: How reliable is Amazon S3?

Amazon S3 gives any developer access to the same highly scalable, highly available, fast, inexpensive data storage infrastructure that Amazon uses to run its own global network of web sites. The S3 Standard storage class is designed for 99.99% availability, the S3 Standard-IA storage class, S3 Intelligent-Tiering storage class, and the S3 Glacier Instant Retrieval storage classes are designed for 99.9% availability, the S3 One Zone-IA storage class is designed for 99.5% availability, and the S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive class are designed for 99.99% availability and an SLA of 99.9%. All of these storage classes are backed by the Amazon S3 Service Level Agreement.

Q: How will Amazon S3 perform if traffic from my application suddenly spikes?

Amazon S3 was designed from the ground up to handle traffic for any internet application. Pay-as-you-go pricing and unlimited capacity ensures that your incremental costs don’t change and that your service is not interrupted. Amazon S3’s massive scale lets you spread the load evenly, so that no individual application is affected by traffic spikes.

Q: Does Amazon S3 offer a Service Level Agreement (SLA)?

Yes. The Amazon S3 SLA provides for a service credit if a customer's monthly uptime percentage is below our service commitment in any billing cycle.

Q: What is the consistency model for Amazon S3?

Amazon S3 delivers strong read-after-write consistency automatically, without changes to performance or availability, without sacrificing regional isolation for applications, and at no additional cost.

After a successful write of a new object or an overwrite of an existing object, any subsequent read request immediately receives the latest version of the object. S3 also provides strong

consistency for list operations, so after a write, you can immediately perform a listing of the objects in a bucket with any changes reflected.

Q: Why does strong read-after-write consistency help me?

Strong read-after-write consistency helps when you need to immediately read an object after a write -- for example, when you often read and list immediately after writing objects. High-performance computing workloads also benefit in that when an object is overwritten and then read many times simultaneously, strong read-after-write consistency provides assurance that the latest write is read across all reads. These applications automatically and immediately benefit from strong read-after-write consistency. The strong consistency of S3 also reduces costs by removing the need for extra infrastructure to provide strong consistency.

AWS Regions

Q: Where is my data stored?

You specify an AWS Region when you create your Amazon S3 bucket. For S3 Standard, S3 Standard-IA, S3 Intelligent-Tiering, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive storage classes, your objects are automatically stored across multiple devices spanning a minimum of three Availability Zones, each separated by miles across an AWS Region. Objects stored in the S3 One Zone-IA storage class are stored redundantly within a single Availability Zone in the AWS Region you select. For S3 on Outposts, your data is stored in your Outpost on-premises environment, unless you manually choose to transfer it to an AWS Region. Please refer to Regional Products and Services for details of Amazon S3 service availability by AWS Region.

Q: What is an AWS Region?

An AWS Region is a geographic location where AWS provides multiple, physically separated, and isolated Availability Zones which are connected with low latency, high throughput, and highly redundant networking.

Q: What is an AWS Availability Zone (AZ)?

An AWS Availability Zone is a physically isolated location within an AWS Region. Within each AWS Region, S3 operates in a minimum of three AZs, each separated by miles to protect against local events like fires, floods, etc.

Amazon S3 Standard, S3 Standard-Infrequent Access, S3 Intelligent-Tiering, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive storage classes replicate data across a minimum of three AZs to protect against the loss of one entire AZ. This remains true in Regions where fewer than three AZs are publicly available. Objects stored in these storage classes are available for access from all of the AZs in an AWS Region.

The Amazon S3 One Zone-IA storage class replicates data within a single AZ. The data stored in S3 One Zone-IA is not resilient to the physical loss of an Availability Zone resulting from disasters, such as earthquakes and floods.

Q: How do I decide which AWS Region to store my data in?

There are several factors to consider based on your specific application. For instance, you may want to store your data in a Region that is near your customers, your data centers, or other

AWS resources to reduce data access latencies. You may also want to store your data in a Region that is remote from your other operations for geographic redundancy and disaster recovery purposes. You should also consider Regions that let you address specific legal and regulatory requirements and/or reduce your storage costs—you can choose a lower priced Region to save money.

Q: In which parts of the world is Amazon S3 available?

Amazon S3 is available in AWS Regions worldwide, and you can use Amazon S3 regardless of your location. You just have to decide which AWS Region(s) you want to store your Amazon S3 data. See the AWS Regional Availability Table for a list of AWS Regions in which S3 is available today.

Amazon S3 and IPv6

Q: What is IPv6?

Every server and device connected to the internet must have a unique address. Internet Protocol Version 4 (IPv4) was the original 32-bit addressing scheme. However, the continued growth of the internet means that all available IPv4 addresses will be utilized over time. Internet Protocol Version 6 (IPv6) is an addressing mechanism designed to overcome the global address limitation on IPv4.

Q: What can I do with IPv6?

Using IPv6 support for Amazon S3, applications can connect to Amazon S3 without the need for any IPv6 to IPv4 translation software or systems. You can meet compliance requirements, more easily integrate with existing IPv6-based on-premises applications, and remove the need for expensive networking equipment to handle the address translation. You can also now utilize the existing source address filtering features in IAM policies and bucket policies with IPv6 addresses, expanding your options to secure applications interacting with Amazon S3.

Q: How do I get started with IPv6 on Amazon S3?

You can get started by pointing your application to Amazon S3's new "dual-stack" endpoint, which supports access over both IPv4 and IPv6. In most cases, no further configuration is required for access over IPv6, because most network clients prefer IPv6 addresses by default. Applications that are impacted by using IPv6 can switch back to the standard IPv4-only endpoints at any time. IPv6 with Amazon S3 is supported in all commercial AWS Regions, including AWS GovCloud (US) Regions, Amazon Web Services China (Beijing) Region, operated by Sinnet and Amazon Web Services China (Ningxia) Region, operated by NWCD.

Q: Should I expect a change in Amazon S3 performance when using IPv6?

No, you will see the same performance when using either IPv4 or IPv6 with Amazon S3.

S3 Event Notifications

Q: What are Amazon S3 Event Notifications?

You can enable Amazon S3 Event Notifications and receive them in response to specific events in your S3 bucket, such as PUT, POST, COPY, and DELETE events. You can publish notifications to Amazon EventBridge, Amazon SNS, Amazon SQS, or directly to AWS Lambda.

Q: What can I do with Amazon S3 Event Notifications?

Amazon S3 Event Notifications let you to run workflows, send alerts, or perform other actions in response to changes in your objects stored in S3. You can use S3 Event Notifications to set up triggers to perform actions including transcoding media files when they are uploaded, processing data files when they become available, and synchronizing S3 objects with other data stores. You can also set up event notifications based on object name prefixes and suffixes. For example, you can choose to receive notifications on object names that start with “images/."

Q: What is included in an Amazon S3 Event Notifications?

For a detailed description of the information included in Amazon S3 Event Notification messages, please refer to the Configuring Amazon S3 Event Notifications topic in the Amazon S3 Developer Guide.

Q: How do I set up Amazon S3 Event Notifications?

For a detailed description of how to configure event notifications, please refer to the Configuring Amazon S3 Event Notifications topic in the Amazon S3 Developer Guide. You can learn more about AWS messaging services in the Amazon SNS Documentation and the Amazon SQS Documentation.

Q: What does it cost to use Amazon S3 Event Notifications?

There are no additional charges for using Amazon S3 for event notifications. You pay only for use of Amazon SNS or Amazon SQS to deliver event notifications, or for the cost of running an AWS Lambda function.

Amazon S3 Transfer Acceleration

Q: What is S3 Transfer Acceleration?

Amazon S3 Transfer Acceleration creates fast, easy, and secure transfers of files over long distances between your client and your Amazon S3 bucket. S3 Transfer Acceleration leverages Amazon CloudFront’s globally distributed AWS Edge Locations. As data arrives at an AWS Edge Location, data is routed to your Amazon S3 bucket over an optimized network path.

Q: How do I get started with S3 Transfer Acceleration?

To get started with S3 Transfer Acceleration enable S3 Transfer Acceleration on an S3 bucket using the Amazon S3 console, the Amazon S3 API, or the AWS CLI. After S3 Transfer Acceleration is enabled, you can point your Amazon S3 PUT and GET requests to the s3-accelerate endpoint domain name. Your data transfer application must use one of the following two types of endpoints to access the bucket for faster data transfer: .s3-accelerate.amazonaws.com or .s3-accelerate.dualstack.amazonaws.com for the “dual-stack” endpoint. If you want to use standard data transfer, you can continue to use the regular endpoints.

There are certain restrictions on which buckets will support S3 Transfer Acceleration. For details, please refer the Amazon S3 developer guide.

Q: How fast is S3 Transfer Acceleration?

S3 Transfer Acceleration helps you fully use your bandwidth, minimize the effect of distance on throughput, and is designed to ensure consistently fast data transfer to Amazon S3 regardless

of your client's location. The amount of acceleration primarily depends on your available bandwidth, the distance between the source and destination, and packet loss rates on the network path. Generally, you will see more acceleration when the source is farther from the destination, when there is more available bandwidth, and/or when the object size is bigger.

One customer measured a 50% reduction in their average time to ingest 300 MB files from a global user base spread across the US, Europe, and parts of Asia to a bucket in the Asia Pacific (Sydney) Region. Another customer observed cases where performance improved in excess of 500% for users in South East Asia and Australia uploading 250 MB files (in parts of 50MB) to an S3 bucket in the US East (N. Virginia) Region.

Try the speed comparison tool to get a preview of the performance benefit from your location.

Q: Who should use S3 Transfer Acceleration?

S3 Transfer Acceleration is designed to optimize transfer speeds from across the world into S3 buckets. If you are uploading to a centralized bucket from geographically dispersed locations or if you regularly transfer GBs or TBs of data across continents, you may save hours or days of data transfer time with S3 Transfer Acceleration.

Q: How secure is S3 Transfer Acceleration?

S3 Transfer Acceleration provides the same security as regular transfers to Amazon S3. All Amazon S3 security features, such as access restriction based on a client's IP address, are supported as well. S3 Transfer Acceleration communicates with clients over standard TCP and does not require firewall changes. No data is ever saved at AWS Edge Locations.

Q: What if S3 Transfer Acceleration is not faster than a regular Amazon S3 transfer?

Each time you use S3 Transfer Acceleration to upload an object, we will check whether S3 Transfer Acceleration is likely to be faster than a regular Amazon S3 transfer. If we determine that S3 Transfer Acceleration is not likely to be faster than a regular Amazon S3 transfer of the same object to the same destination AWS Region, we will not charge for the use of S3 Transfer Acceleration for that transfer, and we may bypass the S3 Transfer Acceleration system for that upload.

Q: Can I use S3 Transfer Acceleration with multipart uploads?

Yes, S3 Transfer Acceleration supports all bucket level features including multipart uploads.

Q: How should I choose between S3 Transfer Acceleration and Amazon CloudFront's PUT/POST?

S3 Transfer Acceleration optimizes the TCP protocol and adds additional intelligence between the client and the S3 bucket, making S3 Transfer Acceleration a better choice if a higher throughput is desired. If you have objects that are smaller than 1 GB or if the data set is less than 1 GB in size, you should consider using Amazon CloudFront's PUT/POST commands for optimal performance.

Q: How should I choose between S3 Transfer Acceleration and AWS Snow Family (Snowball, Snowball Edge, and Snowmobile)?

The AWS Snow Family is ideal for customers moving large batches of data at once. The AWS Snowball has a typical 5—7 days turnaround time. As a rule of thumb, S3 Transfer Acceleration over a fully-utilized 1 Gbps line can transfer up to 75 TBs in the same time period. In general, if it will take more than a week to transfer over the internet, or there are recurring transfer jobs and there is more than 25Mbps of available bandwidth, S3 Transfer Acceleration is a good option.

Another option is to use both: perform initial heavy lift moves with an AWS Snowball (or series of AWS Snowballs) and then transfer incremental ongoing changes with S3 Transfer Acceleration.

Q: Can S3 Transfer Acceleration complement AWS Direct Connect?

AWS Direct Connect is a good choice for customers who have a private networking requirement or who have access to AWS Direct Connect exchanges. S3 Transfer Acceleration is best for submitting data from distributed client locations over the public internet, or where variable network conditions make throughput poor. Some AWS Direct Connect customers use S3 Transfer Acceleration to help with remote office transfers where they may suffer from poor internet performance.

Q: Can S3 Transfer Acceleration complement the AWS Storage Gateway or a third-party gateway?

You can benefit from configuring the bucket destination in your third-party gateway to use an S3 Transfer Acceleration endpoint domain.

Visit this File section of the Storage Gateway FAQ to learn more about the AWS implementation.

Q: Can S3 Transfer Acceleration complement third-party integrated software?

Yes. Software packages that connect directly into Amazon S3 can take advantage of S3 Transfer Acceleration when they send their jobs to Amazon S3.

Q: Is S3 Transfer Acceleration HIPAA eligible?

Yes, AWS has expanded its HIPAA compliance program to include Amazon S3 Transfer Acceleration as a HIPAA eligible service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Amazon S3 Transfer Acceleration to make fast, easy, and secure transfers of files, including protected health information (PHI) over long distances between your client and your Amazon S3 bucket.

Security

S3 Access Points

Q: How secure is my data in Amazon S3?

Amazon S3 is secure by default. Upon creation, only you have access to Amazon S3 buckets that you create, and you have complete control over who has access to your data. Amazon S3 supports user authentication to control access to data. You can use access control mechanisms such as bucket policies to selectively grant permissions to users and groups of users. The Amazon S3 console highlights your publicly accessible buckets, indicates the source of public accessibility, and also warns you if changes to your bucket policies or bucket ACLs would make your bucket publicly accessible. You should enable Block Public Access for all accounts and buckets that you do not want publicly accessible.

You can securely upload/download your data to Amazon S3 via SSL endpoints using the HTTPS protocol. If you need extra security you can use the Server-Side Encryption (SSE) option to encrypt data stored at rest. You can configure your Amazon S3 buckets to automatically encrypt objects before storing them if the incoming storage requests do not have any encryption information. Alternatively, you can use your own encryption libraries to encrypt data before storing it in Amazon S3.

For more information on security on AWS please refer to the AWS security page, and for S3 security information visit the S3 security page or the S3 security best practices guide.

Q: How can I control access to my data stored on Amazon S3?

Customers can use a number of mechanisms for controlling access to Amazon S3 resources, including AWS Identity and Access Management (IAM) policies, bucket policies, access point policies, access control lists (ACLs), Query String Authentication, Amazon Virtual Private Cloud (Amazon VPC) endpoint policies, service control policies (SCPs) in AWS Organizations, and Amazon S3 Block Public Access.

IAM
IAM lets organizations with multiple employees create and manage multiple users under a single AWS account. With IAM policies, customers can grant IAM users fine-grained control to their Amazon S3 bucket or objects while also retaining full control over everything the users do.

Bucket and access point policies
With bucket policies and access point policies, customers can define rules which apply broadly across all requests to their Amazon S3 resources, such as granting write privileges to a subset of Amazon S3 resources. Customers can also restrict access based on an aspect of the request, such as HTTP referrer and IP address.

ACLs
Amazon S3 supports our original access control method, access control lists (ACLs). With ACLs, customers can grant specific permissions (i.e. READ, WRITE, FULL_CONTROL) to specific users for an individual bucket or object. For customers who prefer to use exclusively policies for access control, Amazon S3 offers the Object Ownership feature to disable ACLs.

Query String Authentication
With Query String Authentication, customers can create a URL to an Amazon S3 object which is only valid for a limited time. For more information on the various access control policies available in Amazon S3, refer to the Access Control topic in the Amazon S3 Developer Guide.

Amazon VPC
When customers create an Amazon VPC endpoint, they can attach an endpoint policy to it that controls access to the Amazon S3 resources to which they are connecting. Customers can also use Amazon S3 bucket policies to control access to buckets from specific endpoints or specific VPCs.

Service control policies
Service control policies (SCPs) are a type of AWS Organizations policy that customers can use to manage permissions in their organization. SCPs offer central control over the maximum available permissions for all accounts in an organization. With SCPs, customers can ensure their accounts stay within the organization's access control guidelines.

S3 Block Public Access
Amazon S3 Block Public Access provides settings for access points, buckets, and accounts to help customers manage public access to Amazon S3 resources. With S3 Block Public Access, account administrators and bucket owners can easily set up centralized controls to limit public access to their Amazon S3 resources that are enforced regardless of how the resources are created.

Learn more about policies and permissions in the AWS IAM User Guide.

Q: Does Amazon S3 support data access auditing?

Yes, customers can optionally configure an Amazon S3 bucket to create access log records for all requests made against it. Alternatively, customers who need to capture IAM/user identity information in their logs can configure AWS CloudTrail Data Events.

These access log records can be used for audit purposes and contain details about the request, such as the request type, the resources specified in the request, and the time and date the request was processed.

Q: What options do I have for encrypting data stored on Amazon S3?

You can choose to encrypt data using SSE-S3, SSE-C, SSE-KMS, or an encryption client library. All four enable you to store sensitive data encrypted at rest in Amazon S3.

SSE-S3 provides an integrated solution where Amazon handles key management and key protection using multiple layers of security. You should choose SSE-S3 if you prefer to have Amazon manage your keys.

SSE-C lets Amazon S3 perform the encryption and decryption of your objects while retaining control of the keys used to encrypt objects. With SSE-C, you don't need to implement or use a client-side library to perform the encryption and decryption of objects you store in Amazon S3, but you do need to manage the keys that you send to Amazon S3 to encrypt and decrypt objects. Use SSE-C if you want to maintain your own encryption keys, but don't want to implement or leverage a client-side encryption library.

SSE-KMS lets AWS Key Management Service (AWS KMS) manage your encryption keys. Using AWS KMS to manage your keys provides several additional benefits. With AWS KMS, there are separate permissions for the use of the KMS key, providing an additional layer of control and protection against unauthorized access to your objects stored in Amazon S3. AWS KMS provides an audit trail so you can see who used your key to access which object and when, as well as view failed attempts to access data from users without permission to decrypt the data. Also, AWS KMS provides additional security controls to support customer efforts to comply with PCI-DSS, HIPAA/HITECH, and FedRAMP industry requirements.

Using an encryption client library, you retain control of the keys and complete the encryption and decryption of objects client-side using an encryption library of your choice. Some customers prefer full end-to-end control of the encryption and decryption of objects; that way, only encrypted objects are transmitted over the internet to Amazon S3. Use a client-side library if you want to maintain control of your encryption keys, are able to implement or use a client-side encryption library, and need to have your objects encrypted before they are sent to Amazon S3 for storage.

For more information on using Amazon S3 SSE-S3, SSE-C, or SSE-KMS, please refer to the topic on Using Encryption in the Amazon S3 Developer Guide.

Q: Can I comply with EU data privacy regulations using Amazon S3?

Customers can choose to store all data in the EU by using the EU (Frankfurt), EU (Ireland), EU (London), or EU (Paris) Region. You can also use S3 on Outposts to keep all of your data on-premises on the Outpost, and you may choose to transfer data between Outposts or to an AWS Region. It is your responsibility to ensure that you comply with EU privacy laws. Please see the AWS GDPR Center for more information. If you have more specific location requirements or other data privacy regulations that require you to keep data in a location where there is not an AWS Region, you can use S3 on Outposts.

Q: What is an Amazon VPC Endpoint for Amazon S3?

An Amazon VPC Endpoint for Amazon S3 is a logical entity within a VPC that allows connectivity to S3 over the Amazon global network. There are two types of VPC endpoints for S3: gateway VPC endpoints and interface VPC endpoints. Gateway endpoints are a gateway that you specify in your route table to access S3 from your VPC over the Amazon network. Interface endpoints extend the functionality of gateway endpoints by using private IPs to route requests to S3 from within your VPC, on-premises, or from a different AWS Region. For more information, visit the documentation.

Q: Can I allow a specific Amazon VPC Endpoint access to my Amazon S3 bucket?

You can limit access to your bucket from a specific Amazon VPC Endpoint or a set of endpoints using Amazon S3 bucket policies. S3 bucket policies now support a condition, aws:sourceVpce, that you can use to restrict access. For more details and example policies, read Using VPC Endpoints.

Q: What is AWS PrivateLink for Amazon S3?

AWS PrivateLink for S3 provides private connectivity between Amazon S3 and on-premises. You can provision interface VPC endpoints for S3 in your VPC to connect your on-premises applications directly to S3 over AWS Direct Connect or AWS VPN. You no longer need to use public IPs, change firewall rules, or configure an internet gateway to access S3 from on-premises. To learn more visit the documentation.

Q: How do I get started with interface VPC endpoints for S3?

You can create an interface VPC endpoint using the AWS VPC Management Console, AWS Command Line Interface (AWS CLI), AWS SDK or API. To learn more, visit the documentation.

Q: When should I choose gateway VPC endpoints versus AWS PrivateLink-based interface VPC endpoints?

We recommend that you use interface VPC endpoints to access S3 from on-premises or from a VPC in another AWS Region. For resources that are accessing S3 from VPC in the same AWS Region as S3, we recommend using gateway VPC endpoints as they are not billed. To learn more, visit the documentation.

Q: Can I use both Interface Endpoints and Gateway Endpoints for S3 in the same VPC?

Yes. If you have an existing gateway VPC endpoint, create an interface VPC endpoint in your VPC and update your client applications with the VPC endpoint specific endpoint names. For example, if your VPC endpoint id of the interface endpoint is vpce-0fe5b17a0707d6abc-29p5708s in us-east-1 Region, then your endpoint specific DNS name will be vpce-0fe5b17a0707d6abc-29p5708s.s3.us-east-1.vpce.amazonaws.com. In this case, only the requests to the VPC endpoint specific names will route through Interface VPC endpoints to S3 while all other requests would continue to route through the gateway VPC endpoint. To learn more, visit the documentation.

Q: What is Amazon Macie and how can I use it to secure my data?

Amazon Macie is an AI-powered security service that helps you prevent data loss by automatically discovering, classifying, and protecting sensitive data stored in Amazon S3.

Amazon Macie uses machine learning to recognize sensitive data such as personally identifiable information (PII) or intellectual property, assigns a business value, and provides visibility into where this data is stored and how it is being used in your organization. Amazon Macie continuously monitors data access activity for anomalies, and delivers alerts when it detects risk of unauthorized access or inadvertent data leaks.

You can use Amazon Macie to protect against security threats by continuously monitoring your data and account credentials. Amazon Macie gives you an automated and low-touch way to discover and classify your business data. It provides controls via templated Lambda functions to revoke access or trigger password reset policies upon the discovery of suspicious behavior, unauthorized data access to entities, or third-party applications. When alerts are generated, you can use Amazon Macie for incident response, using Amazon CloudWatch Events to swiftly take action to protect your data. For more information, go to the Amazon Macie User Guide.

Q: What is IAM Access Analyzer for Amazon S3 and how does it work?

Access Analyzer for S3 is a feature that helps you simplify permissions management as you set, verify, and refine policies for your S3 buckets and access points.

Access Analyzer for S3 monitors your existing access policies to verify that they provide only the required access to your S3 resources. Access Analyzer for S3 evaluates your bucket access policies and helps you to discover and swiftly make changes to buckets that do not require access.

Access Analyzer for S3 alerts you when you have a bucket that is configured to allow access to anyone on the internet or that is shared with other AWS accounts. You receive *findings* about the source and level of public or shared access. For example, Access Analyzer for S3 will proactively inform you if unrequired read or write access was provided through an access control list or bucket policy. With these findings, you can immediately set or restore the required access policy.

When reviewing results that show potentially shared access to a bucket, you can Block Public Access to the bucket with a single click in the S3 console. You also can drill down into bucket-level permissions settings to configure granular levels of access. For auditing purposes, you can download Access Analyzer for S3 findings as a CSV report.

Additionally, the S3 console reports security warnings, errors, and suggestions from IAM Access Analyzer as you author your S3 policies. The console automatically runs more than 100 policy checks to validate your policies. These checks save you time, guide you to resolve errors, and help you apply security best practices.

For more information, visit the IAM Access Analyzer documentation.

S3 Access Points

Q: What is Amazon S3 Access Points?

Today, customers manage access to their S3 buckets using a single bucket policy that controls access for hundreds of applications with different permission levels.

Amazon S3 Access Points simplifies managing data access at scale for applications using shared data sets on S3. With S3 Access Points, you can now easily create hundreds of access points per bucket, representing a new way of provisioning access to shared data sets. Access Points provide a customized path into a bucket, with a unique hostname and access policy that

enforces the specific permissions and network controls for any request made through the access point. Learn more at the S3 Access Points page and the user guide.

Q: Why should I use an access point?

S3 Access Points simplify how you manage data access for your application set to your shared data sets on S3. You no longer have to manage a single, complex bucket policy with hundreds of different permission rules that need to be written, read, tracked, and audited. With S3 Access Points, you can now create application-specific access points permitting access to shared data sets with policies tailored to the specific application.

Using Access Points, you can decompose one large bucket policy into separate, discrete access point policies for each application that needs to access the shared data set. This makes it simpler to focus on building the right access policy for an application, while not having to worry about disrupting what any other application is doing within the shared data set. You can also create a Service Control Policy (SCP) and require that all access points be restricted to a Virtual Private Cloud (VPC), firewalling your data to within your private networks.

Q: How do S3 Access Points work?

Each S3 Access Point is configured with an access policy specific to a use case or application, and a bucket can have hundreds of access points. For example, you can create an access point for your S3 bucket that grants access for groups of users or applications for your data lake. An Access Point can support a single user or application, or groups of users or applications within and across accounts, allowing separate management of each access point.

Each access point is associated with a single bucket and contains a network origin control, and a Block Public Access control. You can create an access point with a network origin control that only permits storage access from your Virtual Private Cloud, a logically isolated section of the AWS cloud. You can also create an access point with the access point policy configured to only allow access to objects with defined prefixes or to objects with specific tags.

You can access data in shared buckets through an access point in one of two ways. For S3 object operations, you can use the access point ARN in place of a bucket name. For requests requiring a bucket name in the standard S3 bucket name format, you can use an access point alias instead. Aliases for S3 Access Points are automatically generated and are interchangeable with S3 bucket names anywhere you use a bucket name for data access. Every time you create an access point for a bucket, S3 automatically generates a new Access Point Alias. For the full set of compatible operations and AWS services, visit the S3 Documentation.

Q: Is there a quota on how many access points I can create?

By default, you can create 10,000 access points per region per account. Unlike S3 buckets, there is no hard limit on the number of access points per AWS account. Please visit AWS Service Quotas to request an increase in this quota.

Q: When using an access point, how are requests authorized?

S3 access points have their own IAM access point policy. You write access point policies like you would a bucket policy, using the access point ARN as the resource. Access point policies can grant or restrict access to the S3 data requested through the access point. Amazon S3 evaluates all the relevant policies, including those on the user, bucket, access point, VPC

Endpoint, and service control policies as well as Access Control Lists, to decide whether to authorize the request.

Q: How do I write access point policies?

You can write an access point policy just like a bucket policy, using IAM rules to govern permissions and the access point ARN in the policy document.

Q: How is restricting access to specific VPCs using network origin controls on access points different from restricting access to VPCs using the bucket policy?

You can continue to use bucket policies to limit bucket access to specified VPCs. Access points provide an easier, auditable way to lock down all or a subset of data in a shared data set to VPC-only traffic for all applications in your organization using API controls. You can use an AWS Organizations Service Control Policy (SCP) to mandate that any access point created in your organization set the “network origin control” API parameter value to “vpc”. Then, any new access point created automatically restricts data access to VPC-only traffic. No additional access policy is required to make sure that data requests are processed only from specified VPCs.

Q: Can I enforce a “No internet data access” policy for all access points in my organization?

Yes. To enforce a “No internet data access” policy for access points in your organization, you would want to make sure all access points enforce VPC only access. To do so, you will write an AWS SCP that only supports the value “vpc” for the “network origin control” parameter in the create_access_point() API. If you had any internet-facing access points that you created previously, they can be removed. You will also need to modify the bucket policy in each of your buckets to further restrict internet access directly to your bucket through the bucket hostname. Since other AWS services may be directly accessing your bucket, make sure you set up access to allow the AWS services you want by modifying the policy to permit these AWS services. Refer to the S3 documentation for examples of how to do this.

Q: Can I completely disable direct access to a bucket using the bucket hostname?

Not currently, but you can attach a bucket policy that rejects requests not made using an access point. Refer to the S3 Documentation for more details.

Q: Can I replace or remove an access point from a bucket?

Yes. When you remove an access point, any access to the associated bucket through other access points, and through the bucket hostname, will not be disrupted.

Q: What is the cost of Amazon S3 Access Points?

There is no additional charge for access points or buckets that use access points. Usual Amazon S3 request rates apply.

Q: How do I get started with S3 Access Points?

You can start creating S3 Access Points on new buckets as well as your existing buckets through the AWS Management Console, the AWS Command Line Interface (CLI), the Application Programming Interface (API), and the AWS Software Development Kit (SDK) client. To learn more about S3 Access Points, visit the user guide.

Durability & Data Protection

Q: How durable is Amazon S3?

Amazon S3 Standard, S3 Standard–IA, S3 Intelligent-Tiering, S3 One Zone-IA, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive are all designed to provide 99.999999999% (11 9's) of data durability of objects over a given year. This durability level corresponds to an average annual expected loss of 0.000000001% of objects. For example, if you store 10,000,000 objects with Amazon S3, you can on average expect to incur a loss of a single object once every 10,000 years. S3 on Outposts is designed to durably and redundantly store data across multiple devices and servers on your Outpost. In addition, Amazon S3 Standard, S3 Standard-IA, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive are all designed to sustain data in the event of an entire S3 Availability Zone loss.

As with any environment, the best practice is to have a backup and to put in place safeguards against malicious or accidental deletion. For S3 data, that best practice includes secure access permissions, Cross-Region Replication, versioning, and a functioning, regularly tested backup.

Q: How is Amazon S3 designed to achieve 99.999999999% durability?

Amazon S3 Standard, S3 Standard-IA, S3 Intelligent-Tiering, and S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive storage classes redundantly store your objects on multiple devices across a minimum of three Availability Zones (AZs) in an Amazon S3 Region before returning SUCCESS. The S3 One Zone-IA storage class stores data redundantly across multiple devices within a single AZ. These services are designed to sustain concurrent device failures by quickly detecting and repairing any lost redundancy, and they also regularly verify the integrity of your data using checksums.

Q: What checksums does Amazon S3 support for data integrity checking?

Amazon S3 uses a combination of Content-MD5 checksums, secure hash algorithms (SHAs), and cyclic redundancy checks (CRCs) to verify data integrity. Amazon S3 performs these checksums on data at rest and repairs any disparity using redundant data. In addition, S3 calculates checksums on all network traffic to detect alterations of data packets when storing or retrieving data. You can choose from four supported checksum algorithms for data integrity checking on your upload and download requests. You can choose a SHA-1, SHA-256, CRC32, or CRC32C checksum algorithm, depending on your application needs. You can automatically calculate and verify checksums as you store or retrieve data from S3, and can access the checksum information at any time using the GetObjectAttributes S3 API or an S3 Inventory report. Calculating a checksum as you stream data into S3 saves you time as you’re able to both verify and transmit your data in a single pass, instead of as two sequential operations. Using checksums for data validation is a best practice for data durability, and these capabilities increase the performance and reduce the cost to do so.

Q: What is Versioning?

Versioning allows you to preserve, retrieve, and restore every version of every object stored in an Amazon S3 bucket. Once you enable Versioning for a bucket, Amazon S3 preserves existing objects anytime you perform a PUT, POST, COPY, or DELETE operation on them. By default, GET requests will retrieve the most recently written version. Older versions of an overwritten or deleted object can be retrieved by specifying a version in the request.

Q: Why should I use Versioning?

Amazon S3 provides customers with a highly durable storage infrastructure. Versioning offers an additional level of protection by providing a means of recovery when customers accidentally overwrite or delete objects. This allows you to easily recover from unintended user actions and application failures. You can also use Versioning for data retention and archiving.

Q: How do I start using Versioning?

You can start using Versioning by enabling a setting on your Amazon S3 bucket. For more information on how to enable Versioning, please refer to the Amazon S3 Technical Documentation.

Q: How does Versioning protect me from accidental deletion of my objects?

When a user performs a DELETE operation on an object, subsequent simple (un-versioned) requests will no longer retrieve the object. However, all versions of that object will continue to be preserved in your Amazon S3 bucket and can be retrieved or restored. Only the owner of an Amazon S3 bucket can permanently delete a version. You can set Lifecycle rules to manage the lifetime and the cost of storing multiple versions of your objects.

Q: Can I set up a trash, recycle bin, or rollback window on my Amazon S3 objects to recover from deletes and overwrites?

You can use Amazon S3 Lifecycle rules along with S3 Versioning to implement a rollback window for your S3 objects. For example, with your versioning-enabled bucket, you can set up a rule that archives all of your previous versions to the lower-cost S3 Glacier Flexible Retrieval storage class and deletes them after 100 days, giving you a 100-day window to roll back any changes on your data while lowering your storage costs. Additionally, you can save costs by deleting old (noncurrent) versions of an object after five days and when there are at least two newer versions of the object. You can change the number of days or the number of newer versions based on your cost optimization needs. This allows you to retain additional versions of your objects when needed, but saves you cost by transitioning or removing them after a period of time.

Q: How can I ensure maximum protection of my preserved versions?

Versioning's Multi-Factor Authentication (MFA) Delete capability can be used to provide an additional layer of security. By default, all requests to your Amazon S3 bucket require your AWS account credentials. If you enable Versioning with MFA Delete on your Amazon S3 bucket, two forms of authentication are required to permanently delete a version of an object: your AWS account credentials and a valid six-digit code and serial number from an authentication device in your physical possession. To learn more about enabling Versioning with MFA Delete, including how to purchase and activate an authentication device, please refer to the Amazon S3 Technical Documentation.

Q: How am I charged for using Versioning?

Normal Amazon S3 rates apply for every version of an object stored or requested. For example, let's look at the following scenario to illustrate storage costs when utilizing Versioning (let's assume the current month is 31 days long):

1) Day 1 of the month: You perform a PUT of 4 GB (4,294,967,296 bytes) on your bucket.
2) Day 16 of the month: You perform a PUT of 5 GB (5,368,709,120 bytes) within the same bucket using the same key as the original PUT on Day 1.

When analyzing the storage costs of the above operations, please note that the 4 GB object from Day 1 is not deleted from the bucket when the 5 GB object is written on Day 15. Instead, the 4 GB object is preserved as an older version and the 5 GB object becomes the most recently written version of the object within your bucket. At the end of the month:

Total Byte-Hour usage
[4,294,967,296 bytes x 31 days x (24 hours / day)] + [5,368,709,120 bytes x 16 days x (24 hours / day)] = 5,257,039,970,304 Byte-Hours.

Conversion to Total GB-Months
5,257,039,970,304 Byte-Hours x (1 GB / 1,073,741,824 bytes) x (1 month / 744 hours) = 6.581 GB-Month

The cost is calculated based on the current rates for your region on the [Amazon S3 Pricing Page](#).

Storage Classes

Q: What are the Amazon S3 storage classes?

Amazon S3 offers a range of storage classes that you can choose from based on the data access, resiliency, and cost requirements of your workloads. S3 storage classes are purpose-built to provide the lowest cost storage for different access patterns. S3 storage classes are ideal for virtually any use case, including those with demanding performance needs, data residency requirements, unknown or changing access patterns, or archival storage. Each S3 storage class charges a fee to store data and fees to access data. In deciding which S3 storage class best fits your workload, consider the access patterns and retention time of your data to optimize for the lowest total cost over the lifetime of your data.

S3 Storage Classes can be configured at the object level and a single bucket can contain objects stored across all of the storage classes. You can also use S3 Lifecycle policies to automatically transition objects between storage classes without any application changes.

Q: How do I decide which S3 storage class to use?

In deciding which S3 storage class best fits your workload, consider the access patterns and retention time of your data to optimize for the lowest total cost over the lifetime of your data. Many workloads have changing (user-generated content), unpredictable (analytics, data lakes), or unknown (new applications) access patterns, and that is why S3 Intelligent-Tiering should be the default storage class to automatically save on storage costs. If you know the access patterns of your data, you can follow this guidance. The S3 Standard storage class is ideal for frequently accessed data; this is the best choice if you access data more than once a month. S3 Standard-Infrequent Access is ideal for data retained for at least a month and accessed once every month or two.

The Amazon S3 Glacier storage classes are purpose-built for data archiving, providing you with the highest performance, most retrieval flexibility, and the lowest cost archive storage in the cloud. You can now choose from three archive storage classes optimized for different access patterns and storage duration. For archive data that needs immediate access, such as medical images, news media assets, or genomics data, choose the S3 Glacier Instant Retrieval storage class, an archive storage class that delivers the lowest cost storage with milliseconds retrieval. For archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases, choose S3 Glacier Flexible Retrieval (formerly S3 Glacier), with retrieval in minutes or free bulk retrievals in 5—12 hours. To save even more on long-lived archive storage such as compliance archives and digital media preservation, choose S3 Glacier Deep Archive, the lowest cost storage in the cloud with data retrieval within 12 hours. All these storage classes provide multi-Availability Zone (AZ) resiliency by redundantly storing data on multiple devices and physically separated AWS Availability Zones in an AWS Region.

For data that has a lower resiliency requirement, you can reduce costs by selecting a single-AZ storage class, like S3 One Zone-Infrequent Access. If you have data residency or latency requirements that can't be met by an existing AWS Region, you can choose S3 on Outposts to store data on-premises.

S3 Intelligent-Tiering

Q: What is S3 Intelligent-Tiering?

S3 Intelligent-Tiering is the first cloud storage that automatically reduces your storage costs on a granular object level by automatically moving data to the most cost-effective access tier based on access frequency, without performance impact, retrieval fees, or operational overhead. S3 Intelligent-Tiering delivers milliseconds latency and high throughput performance for frequently, infrequently, and rarely accessed data in the Frequent, Infrequent, and Archive Instant Access tiers. For a small monthly object monitoring and automation charge, S3 Intelligent-Tiering monitors the access patterns and moves the objects automatically from one tier to another. There are no retrieval charges in S3 Intelligent-Tiering, so you won't see unexpected increases in storage bills when access patterns change.

You can use S3 Intelligent-Tiering as the default storage class for virtually any workload, especially data lakes, data analytics, machine learning, new applications, and user-generated content.

Q: How does S3 Intelligent-Tiering work?

The Amazon S3 Intelligent-Tiering storage class is designed to optimize storage costs by automatically moving data to the most cost-effective access tier when access patterns change. For a low monthly object monitoring and automation charge, S3 Intelligent-Tiering monitors access patterns and automatically moves objects that have not been accessed for 30 consecutive days to the Infrequent Access tier to save up to 40% on storage costs. After 90 days consecutive days of no access, objects are moved to the Archive Instant Access tier to save up to 68% on storage costs. There is no impact on performance and there are no retrieval charges in S3 Intelligent-Tiering. If an object in the Infrequent Access tier or Archive Instant Access tier is accessed later, it is automatically moved back to the Frequent Access tier.

To get the lowest storage cost on data that can be accessed asynchronously, you can choose to activate additional archiving capabilities. Once you enable one or both of the asynchronous archive access tiers, S3 Intelligent-Tiering will move objects that have not been accessed for a minimum of 90 days to the Archive Access tier to save up to 71% and after 180 days of no

access to the Deep Archive Access tier to save up to 95% for rarely accessed objects. If an object in the optional Archive or Deep Access tiers is restored later, it is moved back to the Frequent Access tier, and before you can retrieve the object you must first restore the object using RestoreObject. For information about restoring archived objects, see Restoring Archived Objects. There are no retrieval charges in S3 Intelligent-Tiering. No additional tiering or lifecycle charges apply when objects are moved between access tiers within the S3 Intelligent-Tiering storage class.

There is no minimum object size for S3 Intelligent-Tiering, but objects smaller than 128KB are not eligible for auto-tiering. These smaller objects may be stored in S3 Intelligent-Tiering, but will always be charged at the Frequent Access tier rates, and are not charged the monitoring and automation charge.

If you would like to standardize on S3 Intelligent-Tiering as the default storage class for newly created data, you can modify your applications by specifying INTELLIGENT-TIERING on your S3 PUT API request header. S3 Intelligent-Tiering is designed for 99.9% availability and 99.999999999% durability, and automatically offers the same low latency and high throughput performance of S3 Standard. You can use AWS Cost Explorer to measure the additional savings from the Archive Instant Access tier.

Q: Why would I choose to use S3 Intelligent-Tiering?

You can use S3 Intelligent-Tiering as the default storage class for virtually any workload, especially data lakes, data analytics, machine learning, new applications, and user-generated content. S3 Intelligent-Tiering is the first cloud storage that automatically reduces your storage costs on a granular object level by automatically moving data to the most cost-effective access tier based on access frequency, without performance impact, retrieval fees, or operational overhead. If you have data with unknown or changing access patterns, including data lakes, data analytics, and new applications, we recommend using S3 Intelligent-Tiering. If you have data that does not require immediate retrieval, we recommend activating the Deep Archive Access tier where you pay as little as $1 per TB per month for data that may become rarely accessed over long periods of time. S3 Intelligent-Tiering is for data with unknown or changing access patterns. There are no retrieval fees when using the S3 Intelligent-Tiering storage class.

Q: What performance does S3 Intelligent-Tiering offer?

S3 Intelligent-Tiering automatically optimizes your storage costs without an impact to your performance. The S3 Intelligent-Tiering Frequent, Infrequent, and Archive Instant Access tiers provide milliseconds latency and high throughput performance.

Q: What performance do the optional Archive Access and Deep Archive Access tiers provide?

For data that can be accessed asynchronously, the optional Archive Access tier has the same performance as S3 Glacier Flexible Retrieval, and the Deep Archive Access tier has the same performance as the S3 Glacier Deep Archive storage class. You should only activate the asynchronous archive capabilities if your application can wait minutes to hours. If the object you are retrieving is stored in the Archive or Deep Archive Access tiers, before you can retrieve the object you must first restore an object using RestoreObject. For information about restoring archived objects, see Restoring Archived Objects. Objects in the Archive Access tier are moved to the Frequent Access tier in 3—5 hours and within 12 hours if they are in the Deep Archive Access tier. If you need faster access to an object in the Archive Access tier, you can pay for faster retrieval by using the console to select the expedited retrieval speed option.

Q: How durable and available is S3 Intelligent-Tiering?

S3 Intelligent-Tiering is designed for the same 99.999999999% durability as the S3 Standard storage class. S3 Intelligent-Tiering is designed for 99.9% availability, and carries a service level agreement providing service credits if availability is less than our service commitment in any billing cycle.

Q: How do I get my data into S3 Intelligent-Tiering?

There are two ways to get data into S3 Intelligent-Tiering. You can directly PUT into S3 Intelligent-Tiering by specifying INTELLIGENT_TIERING in the x-amz-storage-class header or set lifecycle policies to transition objects from S3 Standard or S3 Standard-IA to S3 INTELLIGENT_TIERING.

Q: How am I charged for S3 Intelligent-Tiering?

S3 Intelligent-Tiering charges you for monthly storage, requests, and data transfer, and charges a small monthly charge for monitoring and automation per object. The S3 Intelligent-Tiering storage class automatically stores objects in three access tiers: a Frequent Access tier priced at S3 Standard storage rates, an Infrequent Access tier priced at S3 Standard-Infrequent Access storage rates, and an Archive Instant Access tier priced at the S3 Glacier Instant Retrieval storage rates. S3 Intelligent-Tiering also has two optional archive tiers designed for asynchronous access, an Archive Access tier priced at S3 Glacier Flexible Retrieval storage rates, and a Deep Archive Access tier priced at S3 Glacier Deep Archive storage rates.

For a small monitoring and automation fee, S3 Intelligent-Tiering monitors access patterns and automatically moves objects through low latency and high throughput access tiers, as well as two opt in asynchronous archive access tiers where customers get the lowest storage costs in the cloud for data that can be accessed asynchronously.

There is no minimum billable object size in S3 Intelligent-Tiering, but objects smaller than 128KB are not eligible for auto-tiering. These small objects will not be monitored and will always be charged at the Frequent Access tier rates, with no monitoring and automation charge. For each object archived to the Archive Access tier or Deep Archive Access tier in S3 Intelligent-Tiering, Amazon S3 uses 8 KB of storage for the name of the object and other metadata (billed at S3 Standard storage rates) and 32 KB of storage for index and related metadata (billed at S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive storage rates).

Q: Is there a charge to retrieve data from S3 Intelligent-Tiering?

No. There are no retrieval fees for S3 Intelligent-Tiering. S3 Intelligent-Tiering monitors the access patterns of your data and if you access an object in the Infrequent Access, Archive Instant Access, or the asynchronous archive tiers, S3 Intelligent-Tiering automatically moves that object to the Frequent Access tier.

Q: How do I activate S3 Intelligent-Tiering archive access tiers?

You can activate the Archive Access tier and Deep Archive Access tier by creating a bucket, prefix, or object tag level configuration using the Amazon S3 API, CLI, or S3 management console. You should only activate one or both of the archive access tiers if your objects can be accessed asynchronously by your application.

Q: Can I extend the time before objects are archived within S3 Intelligent-Tiering storage class?

Yes. In the bucket, prefix, or object tag level configuration, you can extend the last access time for archiving objects in S3 Intelligent-Tiering. When enabled, by default objects that haven't been accessed for a minimum of 90 consecutive days automatically move to the Archive Access tier, skipping the Archive Instant Access tier. Objects that haven't been accessed for a minimum of 180 consecutive days automatically move to the Deep Archive Access tier. The default configuration for the consecutive days since last access before automatic archiving in S3 Intelligent-Tiering can be extended for up to 2 years.

Q: How do I access an object from the Archive Access or Deep Archive Access tiers in the S3 Intelligent-Tiering storage class?

To access an object in the Archive or Deep Archive Access tiers, you need to issue a Restore request and the object will begin moving back to the Frequent Access tier, all within the S3 Intelligent-Tiering storage class. Objects in the Archive Access Tier are moved to the Frequent Access tier in 3-5 hours, objects in the Deep Archive Access tier are moved to the Frequent Access tier within 12 hours. Once the object is in the Frequent Access tier, you can issue a GET request to retrieve the object.

Q: How do I know in which S3 Intelligent-Tiering access tier my objects are stored in?

You can use Amazon S3 Inventory to report the access tier of objects stored in the S3 Intelligent-Tiering storage class. Amazon S3 Inventory provides CSV, ORC, or Parquet output files that list your objects and their corresponding metadata on a daily or weekly basis for an S3 bucket or a shared prefix. You can also make a HEAD request on your objects to report the S3 Intelligent-Tiering archive access tiers.

Q: Can I lifecycle objects from S3 Intelligent-Tiering to another storage class?

Yes. You can lifecycle objects from S3 Intelligent-Tiering Frequent Access, Infrequent, and Archive Instant Access tiers to S3 One-Zone Infrequent Access, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive. In addition, you can lifecycle objects from the S3 Intelligent-Tiering optional archive access tiers to S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive, and from the S3 Intelligent-Tiering Deep Archive Access tier to S3 Glacier Deep Archive.

Q: Is there a minimum duration for S3 Intelligent-Tiering?

No. The S3 Intelligent-Tiering storage class has no minimum storage duration.

Q: Is there a minimum billable object size for S3 Intelligent-Tiering?

No. The S3 Intelligent-Tiering storage class has no minimum billable object size, but objects smaller than 128KB are not eligible for auto-tiering. These smaller objects will always be charged at the Frequent Access tier rates, with no monitoring and automation charge. For each object archived to the opt-in Archive Access tier or Deep Archive Access tier in S3 Intelligent-Tiering, Amazon S3 uses 8 KB of storage for the name of the object and other metadata (billed at S3 Standard storage rates) and 32 KB of storage for index and related metadata (billed at S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive storage rates).

S3 Standard

Q: What is S3 Standard?

Amazon S3 Standard delivers durable storage with millisecond access latency and high throughput performance for frequently accessed data, typically more than once per month. S3 Standard is designed for performance-sensitive uses cases, such as data lakes, cloud-native applications, dynamic websites, content distribution, mobile and gaming applications, analytics, and machine learning models. S3 Standard is designed for 99.99% data availability and durability of 99.999999999% of objects across multiple Availability Zones in a given year. You can use S3 Lifecycle policies to control exactly when data is transitioned between S3 Standard and lower costs storage classes without any application changes.

Q: Why would I choose to use S3 Standard?

S3 Standard is ideal for your most frequently accessed or modified data that requires access in milliseconds and high throughput performance. S3 Standard is ideal for data that is read or written very often, as there are no retrieval charges. S3 Standard is optimized for a wide variety of use cases, including data lakes, cloud native applications, dynamic websites, content distribution, mobile and gaming applications, and analytics.

S3 Standard-Infrequent Access (S3 Standard-IA)

Q: What is S3 Standard-Infrequent Access?

Amazon S3 Standard-Infrequent Access (S3 Standard-IA) is an Amazon S3 storage class for data that is accessed less frequently but requires rapid access when needed. S3 Standard-IA offers the high durability, throughput, and low latency of the Amazon S3 Standard storage class, with a low per-GB storage price and per-GB retrieval charge. This combination of low cost and high performance make S3 Standard-IA ideal for long-term storage, backups, and as a data store for disaster recovery. The S3 Standard-IA storage class is set at the object level and can exist in the same bucket as the S3 Standard or S3 One Zone-IA storage classes, allowing you to use S3 Lifecycle policies to automatically transition objects between storage classes without any application changes.

Q: Why would I choose to use S3 Standard-IA?

S3 Standard-IA is ideal for data that is accessed less frequently, but requires rapid access when needed. S3 Standard-IA is ideally suited for long-term file storage, older sync and share storage, and other aging data.

Q: What performance does S3 Standard-IA offer?

S3 Standard-IA provides the same milliseconds latency and high throughput performance as the S3 Standard storage class.

Q: How do I get my data into S3 Standard-IA?

There are two ways to get data into S3 Standard-IA. You can directly PUT into S3 Standard-IA by specifying STANDARD_IA in the x-amz-storage-class header. You can also set Lifecycle policies to transition objects from the S3 Standard to the S3 Standard-IA storage class.

Q: What charges will I incur if I change the storage class of an object from S3 Standard-IA to S3 Standard with a COPY request?

You will incur charges for an S3 Standard-IA COPY request and an S3 Standard-IA data retrieval.

Q: Is there a minimum storage duration charge for S3 Standard-IA?

S3 Standard-IA is designed for long-lived, infrequently accessed data that is retained for months or years. Data that is deleted from S3 Standard-IA within 30 days will be charged for a full 30 days.

Q: Is there a minimum object storage charge for S3 Standard-IA?

S3 Standard-IA is designed for larger objects and has a minimum object storage charge of 128KB. Objects smaller than 128KB in size will incur storage charges as if the object were 128KB. For example, a 6KB object in S3 Standard-IA will incur S3 Standard-IA storage charges for 6KB and an additional minimum object size charge equivalent to 122KB at the S3 Standard-IA storage price.

Q: Can I tier objects from S3 Standard-IA to S3 One Zone-IA or to the S3 Glacier Flexible Retrieval storage class?

Yes. In addition to using Lifecycle policies to migrate objects from S3 Standard to S3 Standard-IA, you can also set up Lifecycle policies to tier objects from S3 Standard-IA to S3 One Zone-IA, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and the S3 Glacier Deep Archive storage class.

S3 One Zone-Infrequent Access (S3 One Zone-IA)

Q: What is S3 One Zone-IA storage class?

S3 One Zone-IA storage class is an Amazon S3 storage class that customers can choose to store objects in a single availability zone. S3 One Zone-IA storage redundantly stores data within that single Availability Zone to deliver storage at 20% less cost than geographically redundant S3 Standard-IA storage, which stores data redundantly across multiple geographically separate Availability Zones.

S3 One Zone-IA offers a 99% available SLA and is also designed for eleven 9's of durability within the Availability Zone. However, the data in the S3 One Zone-IA storage class is not resilient to the physical loss of an entire Availability Zone.

S3 One Zone-IA storage offers the same Amazon S3 features as S3 Standard and S3 Standard-IA and is used through the Amazon S3 API, CLI and console. S3 One Zone-IA storage class is set at the object level and can exist in the same bucket as S3 Standard and S3 Standard-IA storage classes. You can use S3 Lifecycle policies to automatically transition objects between storage classes without any application changes.

Q: What use cases are best suited for S3 One Zone-IA storage class?

Customers can use S3 One Zone-IA for infrequently-accessed storage, like backup copies, disaster recovery copies, or other easily re-creatable data.

Q: What performance does S3 One Zone-IA storage offer?

S3 One Zone-IA storage class offers the same latency and throughput performance as the S3 Standard and S3 Standard-Infrequent Access storage classes.

Q: How durable is the S3 One Zone-IA storage class?

S3 One Zone-IA storage class is designed for 99.999999999% of durability within an Availability Zone. However, data in the S3 One Zone-IA storage class is not resilient to the loss of availability or physical loss of an Availability Zone. In contrast, S3 Standard, S3 Intelligent-Tiering, S3 Standard-Infrequent Access, and the S3 Glacier storage classes are designed to withstand loss of availability or the destruction of an Availability Zone. S3 One Zone-IA can deliver the same or better durability and availability than most modern, physical data centers, while providing the added benefit of elasticity of storage and the Amazon S3 feature set.

Q: Is an S3 One Zone-IA “Zone” the same thing as an AWS Availability Zone?

Yes. Each AWS Region is a separate geographic area. Each Region has multiple, isolated locations known as Availability Zones. The Amazon S3 One Zone-IA storage class uses an individual AWS Availability Zone within the Region.

Q: How much disaster recovery protection do I forgo by using S3 One Zone-IA?

Each Availability Zone uses redundant power and networking. Within an AWS Region, Availability Zones are on different flood plains, earthquake fault zones, and geographically separated for fire protection. S3 Standard and S3 Standard-IA storage classes offer protection against these sorts of disasters by storing your data redundantly in multiple Availability Zones. S3 One Zone-IA offers protection against equipment failure within an Availability Zone, but the data is not resilient to the physical loss of the Availability Zone resulting from disasters, such as earthquakes and floods. Using S3 One Zone-IA, S3 Standard, and S3 Standard-IA options, you can choose the storage class that best fits the durability and availability needs of your storage.

Amazon S3 Glacier Instant Retrieval storage class

Q: What is the S3 Glacier Instant Retrieval storage class?

The S3 Glacier Instant Retrieval storage class delivers the lowest cost storage for long-lived data that is rarely accessed and requires milliseconds retrieval. S3 Glacier Instant Retrieval delivers the fastest access to archive storage, with the same throughput and milliseconds access as S3 Standard and S3 Standard-IA storage classes. S3 Glacier Instant Retrieval is designed for 99.999999999% (11 9s) of data durability and 99.9% availability by redundantly storing data across a minimum of three physically separated AWS Availability Zones.

Q: Why would I choose to use S3 Glacier Instant Retrieval?

S3 Glacier Instant Retrieval is ideal if you have data that is rarely accessed (once a quarter) and requires milliseconds retrieval times. It’s the ideal storage class if you want the same low latency and high throughput performance as S3 Standard-IA, but store data that is accessed less frequently than S3 Standard-IA, with a lower storage price and slightly higher data access costs.

Q: How available and durable is S3 Glacier Instant Retrieval?

S3 Glacier Instant Retrieval is designed for 99.999999999% (11 9s) of durability and 99.9% availability, the same as S3 Standard-IA, and carries a service level agreement providing service credits if availability is less than 99% in any billing cycle.

Q: What performance does S3 Glacier Instant Retrieval offer?

S3 Glacier Instant Retrieval provides the same milliseconds latency and high throughput performance as the S3 Standard and S3 Standard-IA storage classes. Unlike the S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive storage classes, which are designed for asynchronous access, you do not need to issue a Restore request before accessing an object stored in S3 Glacier Instant Retrieval.

Q: How do I get my data into S3 Glacier Instant Retrieval?

There are two ways to get data into S3 Glacier Instant Retrieval. You can directly PUT into S3 Glacier Instant retrieval by specifying GLACIER_IR in the x-amz-storage-class header or set S3 Lifecycle policies to transition objects from S3 Standard or S3 Standard-IA to S3 Glacier Instant Retrieval.

Q: Is there a minimum storage duration charge for Amazon S3 Glacier Instant Retrieval?

S3 Glacier Instant Retrieval is designed for long-lived, rarely accessed data that is retained for months or years. Objects that are archived to S3 Glacier Instant Retrieval have a minimum of 90 days of storage, and objects deleted, overwritten, or transitioned before 90 days incur a pro-rated charge equal to the storage charge for the remaining days.

Q: Is there a minimum object size charge for Amazon S3 Glacier Instant Retrieval?

S3 Glacier Instant Retrieval is designed for larger objects and has a minimum object storage charge of 128KB. Objects smaller than 128KB in size will incur storage charges as if the object were 128KB. For example, a 6KB object in S3 Glacier Instant Retrieval will incur S3 Glacier Instant Retrieval storage charges for 6KB and an additional minimum object size charge equivalent to 122KB at the S3 Glacier Instant Retrieval storage price.

Q: How am I charged for S3 Glacier Instant Retrieval?

S3 Glacier Instant Retrieval charges you for monthly storage, requests based on the request type, and data retrievals. The volume of storage billed in a month is based on average storage used throughout the month, measured in gigabyte per month (GB-Month). You are charged for requests based on the request type—such as PUTs, COPYs, and GETs. You also pay a per GB fee for every gigabyte of data returned to you.

Amazon S3 Glacier Flexible Retrieval storage class (Formerly S3 Glacier)

Q: What is the S3 Glacier Flexible Retrieval storage class?

The S3 Glacier Flexible Retrieval storage class delivers low-cost storage, up to 10% lower cost (than S3 Glacier Instant Retrieval), for archive data that is accessed 1-2 times per year and is retrieved asynchronously, with free bulk retrievals. For archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases, S3 Glacier Flexible Retrieval (formerly S3 Glacier) is the ideal storage class. S3 Glacier Flexible Retrieval delivers the most flexible retrieval options that balance cost with access times ranging from minutes to hours and with free bulk retrievals. It is an ideal solution for backup, disaster recovery, offsite data storage needs, and for when some data needs to occasionally retrieved in minutes, and you don't want to worry about costs. S3 Glacier Flexible Retrieval is designed for 99.999999999% (11 9s) of data durability and 99.99%

availability by redundantly storing data across multiple physically separated AWS Availability Zones in a given year.

Q: Why would I choose to use S3 Glacier Flexible Retrieval storage class?

For archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases, S3 Glacier Flexible Retrieval (formerly S3 Glacier) is the ideal storage class. S3 Glacier Flexible Retrieval delivers the most flexible retrieval options that balance cost with access times ranging from minutes to hours and with free bulk retrievals. It is an ideal solution for backup, disaster recovery, offsite data storage needs, and for when some data needs to occasionally retrieved in minutes, and you don't want to worry about costs to retrieve the data.

Q: How do I get my into S3 Glacier Flexible Retrieval?

There are two ways to get data into S3 Glacier Flexible Retrieval. You can directly PUT into S3 Glacier Flexible Retrieval by specifying GLACIER in the x-amz-storage-class header. You can also use S3 Lifecycle rules to transition objects from any of the S3 storage classes for active data (S3 Standard, S3 Intelligent-Tiering, S3 Standard-IA, S3 One Zone-IA, and S3 Glacier Instant Retrieval) to Amazon S3 Glacier Flexible Retrieval based on object age. Use the Amazon S3 Management Console, the AWS SDKs, or the Amazon S3 APIs to directly PUT into Amazon S3 Glacier or define rules for archival.

Note: S3 Glacier Flexible Retrieval (Formerly S3 Glacier) is also available through the original direct Glacier APIs and through the Amazon S3 Glacier Management Console. For an enhanced experience complete with access to the full S3 feature set including lifecycle management, S3 Replication, S3 Storage Lens, and more, we recommend using S3 APIs and the S3 Management Console to use S3 Glacier features.
Q: How can I retrieve my objects that are archived in S3 Glacier Flexible Retrieval and will I be notified when the object is restored?

Objects that are archived in S3 Glacier Flexible Retrieval are accessed asynchronously. To retrieve data stored in S3 Glacier Flexible Retrieval, initiate a retrieval request using the Amazon S3 APIs or the Amazon S3 console. The retrieval request creates a temporary copy of your data in the S3 Standard storage class while leaving the archived data intact in S3 Glacier Flexible Retrieval. You can specify the amount of time in days for which the temporary copy is stored in Amazon S3. You can then access your temporary copy from S3 through an Amazon S3 GET request on the archived object.

With restore notifications, you can now be notified with an S3 Event Notification when an object has successfully restored from S3 Glacier Flexible Retrieval and the temporary copy is made available to you. The bucket owner (or others, as permitted by an IAM policy) can arrange for notifications to be issued to Amazon Simple Queue Service (SQS) or Amazon Simple Notification Service (SNS). Notifications can also be delivered to AWS Lambda for processing by a Lambda function.
Q: How long will it take to restore my objects archived in Amazon S3 Glacier Flexible Retrieval?

When processing a retrieval job, Amazon S3 first retrieves the requested data from S3 Glacier Flexible Retrieval, and then creates a temporary copy of the requested data in Amazon S3. This typically takes a few minutes. The access time of your request depends on the retrieval option you choose: Expedited, Standard, or Bulk retrievals. For all but the largest objects (250MB+), data accessed using Expedited retrievals are typically made available within 1-5 minutes. Objects retrieved using Standard retrievals typically complete between 3-5 hours. Bulk retrievals

typically complete within 5—12 hours, and are free of charge. For more information about the S3 Glacier Flexible Retrieval options, refer to restoring an archived object in the S3 user guide.

With S3 Glacier storage class provisioned capacity units, you can pay a fixed upfront fee for a given month to ensure the availability of retrieval capacity for expedited retrievals from S3 Glacier Flexible Retrieval. You can purchase two provisioned capacity units per month to increase the amount of data you can retrieve. Each unit of capacity ensures that at least three expedited retrievals can be performed every five minutes, and it provides up to 150 MB/s of retrieval throughput. If your workload requires highly reliable and predictable access to a subset of your data in minutes, you should purchase provisioned retrieval capacity. Without provisioned capacity, expedited retrievals might not be accepted during periods of high demand. If you require access to expedited retrievals under any circumstance, we recommend that you purchase provisioned retrieval capacity.

You can purchase provisioned capacity using the Amazon S3 console, the purchase provisioned capacity REST API, the AWS SDKs, or the AWS CLI. A provisioned capacity unit lasts for one month starting at the date and time of purchase, which is the start date. The unit expires on the expiration date, which is exactly one month after the start date to the nearest second.

Q: How is my storage charge calculated for Amazon S3 objects archived to S3 Glacier Flexible Retrieval?

The volume of storage billed in a month is based on average storage used throughout the month, measured in gigabyte-months (GB-Months). Amazon S3 calculates the object size as the amount of data you stored, plus an additional 32 KB of S3 Glacier data, plus an additional 8 KB of Amazon S3 Standard storage class data. S3 Glacier Flexible Retrieval requires an additional 32 KB of data per object for S3 Glacier's index and metadata so you can identify and retrieve your data. Amazon S3 requires 8 KB to store and maintain the user-defined name and metadata for objects archived to S3 Glacier Flexible Retrieval. This enables you to get a real-time list of all of your Amazon S3 objects, including those stored using S3 Glacier Flexible Retrieval, using the Amazon S3 LIST API, or the S3 inventory report.

For example, if you have archived 100,000 objects that are 1 GB each, your billable storage would be:
1.000032 gigabytes for each object x 100,000 objects = 100,003.2 gigabytes of S3 Glacier storage.
0.000008 gigabytes for each object x 100,000 objects = 0.8 gigabytes of S3 Standard storage.

The fee is calculated based on the current rates for your AWS Region on the Amazon S3 pricing page.

Q: Are there minimum storage duration and minimum object storage charges for Amazon S3 Glacier Flexible Retrieval?

Amazon S3 Glacier Flexible Retrieval is designed for long-lived but rarely accessed data that is retained for months or years. Objects that are archived to S3 have a minimum of 90 days of storage, and objects deleted, overwritten, or transitioned before 90 days incur a pro-rated charge equal to the storage charge for the remaining days.

Amazon S3 Glacier Flexible Retrieval has a minimum billable object storage size of 40 KB. Objects smaller than 40 KB in size may be stored but will be charged for 40 KB of storage.

Q: How much does it cost to retrieve data from Amazon S3 Glacier Flexible Retrieval?

There are three ways to retrieve data from S3 Glacier Flexible Retrieval: Expedited, Standard, and Bulk Retrievals. Expedited and Standard have a per-GB retrieval fee and per-request fee (i.e., you pay for requests made against your Amazon S3 objects). Bulk Retrievals from S3 Glacier Flexible Retrieval are free.

Q: Does Amazon S3 provide capabilities for archiving objects to lower cost storage classes?

The Amazon S3 Glacier storage classes are purpose-built for data archiving, providing you with the highest performance, most retrieval flexibility, and the lowest cost archive storage in the cloud. You can now choose from three archive storage classes optimized for different access patterns and storage duration. For archive data that needs immediate access, such as medical images, news media assets, or genomics data, choose the S3 Glacier Instant Retrieval storage class, an archive storage class that delivers the lowest cost storage with milliseconds retrieval. For archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases, choose S3 Glacier Flexible Retrieval (formerly S3 Glacier), with retrieval in minutes or free bulk retrievals in 5—12 hours. To save even more on long-lived archive storage such as compliance archives and digital media preservation, choose S3 Glacier Deep Archive, the lowest cost storage in the cloud with data retrieval within 12 hours.

Q: What is the backend infrastructure supporting the S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive storage class?

We prefer to focus on the customer outcomes of performance, durability, availability, and security. However, this question is often asked by our customers. We use a number of different technologies which allow us to offer the prices we do to our customers. Our services are built using common data storage technologies specifically assembled into purpose-built, cost-optimized systems using AWS-developed software. The S3 Glacier storage classes benefit from our ability to optimize the sequence of inputs and outputs to maximize efficiency accessing the underlying storage.

Amazon S3 Glacier Deep Archive

Q: What is the Amazon S3 Glacier Deep Archive storage class?

S3 Glacier Deep Archive is an Amazon S3 storage class that provides secure and durable object storage for long-term retention of data that is accessed once or twice in a year. From just $0.00099 per GB-month (less than one-tenth of one cent, or about $1 per TB-month), S3 Glacier Deep Archive offers the lowest cost storage in the cloud, at prices significantly lower than storing and maintaining data in on-premises magnetic tape libraries or archiving data off-site.

Q: What use cases are best suited for the S3 Glacier Deep Archive storage class?

S3 Glacier Deep Archive is an ideal storage class to provide offline protection of your company's most important data assets, or when long-term data retention is required for corporate policy, contractual, or regulatory compliance requirements. Customers find S3 Glacier Deep Archive to be a compelling choice to protect core intellectual property, financial and medical records, research results, legal documents, seismic exploration studies, and long-term backups, especially in highly regulated industries, such as Financial Services, Healthcare, Oil & Gas, and Public Sectors. In addition, there are organizations, such as media and entertainment

companies, that want to keep a backup copy of core intellectual property. Frequently, customers using S3 Glacier Deep Archive can reduce or discontinue the use of on-premises magnetic tape libraries and off-premises tape archival services.

Q: How does the S3 Glacier Deep Archive storage class differ from the S3 Glacier Instant Retrieval, and S3 Glacier Flexible Retrieval storage classes?

S3 Glacier Deep Archive expands our data archiving offerings, enabling you to select the optimal storage class based on storage and retrieval costs, and retrieval times. Choose the S3 Glacier Instant Retrieval storage class when you need milliseconds access to low cost archive data. For archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases, choose S3 Glacier Flexible Retrieval (formerly S3 Glacier), with retrieval in minutes or free bulk retrievals in 5-12 hours. S3 Glacier Deep Archive, in contrast, is designed for colder data that is very unlikely to be accessed, but still requires long-term, durable storage. S3 Glacier Deep Archive is up to 75% less expensive than S3 Glacier Flexible Retrieval and provides retrieval within 12 hours using the Standard retrieval speed. You may also reduce retrieval costs by selecting Bulk retrieval, which will return data within 48 hours.

Q: How do I get started using S3 Glacier Deep Archive?

The easiest way to store data in S3 Glacier Deep Archive is to use the S3 API to upload data directly. Just specify “S3 Glacier Deep Archive” as the storage class. You can accomplish this using the AWS Management Console, S3 REST API, AWS SDKs, or AWS Command Line Interface.

You can also begin using S3 Glacier Deep Archive by creating policies to migrate data using S3 Lifecycle, which provides the ability to define the lifecycle of your object and reduce your cost of storage. These policies can be set to migrate objects to S3 Glacier Deep Archive based on the age of the object. You can specify the policy for an S3 bucket, or for specific prefixes. Lifecycle transitions are billed at the S3 Glacier Deep Archive Upload price.

Tape Gateway, a cloud-based virtual tape library feature of AWS Storage Gateway, now integrates with S3 Glacier Deep Archive, enabling you to store your virtual tape-based, long-term backups and archives in S3 Glacier Deep Archive, thereby providing the lowest cost storage for this data in the cloud. To get started, create a new virtual tape using AWS Storage Gateway Console or API, and set the archival storage target either to S3 Glacier Flexible Retrieval or S3 Glacier Deep Archive. When your backup application ejects the tape, the tape will be archived to your selected storage target.

Q: How do you recommend migrating data from my existing tape archives to S3 Glacier Deep Archive?

There are multiple ways to migrate data from existing tape archives to S3 Glacier Deep Archive. You can use the AWS Tape Gateway to integrate with existing backup applications using a virtual tape library (VTL) interface. This interface presents virtual tapes to the backup application. These can be immediately used to store data in Amazon S3, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive.

You can also use AWS Snowball or Snowmobile to migrate data. Snowball and Snowmobile accelerate moving terabytes to petabytes of data into and out of AWS using physical storage devices designed to be secure for transport. Using Snowball and Snowmobile helps to eliminate

challenges that can be encountered with large-scale data transfers including high network costs, long transfer times, and security concerns.

Finally, you can use AWS Direct Connect to establish dedicated network connections from your premises to AWS. In many cases, Direct Connect can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than internet-based connections.

Q: How can I retrieve my objects stored in S3 Glacier Deep Archive?

To retrieve data stored in S3 Glacier Deep Archive, initiate a “Restore” request using the Amazon S3 APIs or the Amazon S3 Management Console. The Restore creates a temporary copy of your data in the S3 Standard storage class while leaving the archived data intact in S3 Glacier Deep Archive. You can specify the amount of time in days for which the temporary copy is stored in S3. You can then access your temporary copy from S3 through an Amazon S3 GET request on the archived object.

When restoring an archived object, you can specify one of the following options in the Tier element of the request body: Standard is the default tier and lets you access any of your archived objects within 12 hours, and Bulk lets you retrieve large amounts, even petabytes of data inexpensively and typically completes within 48 hours.

Q: How am I charged for using S3 Glacier Deep Archive?

S3 Glacier Deep Archive storage is priced based on the amount of data you store in GBs, the number of PUT/lifecycle transition requests, retrievals in GBs, and number of restore requests. This pricing model is similar to S3 Glacier Flexible Retrieval.

Q: How will S3 Glacier Deep Archive usage show up on my AWS bill and in the AWS Cost Management tool?

S3 Glacier Deep Archive usage and cost will show up as an independent service line item on your monthly AWS bill, separate from your Amazon S3 usage and costs. However, if you are using the AWS Cost Management tool, S3 Glacier Deep Archive usage and cost will be included under the Amazon S3 usage and cost in your detailed monthly spend reports, and not broken out as a separate service line item.

Q: Are there minimum storage duration and minimum object storage charges for S3 Glacier Deep Archive?

S3 Glacier Deep Archive is designed for long-lived but rarely accessed data that is retained for 7—10 years or more. Objects that are archived to S3 Glacier Deep Archive have a minimum of 180 days of storage, and objects deleted before 180 days incur a pro-rated charge equal to the storage charge for the remaining days.

S3 Glacier Deep Archive has a minimum billable object storage size of 40KB. Objects smaller than 40KB in size may be stored but will be charged for 40KB of storage.

Q: How does S3 Glacier Deep Archive integrate with other AWS Services?

S3 Glacier Deep Archive is integrated with Amazon S3 features, including S3 Object Tagging, S3 Lifecycle policies, S3 Object Lock, and S3 Replication. With S3 storage management features, you can use a single Amazon S3 bucket to store a mixture of S3 Glacier Deep Archive, S3 Standard, S3 Standard-IA, S3 One Zone-IA, and S3 Glacier Flexible Retrieval data. This allows storage administrators to make decisions based on the nature of the data and data access patterns. Customers can use Amazon S3 Lifecycle policies to automatically migrate data to lower-cost storage classes as the data ages, or S3 Cross-Region Replication or Same-Region Replication policies to replicate data to the same or a different region.

AWS Storage Gateway service integrates Tape Gateway with S3 Glacier Deep Archive storage class, allowing you to store virtual tapes in the lowest-cost Amazon S3 storage class, reducing the monthly cost to store your long-term data in the cloud by 75%. With this feature, Tape Gateway supports archiving your new virtual tapes directly to S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive, helping you meet your backup, archive, and recovery requirements. Tape Gateway helps you move tape-based backups to AWS without making any changes to your existing backup workflows. Tape Gateway supports most of the leading backup applications such as Veritas, Veeam, Commvault, Dell EMC NetWorker, IBM Spectrum Protect (on Windows OS), and Microsoft Data Protection Manager.

S3 on Outposts

Q: What is Amazon S3 on Outposts?

Amazon S3 on Outposts delivers object storage in your on-premises environment, using the S3 APIs and capabilities that you use in AWS today. AWS Outposts is a fully managed service that extends AWS infrastructure, AWS services, APIs, and tools to virtually any datacenter, co-location space, or on-premises facility. Using S3 on Outposts, you can securely process and store customer data generated on-premises before moving it to an AWS Region, access data locally for applications that run on-premises, or store data on your Outpost for companies in locations with data residency requirements, and or those in regulated industries. To learn more about S3 on Outposts, visit the overview page.

Storage Management

S3 Object Tags

Q: What are S3 Object Tags?

S3 Object Tags are key-value pairs applied to S3 objects which can be created, updated or deleted at any time during the lifetime of the object. With these, you have the ability to create Identity and Access Management (IAM) policies, set up S3 Lifecycle policies, and customize storage metrics. These object-level tags can then manage transitions between storage classes and expire objects in the background. You can add tags to new objects when you upload them or you can add them to existing objects. Up to ten tags can be added to each S3 object and you can use either the AWS Management Console, the REST API, the AWS CLI, or the AWS SDKs to add object tags.

Q: Why should I use object tags?

Object tags are a tool you can use to enable simple management of your S3 storage. With the ability to create, update, and delete tags at any time during the lifetime of your object, your storage can adapt to the needs of your business. These tags allow you to control access to objects tagged with specific key-value pairs, allowing you to further secure confidential data for only a select group or user. Object tags can also be used to label objects that belong to a specific project or business unit, which could be used in conjunction with S3 Lifecycle policies

to manage transitions to other storage classes (S3 Standard-IA, S3 One Zone-IA, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, and S3 Glacier Deep Archive) or with S3 Replication to selectively replicate data between AWS Regions.

Q: How can I update the object tags on my objects?

Object tags can be changed at any time during the lifetime of your S3 object, you can use either the AWS Management Console, the REST API, the AWS CLI, or the AWS SDKs to change your object tags. Note that all changes to tags outside of the AWS Management Console are made to the full tag set. If you have five tags attached to a particular object and want to add a sixth, you need to include the original five tags in that request.

Q: How much do object tags cost?

Object tags are priced based on the quantity of tags and a request cost for adding tags. The requests associated with adding and updating Object Tags are priced the same as existing request prices.

Q: How do I get started with Storage Class Analysis?

You can use the AWS Management Console or the S3 PUT Bucket Analytics API to configure a Storage Class Analysis policy to identify infrequently accessed storage that can be transitioned to the S3 Standard-IA or S3 One Zone-IA storage class or archived to the S3 Glacier storage classes. You can navigate to the “Management” tab in the S3 Console to manage Storage Class Analysis, S3 Inventory, and S3 CloudWatch metrics.

S3 Inventory

Q: What is S3 Inventory?

The S3 Inventory report provides a scheduled alternative to Amazon S3’s synchronous List API. You can configure S3 Inventory to provide a CSV, ORC, or Parquet file output of your objects and their corresponding metadata on a daily or weekly basis for an S3 bucket or prefix. You can simplify and speed up business workflows and big data jobs with S3 Inventory. You can also use S3 inventory to verify encryption and replication status of your objects to meet business, compliance, and regulatory needs.

Q: How do I get started with S3 Inventory?

You can use the AWS Management Console or the PUT Bucket Inventory Configuration API to configure a daily or weekly inventory report for all the objects within your S3 bucket or a subset of the objects under a shared prefix. As part of the configuration, you can specify a destination S3 bucket for your S3 Inventory report, the output file format (CSV, ORC, or Parquet), and specific object metadata necessary for your business application, such as object name, size, last modified date, storage class, version ID, delete marker, non-current version flag, multipart upload flag, replication status, or encryption status. You can use S3 Inventory as a direct input into your application workflows or Big Data jobs. You can also query S3 Inventory using Standard SQL language with Amazon Athena, Amazon Redshift Spectrum, and other tools such as Presto, Hive, and Spark.

Q: How am I charged for using S3 Inventory?

Please see the Amazon S3 pricing page for S3 Inventory pricing. Once you configure encryption using SSE-KMS, you will incur KMS charges for encryption, refer to the KMS pricing page for detail.

S3 Batch Operations

Q: What is S3 Batch Operations?

S3 Batch Operations is a feature that you can use to automate the execution of a single operation (like copying an object, or executing an AWS Lambda function) across many objects. With S3 Batch Operations, you can, with a few clicks in the S3 console or a single API request, make a change to billions of objects without having to write custom application code or run compute clusters for storage management applications. Not only does S3 Batch Operations administer your storage operation across many objects, S3 Batch Operations manages retries, displays progress, delivers notifications, provides a completion report, and sends events to AWS CloudTrail for all operations performed on your target objects. S3 Batch Operations can be used from the S3 console, or through the AWS CLI and SDK.

Q: How do I get started with S3 Batch Operations?

You can get started with S3 Batch Operations by going into the Amazon S3 console or using the AWS CLI or SDK to create your first S3 Batch Operations job. A S3 Batch Operations job consists of the list of objects to act upon and the type of operation to be performed (see the full list of available operations). Start by selecting an S3 Inventory report or providing your own custom list of objects for S3 Batch Operations to act upon. An S3 Inventory report is a file listing all objects stored in an S3 bucket or prefix. Next, you choose from a set of S3 operations supported by S3 Batch Operations, such as replacing tag sets, changing ACLs, copying storage from one bucket to another, or initiating a restore from S3 Glacier Flexible Retrieval to S3 Standard storage class. You can then customize your S3 Batch Operations jobs with specific parameters such as tag values, ACL grantees, and restoration duration. To further customize your storage actions, you can write your own Lambda function and invoke that code through S3 Batch Operations.

Once you create your S3 Batch Operations job, S3 Batch Operations will process your list of objects and send the job to the “awaiting confirmation” state if required. After you confirm the job details, S3 Batch Operations will begin executing the operation you specified. You can view your job’s progress programmatically or through the S3 console, receive notifications on completion, and review a completion report that itemizes the changes made to your storage.

S3 Object Lock

Q: What is Amazon S3 Object Lock?

Amazon S3 Object Lock is an Amazon S3 feature that prevents an object version from being deleted or overwritten for a fixed amount of time or indefinitely, so that you can enforce retention policies as an added layer of data protection or for regulatory compliance. You can migrate workloads from existing write-once-read-many (WORM) systems into Amazon S3, and configure S3 Object Lock at the object- and bucket-level to prevent object version deletions prior to pre-defined Retain Until Dates or indefinitely (Legal Hold Dates). S3 Object Lock protection is maintained regardless of which storage class the object version resides in and throughout S3 Lifecycle transitions between storage classes.

You should use S3 Object Lock if you have regulatory requirements that specify that data must be WORM protected, or if you want to add an additional layer of protection to data in Amazon S3. S3 Object Lock can help you to meet regulatory requirements that specify that data should

be stored in an immutable format, and also can protect against accidental or malicious deletion for data in Amazon S3.

Q: How does Amazon S3 Object Lock work?

Amazon S3 Object Lock prevents deletion of an object version for the duration of a specified retention period or indefinitely until a legal hold is removed. With S3 Object Lock, you're able to ensure that an object version remains immutable for as long as WORM protection is applied. You can apply WORM protection by either assigning a Retain Until Date or a Legal Hold to an object version using the AWS SDK, CLI, REST API, or the S3 Management Console. You can apply retention settings within a PUT request, or apply them to an existing object after it has been created.

The Retain Until Date defines the length of time for which an object version will remain immutable. Once a Retain Until Date has been assigned to an object, that object version cannot be modified or deleted until the Retain Until Date has passed. If a user attempts to delete an object before the Retain Until Date, the operation will be denied.

Alternatively, you can make an object immutable by applying a Legal Hold. A Legal Hold prevents an object version from being modified or deleted indefinitely until it is explicitly removed. In order to place and remove Legal Holds, your AWS account must have write permission for the PutObjectLegalHold action. Legal Hold can be applied to any object in an S3 Object Lock enabled bucket, whether or not that object is currently WORM-protected by a retention period.

S3 Object Lock can be configured in one of two Modes. When deployed in Governance Mode, AWS accounts with specific IAM permissions are able to remove WORM protection from an object version. If you require stronger immutability in order to comply with regulations, you can use Compliance Mode. In Compliance Mode, WORM protection cannot be removed by any user, including the root account.
Q: What AWS electronic storage services have been assessed based on financial services regulations?

For customers in the financial services industry, S3 Object Lock provides added support for broker-dealers who must retain records in a non-erasable and non-rewritable format to satisfy regulatory requirements of SEC Rule 17a-4(f), FINRA Rule 4511, or CFTC Regulation 1.31. You can easily designate the records retention time frame to retain regulatory archives in the original form for the required duration, and also place legal holds to retain data indefinitely until the hold is removed.
Q: What AWS documentation supports the SEC 17a-4(f)(2)(i) and CFTC 1.31(c) requirement for notifying my regulator?

Provide notification to your regulator or "Designated Examining Authority (DEA)" of your choice to use Amazon S3 for electronic storage along with a copy of the Cohasset Assessment. For the purposes of these requirements, AWS is not a designated third party (D3P). Be sure to select a D3P and include this information in your notification to your DEA.

S3 CloudWatch Metrics

Q: How do I get started with S3 CloudWatch Metrics?

You can use the AWS Management Console to enable the generation of one-minute CloudWatch request metrics for your S3 bucket or configure filters for the metrics using a prefix or object tag, or access point. Alternatively, you can call the S3 PUT Bucket Metrics API to enable and configure publication of S3 storage metrics. CloudWatch Request Metrics will be

available in CloudWatch within 15 minutes after they are enabled. CloudWatch Storage Metrics are enabled by default for all buckets, and reported once per day.

Q: What alarms can I set on my storage metrics?

You can use CloudWatch to set thresholds on any of the storage metrics counts, timers, or rates and trigger an action when the threshold is breached. For example, you can set a threshold on the percentage of 4xx Error Responses and when at least three data points are above the threshold trigger a CloudWatch alarm to alert a DevOps engineer.

Q: How am I charged for using S3 CloudWatch Metrics?

CloudWatch storage metrics are provided free. Cloudwatch request metrics are priced as custom metrics for Amazon CloudWatch.

S3 Lifecycle Management

Q: What is S3 Lifecycle management?

S3 Lifecycle management provides the ability to define the lifecycle of your object with a predefined policy and reduce your cost of storage. You can set a lifecycle transition policy to automatically migrate objects stored in the S3 Standard storage class to the S3 Standard-IA, S3 One Zone-IA, and/or S3 Glacier storage classes based on the age of the data. You can also set lifecycle expiration policies to automatically remove objects based on the age of the object. You can set a policy for multipart upload expiration, which expires incomplete multipart uploads based on the age of the upload.

Q: How do I set up an S3 Lifecycle management policy?

You can set up and manage Lifecycle policies in the AWS Management Console, S3 REST API, AWS SDKs, or AWS Command Line Interface (CLI). You can specify the policy at the prefix or at the bucket level.

Q: How can I use Amazon S3 Lifecycle management to help lower my Amazon S3 storage costs?

With Amazon S3 Lifecycle policies, you can configure your objects to be migrated from the S3 Standard storage class to S3 Standard-IA or S3 One Zone-IA and/or archived to S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, or S3 Glacier Deep Archive storage classes.

You can also specify an S3 Lifecycle policy to delete objects after a specific period of time. You can use this policy-driven automation to quickly and easily reduce storage costs as well as save time. In each rule you can specify a prefix, a time period, a transition to S3 Standard-IA, S3 One Zone-IA, S3 Glacier Instant Retrieval, S3 Glacier Flexible Retrieval, S3 Glacier Deep Archive, and/or an expiration. For example, you could create a rule that archives into S3 Glacier Flexible Retrieval all objects with the common prefix “logs/” 30 days from creation and expires these objects after 365 days from creation.

You can also create a separate rule that only expires all objects with the prefix “backups/” 90 days from creation. S3 Lifecycle policies apply to both existing and new S3 objects, helping you optimize storage and maximize cost savings for all current data and any new data placed in S3 without time-consuming manual data review and migration.

Within a lifecycle rule, the prefix field identifies the objects subject to the rule. To apply the rule

to an individual object, specify the key name. To apply the rule to a set of objects, specify their common prefix (e.g. “logs/”). You can specify a transition action to have your objects archived and an expiration action to have your objects removed. For time period, provide the creation date (e.g. January 31, 2015) or the number of days from creation date (e.g. 30 days) after which you want your objects to be archived or removed. You may create multiple rules for different prefixes.

Q: How much does it cost to use S3 Lifecycle management?

There is no additional cost to set up and apply Lifecycle policies. A transition request is charged per object when an object becomes eligible for transition according to the Lifecycle rule.

Q: Why would I use an S3 Lifecycle policy to expire incomplete multipart uploads?

The S3 Lifecycle policy that expires incomplete multipart uploads allows you to save on costs by limiting the time non-completed multipart uploads are stored. For example, if your application uploads several multipart object parts, but never commits them, you will still be charged for that storage. This policy can lower your S3 storage bill by automatically removing incomplete multipart uploads and the associated storage after a predefined number of days.

Q: Can I set up Amazon S3 Event Notifications to send notifications when S3 Lifecycle transitions or expires objects?

Yes, you can set up Amazon S3 Event Notifications to notify you when S3 Lifecycle transitions or expires objects. For example, you can send S3 Event Notifications to an Amazon SNS topic, Amazon SQS queue, or AWS Lambda function when S3 Lifecycle moves objects to a different S3 storage class or expires objects.

Storage Analytics & Insights

Q: What features are available to analyze my storage usage on Amazon S3?

S3 Storage Lens delivers organization-wide visibility into object storage usage, activity trends, and makes actionable recommendations to optimize costs and apply data protection best practices. S3 Storage Class Analysis enables you to monitor access patterns across objects to help you decide when to transition data to the right storage class to optimize costs. You can then use this information to configure an S3 Lifecycle policy that makes the data transfer. Amazon S3 Inventory provides a report of your objects and their corresponding metadata on a daily or weekly basis for an S3 bucket or prefix. This report can be used to help meet business, compliance, and regulatory needs by verifying the encryption, and replication status of your objects.

Q: What is Amazon S3 Storage Lens?

Amazon S3 Storage Lens provides organization-wide visibility into object storage usage and activity trends, as well as actionable recommendations to optimize costs and apply data protection best practices. Storage Lens offers an interactive dashboard containing a single view of your object storage usage and activity across tens or hundreds of accounts in your organization, with drill-downs to generate insights at multiple aggregation levels. This includes metrics like bytes, object counts, and requests, as well as metrics detailing S3 feature utilization, such as encrypted object counts and S3 Lifecycle rule counts. S3 Storage Lens also delivers contextual recommendations to find ways for you to reduce storage costs and apply best practices on data protection across tens or hundreds of accounts and buckets. S3 Storage

Lens free metrics are enabled by default for all Amazon S3 users. If you want to get more out of S3 Storage Lens, you can activate advanced metrics and recommendations.

Q: How does S3 Storage Lens work?

S3 Storage Lens aggregates your storage usage and activity metrics on a daily basis to be visualized in the S3 Storage Lens interactive dashboard, or available as a metrics export in CSV or Parquet file format. A default dashboard is created for you automatically at the account level, and you have the option to create additional custom dashboards. S3 Storage Lens dashboards can be scoped to your AWS organization or specific accounts, Regions, buckets, or even prefix level (available with S3 Storage Lens advanced metrics). In configuring your dashboard you can use the default metrics selection, or upgrade to receive 35 additional metrics and recommendations for an additional cost. Also, S3 Storage Lens provides recommendations contextually with storage metrics in the dashboard, so you can take action to optimize your storage based on the metrics.

Q: What are the key questions that can be answered using S3 Storage Lens metrics?

The S3 Storage Lens dashboard is organized around four main types of questions that can be answered about your storage. With the Summary filter, top-level questions related to overall storage usage and activity trends can be explored. For example, “How rapidly is my overall byte count and request count increasing over time?” With the Cost Optimization filter, you can explore questions related to storage cost reduction, for example, “Is it possible for me to save money by retaining fewer non-current versions?” With the Data Protection and Access Management filters you can answer questions about securing your data, for example, “Is my storage protected from accidental or intentional deletion?” Finally, with the Performance and Events filters you can explore ways to improve performance of workflows. Each of these questions represent a first layer of inquiry that would likely lead to drill-down analysis.

Q: What metrics are available in S3 Storage Lens?

S3 Storage Lens contains more than 60 metrics, grouped into free metrics and advanced metrics (available for an additional cost). Within free metrics, you receive metrics to analyze usage (based on a daily snapshot of your objects), which are organized into the categories of cost optimization, data protection, access management, performance, and events. Within advanced metrics, you receive metrics related to activity (such as request counts), deeper cost optimization (such as S3 Lifecycle rule counts), additional data protection (such as S3 Replication rule counts), and detailed status codes (such as 403 authorization errors). In addition, derived metrics are also provided by combining any base metrics. For example, “Retrieval Rate" is a metric calculated by dividing the "Bytes Downloaded Count" by the "Total Storage.” To view the complete list of metrics, please visit the S3 Storage Lens documentation.

Q: What are my dashboard configuration options?

A default dashboard is configured automatically provided for your entire account, and you have the option to create additional custom dashboards that can be scoped to your AWS organization, specific regions, or buckets within an account. You can set up multiple custom dashboards, which can be useful if you require some logical separation in your storage analysis, such as segmenting on buckets to represent various internal teams. By default, your dashboard will receive the S3 Storage Lens free metrics, but you have the option to upgrade to receive S3 Storage Lens advanced metrics and recommendations (for an additional cost). S3 Storage Lens advanced metrics have 6 distinct options: Activity metrics, Advanced Cost Optimization metrics, Advanced Data Protection metrics, Detailed Status Code metrics, Prefix aggregation, and CloudWatch publishing. Additionally, for each dashboard you can enable metrics export, with additional options to specify destination bucket and encryption type.

Q: How much historical data is available in S3 Storage Lens?

For metrics displayed in the interactive dashboard, Storage Lens free metrics retains 14 days of historical data, and Storage Lens advanced metrics (for an additional cost) retains 15 months of historical data. For the optional metrics export, you can configure any retention period you wish, and standard S3 storage charges will apply.

Q: How will I be charged for S3 Storage Lens?

S3 Storage Lens is available in two tiers of metrics. The free metrics are enabled by default and available at no additional charge to all S3 customers. The S3 Storage Lens advanced metrics and recommendations pricing details are available on the S3 pricing page. With S3 Storage Lens free metrics you receive 28 metrics at the bucket level, and can access 14 days of historical data in the dashboard. With S3 Storage Lens advanced metrics and recommendations you receive 35 additional metrics, prefix-level aggregation, CloudWatch metrics support and can access 15 months of historical data in the dashboard.

Q: What is the difference between S3 Storage Lens and S3 Inventory?

S3 Inventory provides a list of your objects and their corresponding metadata for an S3 bucket or a shared prefix, which can be used to perform object-level analysis of your storage. S3 Storage Lens provides metrics aggregated by organization, account, region, storage class, bucket, and prefix levels, which improve organization-wide visibility of your storage.

Q: What is the difference between S3 Storage Lens and S3 Storage Class Analysis (SCA)?

S3 Storage Class Analysis provides recommendations for an optimal storage class by creating object age groups based on object-level access patterns within an individual bucket/prefix/tag for the previous 30 – 90 days. S3 Storage Lens provides daily organization level recommendations on ways to improve cost efficiency and apply data protection best practices, with additional granular recommendations by account, region, storage class, bucket or prefix (available with S3 Storage Lens advanced metrics).

Storage Class Analysis

Q: What is Storage Class Analysis?

With Storage Class Analysis, you can analyze storage access patterns to determine the optimal storage class for your storage. This S3 feature automatically identifies infrequent access patterns to help you transition storage to S3 Standard-IA. You can configure a Storage Class Analysis policy to monitor an entire bucket, prefix, or object tag. Once an infrequent access pattern is observed, you can easily create a new S3 Lifecycle age policy based on the results. Storage Class Analysis also provides daily visualizations of your storage usage on the AWS Management Console and you can also enable an export report to an S3 bucket to analyze using business intelligence tools of your choice such as Amazon QuickSight.

Q: How often is the Storage Class Analysis updated?

Storage Class Analysis is updated on a daily basis in the S3 Management Console, but initial recommendations for storage class transitions are provided after 30 days.

Query in Place

Q: What is "Query in Place" functionality?

Amazon S3 allows customers to run sophisticated queries against data stored without the need to move data into a separate analytics platform. The ability to query this data in place on Amazon S3 can significantly increase performance and reduce cost for analytics solutions leveraging S3 as a data lake. S3 offers multiple query in place options, including S3 Select, Amazon Athena, and Amazon Redshift Spectrum, allowing you to choose one that best fits your use case. You can even use Amazon S3 Select with AWS Lambda to build serverless apps that can take advantage of the in-place processing capabilities provided by S3 Select.

Q: What is S3 Select?

S3 Select is an Amazon S3 feature that makes it easy to retrieve specific data from the contents of an object using simple SQL expressions without having to retrieve the entire object. S3 Select simplifies and improves the performance of scanning and filtering the contents of objects into a smaller, targeted dataset.

You can use S3 Select to retrieve a subset of data using SQL clauses, like SELECT and WHERE, from objects stored in CSV, JSON, or Apache Parquet format. It also works with objects that are compressed with GZIP or BZIP2 (for CSV and JSON objects only), and server-side encrypted objects.

You can use S3 Select to perform operational investigations on log files in Amazon S3 without the need to operate or manage a compute cluster. You can also use S3 Select with open source big data analytics frameworks like Trino to “push down” the computational work of projection operations (for example, SELECT) and predicate operations (for example WHERE) to Amazon S3 for CSV and JSON data formats.

Q: What is Amazon Athena?

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL queries. Athena is serverless, so there is no infrastructure to set up or manage, and you can start analyzing data immediately. You don’t even need to load your data into Athena; it works directly with data stored in any S3 storage class. To get started, just log into the Athena Management Console, define your schema, and start querying. Amazon Athena uses Presto with full standard SQL support and works with a variety of standard data formats, including CSV, JSON, ORC, Apache Parquet and Avro. While Athena is ideal for quick, ad-hoc querying and integrates with Amazon QuickSight for easy visualization, it can also handle complex analysis, including large joins, window functions, and arrays.

Q: What is Amazon Redshift Spectrum?

Amazon Redshift Spectrum is a feature of Amazon Redshift that lets you run queries against exabytes of unstructured data in Amazon S3 with no loading or ETL required. When you issue a query, it goes to the Amazon Redshift SQL endpoint, which generates and optimizes a query plan. Amazon Redshift determines what data is local and what is in Amazon S3, generates a plan to minimize the amount of Amazon S3 data that needs to be read, and requests Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3.

Redshift Spectrum scales out to thousands of instances if needed, so queries run quickly regardless of data size. And, you can use the exact same SQL for Amazon S3 data as you do for your Amazon Redshift queries today and connect to the same Amazon Redshift endpoint using the same business intelligence tools. Redshift Spectrum lets you separate storage and compute, allowing you to scale each independently. You can set up as many Amazon Redshift clusters as you need to query your Amazon S3 data lake, providing high availability and limitless

concurrency. Redshift Spectrum gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need it.

Replication

Q: What is Amazon S3 Replication?

Amazon S3 Replication enables automatic, asynchronous copying of objects across Amazon S3 buckets. Buckets that are configured for object replication can be owned by the same AWS account or by different accounts. You can replicate new objects written to the bucket to one or more destination buckets between different AWS Regions (S3 Cross-Region Replication), or within the same AWS Region (S3 Same-Region Replication). You can also replicate existing bucket contents (S3 Batch Replication), including existing objects, objects that previously failed to replicate, and objects replicated from another source. Learn more by visiting the S3 Replication user guide.

Q: What is Amazon S3 Cross-Region Replication (CRR)?

CRR is an Amazon S3 feature that automatically replicates data between buckets across different AWS Regions. With CRR, you can set up replication at a bucket level, a shared prefix level, or an object level using S3 object tags. You can use CRR to provide lower-latency data access in different geographic regions. CRR can also help if you have a compliance requirement to store copies of data hundreds of miles apart. You can use CRR to change account ownership for the replicated objects to protect data from accidental deletion. To learn more visit the S3 CRR user guide.

Q: What is Amazon S3 Same-Region Replication (SRR)?

SRR is an Amazon S3 feature that automatically replicates data between buckets within the same AWS Region. With SRR, you can set up replication at a bucket level, a shared prefix level, or an object level using S3 object tags. You can use SRR to create one or more copies of your data in the same AWS Region. SRR helps you address data sovereignty and compliance requirements by keeping a copy of your data in a separate AWS account in the same region as the original. You can use SRR to change account ownership for the replicated objects to protect data from accidental deletion. You can also use SRR to easily aggregate logs from different S3 buckets for in-region processing, or to configure live replication between test and development environments. To learn more visit the S3 SRR user guide.

Q: What is Amazon S3 Batch Replication?

Amazon S3 Batch Replication replicates existing objects between buckets. You can use S3 Batch Replication to backfill a newly created bucket with existing objects, retry objects that were previously unable to replicate, migrate data across accounts, or add new buckets to your data lake. You can get started with S3 Batch Replication with just a few clicks in the S3 console or a single API request. To learn more visit the S3 Batch Replication user guide.
Q: How do I enable Amazon S3 Replication (Cross-Region Replication and Same-Region Replication)?

Amazon S3 Replication (CRR and SRR) is configured at the S3 bucket level, a shared prefix level, or an object level using S3 object tags. You add a replication configuration on your source bucket by specifying a destination bucket in the same or different AWS region for replication.

You can use the S3 console, API, the AWS CLI, AWS SDKs, or AWS CloudFormation to enable replication. Versioning must be enabled for both the source and destination buckets to enable

replication. To learn more, please visit overview of setting up S3 Replication in the Amazon S3 Developer Guide.

Q: How do I use S3 Batch Replication?

You would first need to enable S3 Replication at the bucket level. See the previous question for how you can do so. You may then initiate an S3 Batch Replication job in the S3 console after creating a new replication configuration, changing a replication destination in a replication rule from the replication configuration page, or from the S3 Batch Operations Create Job page. Alternatively, you can initiate an S3 Batch Replication jobs via the AWS CLI or SDKs. To learn more, please visit S3 Replication in the Amazon S3 Developer Guide.
Q: Can I use S3 Replication with S3 Lifecycle rules?

With S3 Replication, you can establish replication rules to make copies of your objects into another storage class, in the same or a different region. Lifecycle actions are not replicated, and if you want the same lifecycle configuration applied to both source and destination buckets, enable the same lifecycle configuration on both.

For example, you can configure a lifecycle rule to migrate data from the S3 Standard storage class to the S3 Standard-IA or S3 One Zone-IA storage class or archive data to a S3 Glacier storage class on the destination bucket.

If you have S3 Lifecycle configured for your destination bucket, we recommend disabling Lifecycle rules while the Batch Replication job is active to maintain parity between noncurrent and current versions of objects in the source and destination buckets.

You can find more information about lifecycle configuration and replication on the S3 Replication developer guide.

Q: Can I use S3 Replication to replicate to more than one destination bucket?

Yes. S3 Replication allows customers to replicate their data to multiple destination buckets in the same, or different AWS Regions. When setting up, you simply specify the new destination bucket in your existing replication configuration or create a new replication configuration with multiple destination buckets. For each new destination you specify, you have the flexibility to choose storage class of destination bucket, encryption type, replication metrics and notifications, Replication Time Control (RTC), and other properties.

Q: Can I use S3 Replication to set up two-way replication between S3 buckets?

Yes. To set up two-way replication, you create a replicate rule from bucket A to bucket B and set up another replication rule from bucket B to bucket A. Make sure to enable replica modification sync on both buckets A and B to replicate replica metadata changes like object access control lists (ACLs), object tags, or object locks on the replicated objects.
Q: Can I use replication across AWS accounts to protect against malicious or accidental deletion?

Yes, for CRR and SRR, you can set up replication across AWS accounts to store your replicated data in a different account in the target region. You can use Ownership Overwrite in your replication configuration to maintain a distinct ownership stack between source and destination, and grant destination account ownership to the replicated storage.

Q: Will my object tags be replicated if I use Cross-Region Replication?

Object tags can be replicated across AWS Regions using Cross-Region Replication. For customers with Cross-Region Replication already enabled, new permissions are required in order for tags to replicate. For more information about setting up Cross-Region Replication, please visit How to Set Up Cross-Region Replication in the Amazon S3 Developer Guide.

Q: Can I replicate delete markers from one bucket to another?

Yes, you can replicate delete markers from source to destination if you have delete marker replication enabled in your replication configuration. When you replicate delete markers, Amazon S3 will behave as if the object was deleted in both buckets. You can enable delete marker replication for a new or existing replication rule. You can apply delete marker replication to the entire bucket or to Amazon S3 objects that have a specific prefix, with prefix based replication rules. Amazon S3 Replication does not support delete marker replication for object tag based replication rules. To learn more about enabling delete marker replication see Replicating delete markers from one bucket to another.

Q: Can I replicate data from other AWS Regions to China? Can a customer replicate from one China Region bucket outside of China Regions?

No, Amazon S3 Replication is not available between AWS China Regions and AWS Regions outside of China. You are only able to replicate within the China regions.

Q: Can I replicate existing objects?

Yes. You can use S3 Batch Replication to replicate existing objects between buckets. To learn more, visit the S3 User Guide.

Q: Can I re-try replication if object fail to replicate initially?

Yes. You can use S3 Batch Replication to re-replicate objects that fail to replicate initially. To learn more, visit the S3 User Guide.

Q: What encryption types does S3 Replication support?

S3 Replication supports all encryption types that S3 offers. S3 offers both server-side encryption and client-side encryption – the former requests S3 to encrypt the objects for you, and the latter is for you to encrypt data on the client-side before uploading it to S3. For server-side encryption, S3 offers server-side encryption with Amazon S3-managed keys (SSE-S3), server-side encryption with KMS keys stored in AWS Key Management Service (SSE-KMS), and server-side encryption with customer-provided keys (SSE-C). For further details on these encryption types and how they work, visit the S3 documentation on using encryption.

Q: What is the pricing for cross account data replication?

With S3 Replication, you can configure cross account replication where the source and destination buckets are owned by different AWS accounts. Excluding S3 storage and applicable retrieval charges, customers pay for replication PUT requests and inter-region Data Transfer OUT from S3 to your destination region when using S3 Replication. If you have S3 Replication Time Control (S3 RTC) enabled on your replication rules, you will see a different Data Transfer OUT and replication PUT request charges specific to S3 RTC. For cross account replication, the source account pays for all data transfer (S3 RTC and S3 CRR) and the destination account pays for the replication PUT requests. Data transfer charges only apply for S3 Cross Region

Replication (S3 CRR) and S3 Replication Time Control (S3 RTC), there are no data transfer charges for S3 Same Region Replication (S3 SRR).

If you are using S3 Batch Replication to replicate objects across accounts, you will incur the S3 Batch Operations charges, in addition to the replication PUT requests and Data Transfer OUT charges (note that S3 RTC is not applicable to Batch Replication). The Batch Operations charges include the Job and Object charges, which are respectively based on the number of jobs and number of objects processed. Additionally, if you opt in for the Amazon S3 generated manifest, you will incur a manifest charge based on the number of objects in the source bucket.

S3 Replication Time Control

Q: What is Amazon S3 Replication Time Control?

Amazon S3 Replication Time Control provides predictable replication performance and helps you meet compliance or business requirements. S3 Replication Time Control is designed to replicate most objects in seconds, and 99.99% of objects within 15 minutes. S3 Replication Time Control is backed by a Service Level Agreement (SLA) commitment that 99.9% of objects will be replicated in 15 minutes for each replication region pair during any billing month. Replication Time works with all S3 Replication features. To learn more, please visit the replication developer guide.

Q: How do I enable Amazon S3 Replication Time Control?

You can enable S3 Replication Time Control as an option for each replication rule. You can either create a new S3 Replication rule with S3 Replication Time Control, or enable the feature on an existing replication rule.

You can use the S3 console, API, AWS CLI, AWS SDKs, or AWS CloudFormation to configure replication. To learn more, please visit overview of setting up Replication in the Amazon S3 Developer Guide.

Q: Can I use S3 Replication Time Control to replicate data within and between China Regions?

Yes, you can enable Amazon S3 Replication Time Control to replicate data within and between the AWS China (Ningxia) and China (Beijing) Regions.

Q: What are Amazon S3 Replication metrics and events?

Amazon S3 Replication metrics and events provide visibility into Amazon S3 Replication. With S3 Replication metrics, you can monitor the total number of operations and size of objects that are pending replication, and the replication latency between source and destination buckets for each S3 Replication rule. Replication metrics are available through the Amazon S3 Management Console and through Amazon CloudWatch. S3 Replication events will notify of you of replication failures so you can quickly diagnose and correct issues. If you have S3 Replication Time Control (S3 RTC) enabled you will also receive notifications when an object takes more than 15 minutes to replicate, and when that object replicates successfully to their destination. Like other Amazon S3 events, S3 Replication events are available through Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS), or AWS Lambda.

Q: How do I enable Amazon S3 Replication metrics and events?

You can enable Amazon S3 Replication metrics and events for new or existing replication rules, and they are enabled by default for S3 Replication Time Control enabled rules. You can access S3 Replication metrics through the Amazon S3 Management Console and Amazon CloudWatch. Like other Amazon S3 events, S3 Replication events are available through Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS), or AWS Lambda. To learn more, please visit Monitoring progress with replication metrics and Amazon S3 Event Notifications in the Amazon S3 Developer Guide.

Q: What is the Amazon S3 Replication Time Control Service Level Agreement (SLA)?

Amazon S3 Replication Time Control is designed to replicate 99.99% of your objects within 15 minutes, and is backed by a service level agreement. If fewer than 99.9% of your objects are replicated in 15 minutes for each replication region pair during a monthly billing cycle, the S3 RTC SLA provides a service credit on any object that takes longer than 15 minutes to replicate. The service credit covers a percentage of all replication-related charges associated with the objects that did not meet the SLA, including the RTC charge, replication bandwidth and request charges, and the cost associated with storing your replica in the destination region in the monthly billing cycle affected. To learn more, read the S3 Replication Time Control SLA.

Q: What is the pricing for S3 Replication and S3 Replication Time Control?

For S3 Replication (Cross-Region Replication and Same Region Replication), you pay the S3 charges for storage in the selected destination S3 storage classes, the storage charges for the primary copy, replication PUT requests, and applicable infrequent access storage retrieval charges. For CRR, you also pay for inter-region Data Transfer OUT From S3 to your destination region. S3 Replication Metrics are billed at the same rate as Amazon CloudWatch custom metrics. Additionally, when you use S3 Replication Time Control, you also pay a Replication Time Control Data Transfer charge.

If the source object is uploaded using the multipart upload feature, then it is replicated using the same number of parts and part size. For example, a 100 GB object uploaded using the multipart upload feature (800 parts of 128 MB each) will incur request cost associated with 802 requests (800 Upload Part requests + 1 Initiate Multipart Upload request + 1 Complete Multipart Upload request) when replicated. You will incur a request charge of $0.00401 (802 requests x $0.005 per 1,000 requests) and (if the replication was between different AWS Regions) a charge of $2.00 ($0.020 per GB transferred x 100 GB) for inter-region data transfer. After replication, the 100 GB will incur storage charges based on the destination region.

S3 Multi-Region Access Points

Q: What are S3 Multi-Region Access Points?

Amazon S3 Multi-Region Access Points accelerate performance by up to 60% when accessing data sets that are replicated across multiple AWS Regions. Based on AWS Global Accelerator, S3 Multi-Region Access Points consider factors like network congestion and the location of the requesting application to dynamically route your requests over the AWS network to the lowest latency copy of your data. This automatic routing allows you to take advantage of the global infrastructure of AWS while maintaining a simple application architecture.

Q: Why should I use S3 Multi-Region Access Points?

S3 Multi-Region Access Points accelerate and simplify storage for your multi-region applications. By dynamically routing S3 requests made to a replicated data set, S3 Multi-Region Access Points reduce request latency, so that applications run up to 60% faster. In addition,

you are able to take advantage of the global infrastructure of AWS, while maintaining a simple region-agnostic architecture for your applications.

Q: How do S3 Multi-Region Access Points work?

Multi-Region Access Points dynamically route client requests to one or more underlying S3 buckets. You can configure your Multi-Region Access Point to route across one bucket per AWS Region, in up to 20 AWS Regions. When you create a Multi-Region Access Point, S3 automatically generates a DNS-compatible name. This name is used as a global endpoint that can be used by your clients. When your clients make requests to this endpoint, S3 will dynamically route those requests to one of the underlying buckets that are specified in the configuration of your Multi-Region Access Point.

By default, S3 Multi-Region Access Points route requests to the underlying bucket that is closest to the client, based on network latency. For example, you can configure a Multi-Region Access Point with underlying buckets in US-EAST-1 and in AP-SOUTH-1. With this configuration, your clients in North America will route to US-EAST-1, and your clients in Asia will route to AP-SOUTH-1. This lowers latency for your requests made to S3, improving the performance of your application.

If your application accesses S3 over the internet, then performance will be further improved by S3 Multi-Region Access Points, because S3 will route your requests through the closest AWS Location to your client, and then over the global private AWS network to S3. Requests will avoid congested network segments on the internet, which will reduce network latency and jitter, while improving performance.

Q: How do I get started with S3 Multi-Region Access Points?

The S3 Management Console provides a simple guided workflow that enables you to quickly set up everything you need to run multi-region storage on S3, in just three simple steps. First, a unique global hostname for your Multi-Region Access Point will be automatically generated for you. You can connect your clients and applications without having to specify an AWS Region. Second, you will select one or more buckets in S3 that you would like to route across behind this new hostname. You can choose existing buckets or decide to have new buckets created for you. Third, you will specify S3 Cross-Region Replication rules.

Alternatively, you can use CloudFormation to automate your multi-region storage configuration. All of the building blocks required to set up multi-region storage on S3, including S3 Multi-Region Access Points, are supported by CloudFormation, allowing you to easily automate a repeatable setup process outside of the S3 Management Console.

Q: What is the difference between S3 Cross-Region Replication (S3 CRR) and S3 Multi-Region Access Points?

S3 CRR and S3 Multi-Region Access Points are complementary features that work together to replicate data across AWS Regions and then to automatically route requests to the replicated copy with the lowest latency. S3 S3 Multi-Region Access Points help you to manage requests across AWS Regions, while CRR allows you to move data across AWS Regions to create isolated replicas. You use S3 Multi-Region Access Points and CRR together to create a replicated multi-region dataset that is addressable by a single global endpoint.

Q: How much do S3 Multi-Region Access Points cost?

When you use an S3 Multi-Region Access Point to route requests within AWS, you pay a low per-GB data routing charge for each GB processed, as well as standard charges for S3 requests, storage, data transfer, and replication. If your application runs outside of AWS and

accesses S3 over the internet, S3 Multi-Region Access Points increase performance by automatically routing your requests through an AWS edge location, over the global private AWS network, to the closest copy of your data based on access latency. When you accelerate requests made over the internet, you pay the data routing charge and an internet acceleration charge. S3 Multi-Region Access Points internet Acceleration pricing varies based on whether the source client is in the same or in a different location as the destination AWS Region, and is in addition to standard S3 data transfer pricing.

Q: How is S3 Transfer Acceleration different than S3 Multi-Region Access Points?

You can use S3 Transfer Acceleration to speed up content transfers to and from a single centralized S3 bucket using the AWS global network. This is particularly helpful for long-distance transfers of larger objects or remove web or mobile applications. With S3 Multi-Region Access Points, you can perform similar accelerated transfers using the AWS global network, but across many S3 buckets in multiple AWS Regions for internet-based, VPC-based, and on-premise requests to and from S3. When you combine S3 Multi-Region Access Points with S3 Cross Replication, you provide the capability for S3 Multi-Region Access Points to dynamically route your requests to the lowest latency copy of your data for applications from clients in multiple locations.

Data processing

Object Lambda

Q: What is S3 Object Lambda?

S3 Object Lambda allows you to add your own code to S3 GET, LIST, and HEAD requests to modify and process data as it is returned to an application. You can use custom code to modify the data returned by S3 GET requests to filter rows, dynamically resize images, redact confidential data, and much more. You can also use S3 Object Lambda to modify the output of S3 LIST requests to create a custom view of objects in a bucket and S3 HEAD requests to modify object metadata like object name and size. S3 Object Lambda helps you to easily meet the unique data format requirements of any application without having to build and operate additional infrastructure, such as a proxy layer, or having to create and maintain multiple derivative copies of your data. S3 Object Lambda uses AWS Lambda functions to automatically process the output of a standard S3 GET, LIST, or HEAD request. AWS Lambda is a serverless compute service that runs customer-defined code without requiring management of underlying compute resources.

With just a few clicks in the AWS Management Console, you can configure a Lambda function and attach it to an S3 Object Lambda service Access Point. From that point forward, S3 will automatically call your Lambda function to process any data retrieved through the S3 Object Lambda endpoint, returning a transformed result back to the application. You can author and execute your own custom Lambda functions, tailoring S3 Object Lambda's data transformation to your specific use case.

To get started with S3 Object Lambda, you can use the S3 Management Console, SDK, or API. Learn more on the S3 Object Lambda page, or the S3 Object Lambda user guide.

Q: Why should I use S3 Object Lambda?

You should use S3 Object Lambda if you want to process data inline with an S3 GET, LIST, or HEAD request. You can use S3 Object Lambda to share a single copy of your data across many applications, avoiding the need to build and operate custom processing infrastructure or to store derivative copies of your data. For example, by using S3 Object Lambda to process S3 GET requests, you can mask sensitive data for compliance purposes, restructure raw data for

the purpose of making it compatible with machine learning applications, filter data to restrict access to specific content within an S3 object, or to address a wide range of additional use cases. You can use S3 Object Lambda to enrich your object lists by querying an external index that contains additional object metadata, filter and mask your object lists to only include objects with a specific object tag, or add a file extension to all the object names in your object lists. For example, if you have an S3 bucket with multiple discrete data sets, you can use S3 Object Lambda to filter an S3 LIST response depending on the requester.

S3 Object Lambda can be set up with just a few clicks in the Amazon S3 Management Console. Read the user guide to learn more.

Q: How does S3 Object Lambda work?

S3 Object Lambda uses Lambda functions specified by you to process the output of GET, LIST, and HEAD requests. Once you have defined a Lambda function to process requested data, you can attach that function to an S3 Object Lambda Access Point. GET, LIST, and HEAD requests made through an S3 Object Lambda Access Point will now invoke the specified Lambda function. Lambda will then fetch the S3 object requested by the client and process that object. Once processing has completed, Lambda will stream the processed object back to the calling client. Read the S3 Object Lambda user guide to learn more.

Q: How do I get started with S3 Object Lambda?

S3 Object Lambda can be set up in multiple ways. You can set up S3 Object Lambda in the S3 console by navigating to the Object Lambda Access Point tab. Next, create an S3 Object Lambda Access Point, the Lambda function that you would like S3 to execute against your GET, LIST, and HEAD requests, and a supporting S3 Access Point. Grant permissions to all resources to interact with Object Lambda. Lastly, update your SDK and application to use the new S3 Object Lambda Access Point to retrieve data from S3 using the language SDK of your choice. S3 Object Lambda will begin to process your GET, LIST, and HEAD requests. There are example Lambda function implementations in the AWS documentation to help you get started with the service.

You can also use AWS CloudFormation to automate your S3 Object Lambda configuration. When you use the AWS CloudFormation template, the Lambda function that is deployed in your account will pass S3 objects back to your requesting client or application without any changes. You can add custom code to modify and process data as it is returned to an application. To learn more, visit the S3 Object Lambda User Guide.

Q: What kinds of operations can I perform with S3 Object Lambda?

Any operation supported in a Lambda function is supported with S3 Object Lambda. This gives you a wide range of available options for processing your requests. You supply your own Lambda function to run custom computations against GET, LIST, and HEAD requests, giving you the flexibility to process data according to the needs of your application. Lambda processing time is limited to a maximum of 60 seconds. For more details, please see the S3 Object Lambda documentation here.

Q: Which S3 request types does S3 Object Lambda support?

S3 Object Lambda supports GET, LIST and HEAD requests. Any other S3 API calls made to an S3 Object Lambda Access Point will return the standard S3 API response. Learn more about S3 Object Lambda in the user guide.

Q: What will happen when a S3 Object Lambda function fails?

When a S3 Object Lambda function fails, you will receive a request response detailing the failure. Like other invocations of Lambda functions, AWS also automatically monitors functions on your behalf, reporting metrics through Amazon CloudWatch. To help you troubleshoot failures, Lambda logs all requests processed by your function and automatically stores logs generated by your code with Amazon CloudWatch Logs. For more information about accessing CloudWatch logs for AWS Lambda, please visit CloudWatch documentation.

Q: Does S3 Object Lambda affect the S3 availability SLA or S3 durability?

S3 Object Lambda connects Amazon S3, AWS Lambda, and optionally, other AWS services of your choosing to deliver objects relevant to requesting applications. All AWS services used in connection with S3 Object Lambda will continue to be governed by their respective Service Level Agreements (SLA). For example, in the event that any AWS Service does not meet its Service Commitment, you will be eligible to receive a Service Credit as documented in that service's SLA. Creating an S3 Object Lambda Access Point does not impact the durability of your objects. However, S3 Object Lambda invokes your specified AWS Lambda function and you must ensure your specified Lambda function is intended and correct. See the latest Amazon S3 SLA here.

Q: How much does S3 Object Lambda cost?

When you use S3 Object Lambda, you pay a per GB charge for every gigabyte of data returned to you through S3 Object Lambda. You are also charged for requests based on the request type (GET, LIST, and HEAD requests) and AWS Lambda compute charges for the time your specified function is running to process the requested data.

Amazon S3 Glacier

Amazon Simple Storage Service Glacier (Amazon S3 Glacier) is a storage service optimized for infrequently used data, or "cold data." The service provides durable and extremely low-cost storage with security features for data archiving and backup. With Amazon S3 Glacier, you can store your data cost effectively for months, years, or even decades. Amazon S3 Glacier enables you to offload the administrative burdens of operating and scaling storage to AWS, so you don't have to worry about capacity planning, hardware provisioning, data replication, hardware failure detection and recovery, or time-consuming hardware migrations. For more service highlights and pricing information, go to the Amazon S3 Glacier detail page.

Amazon S3 Glacier

Amazon S3 Glacier is a secure, durable, and extremely low-cost cloud storage service for data archiving and long-term backup. It is designed to deliver 99.999999999 percent durability, with comprehensive security and compliance capabilities that can help meet even the most stringent regulatory requirements. Amazon S3 Glacier provides query-in-place functionality, allowing you to run powerful analytics directly on your archived data at rest.

Data access options

To keep costs low yet suitable for varying retrieval needs, Amazon S3 Glacier provides three options for access to archives that span a few minutes to several hours.

Storage and bandwidth include all file overhead

Rate tiers take into account your aggregate usage for Data Transfer Out to the internet across Amazon EC2, Amazon S3, Amazon Glacier, Amazon RDS, Amazon SimpleDB, Amazon SQS, Amazon SNS, Amazon DynamoDB, and AWS Storage Gateway.

Data transfer

Data transfer in to Amazon S3 is free. Data transfer out of Amazon S3 is priced by Region.

What Is Amazon S3 Glacier?

Amazon S3 Glacier (S3 Glacier) is a secure and durable service for low-cost data archiving and long-term backup.

With S3 Glacier, you can store your data cost effectively for months, years, or even decades. S3 Glacier helps you offload the administrative burdens of operating and scaling storage to AWS, so you don't have to worry about capacity planning, hardware provisioning, data replication, hardware failure detection and recovery, or time-consuming hardware migrations.

Amazon Simple Storage Service (Amazon S3) also provides three Amazon S3 Glacier archive storage classes. These storage classes are designed for different access patterns and storage duration. These storage classes differ as follows:

- **S3 Glacier Instant Retrieval** – Use for archiving data that is rarely accessed and requires milliseconds retrieval.
- **S3 Glacier Flexible Retrieval** (formerly the S3 Glacier storage class) – Use for archives where portions of the data might need to be retrieved in minutes. Data stored in the S3 Glacier Flexible Retrieval storage class can be accessed in as little as 1-5 minutes by using Expedited retrieval. You can also request free Bulk retrievals in up to 5-12 hours.
- **S3 Glacier Deep Archive** – Use for archiving data that rarely needs to be accessed. Data stored in the S3 Glacier Deep Archive storage class has a default retrieval time of 12 hours.

Important

S3 Glacier Flexible Retrieval and S3 Glacier Deep Archive objects are not available for real-time access. You must first restore S3 Glacier Flexible Retrieval and S3

Glacier Deep Archive objects before you can access them. For these storage classes, Amazon S3 supports restore requests at a rate of up to 1,000 transactions per second, per AWS account.

When you choose the S3 Glacier Flexible Retrieval or S3 Glacier Deep Archive storage class, your objects remain in Amazon S3. You can't access them directly through the Amazon S3 Glacier service.

For more information about the Amazon S3 Glacier storage classes, see Storage classes for archiving objects in the *Amazon Simple Storage Service User Guide*.

For an enhanced S3 Glacier experience, you can use the Amazon S3 console instead of the Amazon S3 Glacier service. Accessing and using the Amazon S3 Glacier storage classes through the Amazon S3 APIs and Amazon S3 console provides enhanced functionality for data management and cost optimization.

Amazon S3 supports lifecycle configuration on an S3 bucket, which enables you to transition objects to the S3 Glacier storage classes for archival. For more information about Amazon S3 lifecycle configuration and transitioning objects to the S3 Glacier storage classes, see Object Lifecycle Management and Transitioning Objects in the *Amazon Simple Storage Service User Guide*.

The Amazon S3 Glacier storage classes are only some of the different storage classes for Amazon S3. For a general overview of Amazon S3 core concepts, such as buckets, access points, storage classes and objects, see What is Amazon S3? in the *Amazon Simple Storage Service User Guide*.

Amazon S3 Glacier storage classes

Long-term, secure, durable storage classes for data archiving at the lowest cost and milliseconds access.

The Amazon S3 Glacier storage classes are purpose-built for data archiving, providing you with the highest performance, most retrieval flexibility, and the lowest cost archive storage in the cloud. All S3 Glacier storage classes provide virtually unlimited scalability and are designed for 99.999999999% (11 nines) of data durability. The S3 Glacier storage classes deliver options for the fastest access to your archive data and the lowest-cost archive storage in the cloud.

You can choose from three archive storage classes optimized for different access patterns and storage duration. For archive data that needs immediate access, such

as medical images, news media assets, or genomics data, choose the S3 Glacier Instant Retrieval storage class, an archive storage class that delivers the lowest cost storage with milliseconds retrieval. For archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases, choose S3 Glacier Flexible Retrieval (formerly S3 Glacier), with retrieval in minutes or free bulk retrievals in 5-12 hours. To save even more on long-lived archive storage such as compliance archives and digital media preservation, choose S3 Glacier Deep Archive, the lowest cost storage in the cloud with data retrieval within twelve hours.

Benefits

Retrievals as quick as milliseconds
The Amazon S3 Glacier storage classes provide retrieval options from milliseconds to hours to fit your performance needs. The S3 Glacier Instant Retrieval storage class delivers milliseconds retrieval for archives that need immediate access, such as medical images or news media assets. S3 Glacier Flexible Retrieval provides three retrieval options: expedited retrievals that typically complete in 1–5 minutes, standard retrievals that typically complete in 3–5 hours, and free bulk retrievals that return large amounts of data typically in 5–12 hours. The Amazon S3 Glacier Deep Archive storage class provides two retrieval options ranging from 12-48 hours.

Unmatched durability and scalability
The Amazon S3 Glacier storage classes run on the world's largest global cloud infrastructure with virtually unlimited scalability and are designed for 99.999999999% (11 nines) of durability. Data is redundantly stored across multiple Availability Zones that are physically separated within an AWS Region.

Most comprehensive security and compliance capabilities
The S3 Glacier storage classes offer sophisticated integration with AWS CloudTrail to log, monitor, and retain storage API call activities for auditing, and they support three different forms of encryption. These storage classes also support security standards and compliance certifications, including SEC Rule 17a-4, PCI-DSS, HIPAA/HITECH, FedRAMP, EU GDPR, and FISMA. Amazon S3 Object Lock enables WORM storage capabilities, helping satisfy compliance requirements for virtually every regulatory agency around the globe.

Lowest cost
All S3 Glacier storage classes are designed to be the lowest-cost storage for specific access patterns, allowing you to archive large amounts of data at a very low cost. This makes it feasible to retain all the data you want for use cases like data lakes, analytics, Internet of Things (IoT), machine learning, compliance, and media asset archiving. You pay only for what you need, with no minimum commitments or upfront fees.

Most supported by partners, vendors, and AWS services
In addition to integration with most AWS services, Amazon S3 object storage services include tens of thousands of consulting, systems integrator, and independent software vendor partners, with more joining every month. AWS Partner Network partners have adapted their services and software to work with Amazon S3 storage classes for solutions such as backup and recovery, archiving, and disaster recovery. No other cloud provider has more partners with solutions that are pre-integrated to work with their service.

Consistent across your data lifecycle
All the S3 Glacier storage classes are available in all AWS Regions, use the standard S3 APIs, and support all S3 features—such as S3 Storage Lens to view storage usage and activity metrics, Server-Side Encryption for object data, S3 Object Lock to prevent accidental deletion, and AWS PrivateLink to access S3 through a private endpoint in a VPC. You can also use S3 Lifecyle to transition data to any of the S3 storage classes to lower costs as data becomes infrequently or rarely accessed.

How it works - The S3 Glacier storage classes

Overview

The Amazon S3 Glacier storage classes are purpose-built for data archiving, providing you with the highest performance, most retrieval flexibility, and the lowest cost archive storage in the cloud. You can now choose from three archive storage classes optimized for different access patterns and storage duration.

Amazon S3 Glacier Instant Retrieval storage class

S3 Glacier Instant Retrieval delivers the lowest cost storage, up to 68% lower cost (than S3 Standard-Infrequent Access), for long-lived data that is accessed once per quarter and requires millisecond retrieval. It is designed for rarely accessed data that still needs immediate access in performance-sensitive use cases like image hosting, online file-sharing applications, medical imaging and health records, news media assets, and satellite and aerial imaging. S3 Glacier Instant Retrieval offers the high durability, high throughput, and similar low latency of S3 Standard-IA, with a lower per-GB storage price and slightly higher per-GB retrieval price. S3 Glacier Instant Retrieval is designed for 99.999999999% (11 9s) of data durability and 99.9% availability by redundantly storing data across multiple physically separated AWS Availability Zones in a given year.

Amazon S3 Glacier Flexible Retrieval storage class

S3 Glacier Flexible Retrieval delivers low-cost storage, up to 10% lower cost (than S3 Glacier Instant Retrieval), for archive data that is accessed 1-2 times per year and is retrieved asynchronously. S3 Glacier Flexible Retrieval (formerly S3 Glacier) is the ideal storage class for archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases. S3 Glacier Flexible Retrieval delivers the most flexible retrieval options that balance cost with access times ranging from minutes to hours and with free bulk retrievals. It is an ideal solution for backup, disaster recovery, offsite data storage needs, and for when some data needs to occasionally retrieved in minutes, and you don't want to worry about costs. S3 Glacier Flexible Retrieval is designed for 99.999999999% (11 9s) of data durability and 99.99% availability by redundantly storing data across multiple physically separated AWS Availability Zones in a given year.

Amazon S3 Glacier Deep Archive storage class

S3 Glacier Deep Archive delivers the lowest cost storage, up to 75% lower cost (than S3 Glacier Flexible Retrieval), for long-lived archive data that is accessed less than once per year and is retrieved asynchronously. At just $0.00099 per GB-month

(or $1 per TB-month), S3 Glacier Deep Archive offers the lowest cost storage in the cloud, at prices significantly lower than storing and maintaining data in on-premises tape or archiving data off-site. S3 Glacier Deep Archive is a cost-effective and easy-to-manage alternative to tape. It is designed for customers — particularly those in the financial services, healthcare, media and entertainment and public sector — that retain data sets for 7-10 years or longer to meet customer needs and regulatory compliance requirements. S3 Glacier Deep Archive is designed for 99.999999999% (11 9s) of data durability and 99.99% availability by redundantly storing data across multiple physically separated AWS Availability Zones in a given year.

Use cases

Media asset workflows

Media and entertainment assets such as video and news footage require durable storage and can grow to many petabytes over time. Much of this data needs to be available immediately for breaking news events, video rendering, or content development. The Amazon S3 Glacier Instant Retrieval storage class allows you to archive older media content affordably while still making it available in milliseconds when it's needed. To save even more on storage costs for media archives that don't require milliseconds access, you can use S3 Glacier or S3 Glacier Deep Archive.

Healthcare information archiving

Hospital systems need to retain petabytes of patient records (such as LIS, PACS, and EHR) for decades to meet regulatory requirements. The Amazon S3 Glacier and S3 Glacier Deep Archive storage classes help you reliably archive patient record data securely at a very low cost. The Amazon S3 Glacier Instant Retrieval storage class is ideal for medical images or genomics, where milliseconds retrieval is required.

Regulatory, compliance, and business policy archiving

Many enterprises like financial services and healthcare must retain regulatory and compliance archives for extended durations. Amazon S3 Object Lock helps you set compliance controls to meet your objectives, such as SEC Rule 17a-4(f). Company policies can also require record retention of several years or indefinitely for data such as finance, tax, HR, or email records that can use S3 Glacier storage classes for a long-term content archive.

Scientific data storage

Research organizations generate, analyze, and archive vast amounts of data, such as for genomics or to train machine learning models. With Amazon S3 Glacier Instant Retrieval storage class, you avoid the complexities of hardware and facility management and capacity planning, with the lowest-cost storage and milliseconds retrieval.

Digital preservation
Libraries and government agencies face data integrity challenges in their digital preservation efforts. Unlike traditional systems, which can require laborious data verification and manual repair, Amazon S3 performs regular, systematic data integrity checks and is built to be automatically self-healing. Using S3 Lifecyle, you can transition data to lower-cost storage classes to optimize your savings.

Long-term backup retention
As part of backup plans, businesses store large data backups for many years on expensive on-premises storage systems. The S3 Glacier storage classes deliver the lowest-cost archive storage for these backups, while also making the data available when needed.

Tape replacement
On-premises or offsite tape libraries can lower storage costs but require large upfront investments and specialized maintenance. The Amazon S3 Glacier storage classes have no upfront costs and eliminate the cost and burden of maintenance, while providing faster access than on-premises tape archives.

Amazon Elastic Block Store (EBS)

Easy to use, high performance block storage at any scale.

How it works

Amazon Elastic Block Store (Amazon EBS) is an easy-to-use, scalable, high-performance block-storage service designed for Amazon Elastic Compute Cloud (Amazon EC2).

Use cases

Build your SAN in the cloud for I/O intensive applications
Migrate mid-range, on-premises storage area network (SAN) workloads to the cloud. Attach high-performance and high-availability block storage for mission-critical applications.

Run relational or NoSQL databases
Deploy and scale your choice of databases, including SAP HANA, Oracle, Microsoft SQL Server, PostgreSQL, MySQL, Cassandra, and MongoDB.

Right-size your big data analytics engines

Easily resize clusters for big data analytics engines, such as Hadoop and Spark, and freely detach and reattach volumes.

Amazon Elastic Block Store (Amazon EBS)

Amazon Elastic Block Store (Amazon EBS) provides block level storage volumes for use with EC2 instances. EBS volumes behave like raw, unformatted block devices. You can mount these volumes as devices on your instances. EBS volumes that are attached to an instance are exposed as storage volumes that persist independently from the life of the instance. You can create a file system on top of these volumes, or use them in any way you would use a block device (such as a hard drive). You can dynamically change the configuration of a volume attached to an instance.

We recommend Amazon EBS for data that must be quickly accessible and requires long-term persistence. EBS volumes are particularly well-suited for use as the primary storage for file systems, databases, or for any applications that require fine granular updates and access to raw, unformatted, block-level storage. Amazon EBS is well suited to both database-style applications that rely on random reads and writes, and to throughput-intensive applications that perform long, continuous reads and writes.

With Amazon EBS, you pay only for what you use.

Features of Amazon EBS

- You create an EBS volume in a specific Availability Zone, and then attach it to an instance in that same Availability Zone. To make a volume available outside of the Availability Zone, you can create a snapshot and restore that snapshot to a new volume anywhere in that Region. You can copy snapshots to other Regions and then restore them to new volumes there, making it easier to leverage multiple AWS Regions for geographical expansion, data center migration, and disaster recovery.
- Amazon EBS provides the following volume types: General Purpose SSD, Provisioned IOPS SSD, Throughput Optimized HDD, and Cold HDD. For more information, see EBS volume types.
  The following is a summary of performance and use cases for each volume type.
    - General Purpose SSD volumes (gp2 and gp3) balance price and performance for a wide variety of transactional workloads. These

volumes are ideal for use cases such as boot volumes, medium-size single instance databases, and development and test environments.
    - Provisioned IOPS SSD volumes (io1 and io2) are designed to meet the needs of I/O-intensive workloads that are sensitive to storage performance and consistency. They provide a consistent IOPS rate that you specify when you create the volume. This enables you to predictably scale to tens of thousands of IOPS per instance. Additionally, io2 volumes provide the highest levels of volume durability.
    - Throughput Optimized HDD volumes (st1) provide low-cost magnetic storage that defines performance in terms of throughput rather than IOPS. These volumes are ideal for large, sequential workloads such as Amazon EMR, ETL, data warehouses, and log processing.
    - Cold HDD volumes (sc1) provide low-cost magnetic storage that defines performance in terms of throughput rather than IOPS. These volumes are ideal for large, sequential, cold-data workloads. If you require infrequent access to your data and are looking to save costs, these volumes provides inexpensive block storage.
- You can create your EBS volumes as encrypted volumes, in order to meet a wide range of data-at-rest encryption requirements for regulated/audited data and applications. When you create an encrypted EBS volume and attach it to a supported instance type, data stored at rest on the volume, disk I/O, and snapshots created from the volume are all encrypted. The encryption occurs on the servers that host EC2 instances, providing encryption of data-in-transit from EC2 instances to EBS storage. For more information, see Amazon EBS encryption.
- You can create point-in-time snapshots of EBS volumes, which are persisted to Amazon S3. Snapshots protect data for long-term durability, and they can be used as the starting point for new EBS volumes. The same snapshot can be used to create as many volumes as needed. These snapshots can be copied across AWS Regions. For more information, see Amazon EBS snapshots.
- Performance metrics, such as bandwidth, throughput, latency, and average queue length, are available through the AWS Management Console. These metrics, provided by Amazon CloudWatch, allow you to monitor the performance of your volumes to make sure that you are providing enough performance for your applications without paying for resources you don't

need. For more information, see Amazon EBS volume performance on Linux instances.

What is block storage?

Block storage is technology that controls data storage and storage devices. It takes any data, like a file or database entry, and divides it into blocks of equal sizes. The block storage system then stores the data block on underlying physical storage in a manner that is optimized for fast access and retrieval. Developers prefer block storage for applications that require efficient, fast, and reliable data access. Think of block storage as a more direct pipeline to the data. By contrast, file storage has an extra layer consisting of a file system (NFS, SMB) to process before accessing the data.

What are the benefits of block storage?

Organizations use block level storage because of the following advantages.

Performance

Metadata is additional data that describes the primary data contained in the storage system. Block storage uses limited metadata but relies on unique identifiers assigned to each block for read/write operations. This reduces data transfer overhead and allows the server to efficiently access and retrieve data in block storage. Because block storage metadata is limited, block storage delivers ultra-low latency required for high-performance workloads. This is required for latency sensitive applications like databases. For example, Viasat uses Amazon Elastic Block Store (Amazon EBS) to capture high throughput (highly transactional) data and optimize storage costs. Organizations use Amazon EBS for performance and cost optimization, scale and agility, and for data protection with EBS Snapshots.

Block storage architecture provides multiple paths to the data whereas file storage only provides one path, this is why block storage is preferred for high performance applications.

Flexibility and scalability

Block storage devices are not constrained to specific network environments. Individual blocks can be configured for different operating systems, such as Windows or Linux. Developers can share data across multiple environments to ensure high availability. The block storage architecture is also highly scalable. Developers can add new blocks to existing ones to meet growing capacity needs.

Frequent modification

Block storage supports frequent data writes without affecting performance. Instead of rewriting the entire file, the system identifies the particular block that needs to be amended. Then, it rewrites the selected block with the new data. This makes block storage very efficient for managing large files that require frequent updates.

Granular control

Developers gain a high degree of control over storing data on block storage. For example, they can optimize performance by grouping fast-changing data on specific blocks and storing static files on others. This improves system performance as ongoing updates only affect a small number of data blocks instead of an entire file. For example, block storage gives you the

flexibility to tier fast-changing data on solid state disk (SSD) for the highest performance, and store warm or cold data on lower cost hard drives (HDD).

What are the use cases of block storage?

Block storage's unique characteristics make it the preferred option for transactional, mission-critical, and I/O intensive applications. Block storage Is used for a wide variety of applications including; relational or transactional databases, time series databases, containers, boot disks, and hypervisor files systems.

Storage area networks

Developers often deploy block storage as a storage area network (SAN). SAN is a complex network technology that presents block storage to multiple networked systems as if those blocks were locally attached devices. SAN's typically use fiber channel interconnects. In contrast, a network attached storage (NAS) is a single device that serves files over Ethernet.

The SAN architecture consists of three layers:

- Host layer consists of the servers that manage storage access
- Storage layer consists of physical block storage devices like magnetic tape, disk drives, or optical media
- Fabric layer bridges SAN servers and SAN storage with devices like SAN switches, protocol bridges, routers, cables, and gateway devices

It is Important to note that SAN's employ redundancy using either synchronous or asynchronous replication across long distances. This mitigates downtime in the event that a geographic location can't be accessed.

The SAN architecture can work with several types of storage in a unified environment, including block storage. Block storage provides a high-efficiency alternative to file storage on SANs.

Containers

Developers use block storage to store containerized applications on the cloud. Containers are software packages that contain the application and its resource files for deployment in any computing environment. Like containers, block storage is equally flexible, scalable, and efficient. With block storage, developers can migrate the containers seamlessly between servers, locations, and operating environments.

Transactional workloads

Transactional workloads are sequences of data generated at specific points of business processes. For example, sales records, operation logs, and login alerts are transactional workloads. Organizations that process time-sensitive and mission-critical transactions store such workloads into a low-latency, high-capacity, and fault-tolerant database.

Block storage allows developers to set up a robust, scalable, and highly efficient transactional database. As each block is a self-contained unit, the database performs optimally, even when the stored data grows. Furthermore, individual storage blocks can be hosted at different servers, preventing access bottlenecks.

In mission-critical applications, block storage is secured by a redundant array of independent disks (RAID) to ensure data redundancy. The RAID system backs up data files in secondary

storage and recovers the copy if the primary disk fails. This ensures that the application remains uninterrupted when storing and retrieving transactional workloads on block storage.

Analytics and data warehousing

Block storage is used with Hadoop's HDFS architecture (Hadoop Distributed File System) to store data as independently distributed units enabling performance for Hadoop and Kafka analytics applications.

Virtual machines

A virtual machine (VM) is technology that allows a computer to run a separate operating environment with software-defined computing resources. For example, you can run a Linux operating system on a Windows desktop with a VM. A hypervisor is an abstraction layer responsible for allocating the required memory, drive, and computing services to run the secondary operating environment.

Block storage supports popular VM hypervisors. Users can install the operating system, file system, and other computing resources on a block storage volume. They do so by formatting the block storage volume and turning it into a VM file system. This allows them to easily increase or decrease the virtual drive size and transfer the virtualized storage from one host to another.

How does block storage work?

In a block storage system, you can break the data into independent fixed-size blocks or pieces. Each block is an individual piece of data storage. A complete piece of information, such as a data file, is stored in multiple, nonsequential blocks.

The block storage system does not maintain high-level metadata, such as file type, ownership, and timestamp. Developers must design a data lookup table in the application system to manage the storage of data into respective blocks. The application might store data in different operating environments to increase read/write efficiency.

Data write

During a write sequence, the application splits data into several block-sized sections. It writes the data into multiple blocks and records the block’s identifier in a data lookup table. The lookup table allows the server to calculate the relative address of data stored in the block.

Data read

When users request a specific file from the block storage system, the server uses the data-lookup table to determine where pieces of the data are stored. Then, the application retrieves the data from multiple blocks and merges them in the original sequence.

What other types of storage are available?

In addition to block storage, there are also object and file storage options. Each type offers its own unique advantages.

Object storage

Object storage is a technology that stores and manages data in an unstructured format called objects. Each object is tagged with a unique identifier and contains metadata that describes the underlying content. For example, object storage for photos contains metadata regarding the photographer, resolution, format, and creation time. Developers use object storage to store unstructured data, such as text, video, and images.

Block storage vs object storage

Both storage solutions are beneficial depending on the use case. Block storage provides low latency and high-performance values in various use cases. Its features are primarily useful for structured database storage, VM file system volumes, and high volumes of read and write loads. Object storage is best used for large amounts of unstructured data, especially when durability, unlimited storage, scalability, and complex metadata management are relevant factors for overall performance.

File storage

File storage stores data in a hierarchical structure of files and folders. In network environments, file-based storage often uses network-attached storage (NAS) technology. NAS allows users to access network storage data in similar ways to a local hard drive. File storage is user-friendly and allows users to manage file-sharing control.

Block storage vs file storage

The file storage system stores data in a specific environment, while block storage systems can be integrated with different operating systems. File storage provides an intuitive interface for end-user computing. Meanwhile, you can add new data blocks to the block storage system without increasing operational latency.

Instance storage

An instance store provides temporary block-level storage for your instance. This storage is located on disks that are physically attached to the host computer. Instance store is ideal for temporary storage of information that changes frequently, such as buffers, caches, scratch data, and other temporary content, or for data that is replicated across a fleet of instances, such as a load-balanced pool of web servers.

An instance store consists of one or more instance store volumes exposed as block devices. The size of an instance store as well as the number of devices available varies by instance type

How can AWS support your block storage needs?

Amazon EBS is an easy-to-use block storage solution for cloud workloads. Developers use Amazon EBS to provide a persistent storage service for Amazon Elastic Compute Cloud (Amazon EC2) workloads.

- Amazon EBS provides a highly scalable storage solution for mission-critical and I/O intensive applications.
- Amazon EBS Snapshots provide an easy and secure method for block storage data protection.

- Developers can install various types of databases on Amazon EBS, including SAP HANA, Oracle, Microsoft SQL Server, MySQL, Cassandra, and MongoDB.

Amazon Elastic Block Store (EBS) FAQs

General

Q: Are Amazon EBS volume and snapshot ID lengths changing in 2018?

Yes, please visit the EC2 FAQs page for more details.

Q: What happens to my data when an Amazon EC2 instance terminates?

Unlike the data stored on a local instance store (which persists only as long as that instance is alive), data stored on an Amazon EBS volume can persist independently of the life of the instance. Therefore, we recommend that you use the local instance store only for temporary data. For data requiring a higher level of durability, we recommend using Amazon EBS volumes or backing up the data to Amazon S3. If you are using an Amazon EBS volume as a root partition, set the Delete on termination flag to "No" if you want your Amazon EBS volume to persist outside the life of the instance.

Q: What kind of performance can I expect from Amazon EBS volumes?

Amazon EBS provides seven volume types: Provisioned IOPS SSD (io2 Block Express, io2, and io1), General Purpose SSD (gp3 and gp2), Throughput Optimized HDD (st1) and Cold HDD (sc1). These volume types differ in performance characteristics and price, allowing you to tailor your storage performance and cost to the needs of your applications. The average latency between EC2 instances and EBS is single digit milliseconds. For more performance information see the EBS product details page. For more information about Amazon EBS performance guidelines, see Increasing EBS Performance.

Q: Which volume should I choose?

Amazon EBS includes two major categories of storage: SSD-backed storage for transactional workloads (performance depends primarily on IOPS, latency, and durability) and HDD-backed storage for throughput workloads (performance depends primarily on throughput, measured in MB/s). SSD-backed volumes are designed for transactional, IOPS-intensive database workloads, boot volumes, and workloads that require high IOPS. SSD-backed volumes include Provisioned IOPS SSD (io1 and io2) and General Purpose SSD (gp3 and gp2). Both io2 and io2 Block Express of the Provisioned IOPS SSD volumes are designed to provide 100X durability of 99.999% making it ideal for business-critical applications that need higher uptime. gp3 is the latest generation of General Purpose SSD volumes that provides the right balance of price and performance for most applications that don't require the highest IOPS performance or 99.999% durability. HDD-backed volumes are designed for throughput-intensive and big-data workloads, large I/O sizes, and sequential I/O patterns. HDD-backed volumes include Throughput Optimized HDD (st1) and Cold HDD (sc1).

Q: Since io2 provides higher volume durability, should I still take snapshots and plan to replicate io2 volumes across Availability Zones (AZs) for high durability?

High volume durability, snapshots, and replicating volumes across AZs protect against different types of failures, and customers can choose to use one, two, or all of these approaches based on their data durability requirements. Higher volume durability reduces the probability of losing the primary copy of your data. Snapshots protect against the unlikely event of a volume failure.

Replicating volumes across AZs protects against an AZ level failure and also provides faster recovery in case of failure.

Q. What are best practices for high availability on Amazon EBS?

Amazon EBS volumes are designed to be highly available, reliable, and durable. At no additional charge to you, Amazon EBS volume data is replicated across multiple servers in an Availability Zone to prevent the loss of data from the failure of any single component. Depending on the degree of high availability (HA) that your application requires, we recommend these guidelines to achieve a robust degree of high availability:
1) Design the system to have no single point of failure. For more details, see High Availability and Scaling on AWS.
2) Use automated monitoring, failure detection, and failover mechanisms. See Monitoring the Status of your EBS volumes and Monitoring EBS Volumes using CloudWatch for more details on monitoring your EBS Volume's performance.
3) Prepare operating procedures for manual mechanisms to respond to, mitigate, and recover from any failures. This includes detaching unavailable volumes and attaching a backup recovery volume in cases of failure. For more details, see the documentation on Replacing an EBS volume.

Q: How do I modify the capacity, performance, or type of an existing EBS volume?

Changing a volume configuration is easy. The Elastic Volumes feature allows you to increase capacity, tune performance, or change your volume type with a single CLI call, API call or a few console clicks. For more information about Elastic Volumes, see the Elastic Volumes documentation.

Q: Are EBS Standard Volumes still available?

EBS Standard Volumes have been renamed to EBS Magnetic volumes. Any existing volumes will not have been changed as a result of this and there are no functional differences in the EBS Magnetic offering compared to EBS Standard. The name of this offering was changed to avoid confusion with our General Purpose SSD (gp2) volume type which is our recommended default volume type.

Q: Are Provisioned IOPS SSD (io2 Block Express, io2, and io1) volumes available for all Amazon EC2 instance types?

Provisioned IOPS SSD io1 volumes are available for all Amazon EC2 Instance Types, whereas Provisioned IOPS SSD io2 volumes are available on all EC2 Instances Types, with the exception of X2idn, X2iedn, R5b, C7g, and Trn1 instances. io2 Block Express volumes are currently available only on X2idn, X2iedn, R5b, C7g, and Trn1 instances. Use EBS optimized EC2 instances to deliver consistent and predictable IOPS on io2 and io1 volumes. EBS-optimized instances deliver dedicated throughput between Amazon EC2 and Amazon EBS, with options between 62.5 MB/s and 7,500 MB/s depending on the instance type used. To achieve the limit of 64,000 IOPS and 1,000 MB/s throughput, the volume must be attached to a Nitro System-based EC2 instance.

Q: What is the difference between io2 and io2 Block Express?

io2 volumes offer high performance block storage for all EC2 instances. For applications that require even higher performance, you can attach io2 volumes to X2idn, X2iedn, R5b, C7g, and Trn1 instance types which run on Block Express and provides 4x higher performance than io2. This will enable you to achieve up to 64 TiB capacity, 256,000 IOPS and 4,000 MB/s of throughput from a single io2 volume along with sub-millisecond average IO latency.

Performance

Q: What level of performance consistency can I expect to see from my Provisioned IOPS SSD (io2 and io1) volumes?

When attached to EBS-optimized instances, Provisioned IOPS SSD (io2 and io1) volumes are designed to deliver within 10% of the provisioned IOPS performance 99.9% of the time in a given year. Your exact performance depends on your application's I/O requirements.

Q: What level of performance latency can I expect to see from my Provisioned IOPS SSD (io2 and io1) volumes?

When attached to EBS-optimized instances, Provisioned IOPS (io1 and io2) volumes can achieve single digit millisecond latencies and io2 Block Express volumes can achieve sub-millisecond latency. Your exact performance depends on your application's I/O requirements.

Q: Does the I/O size of my application reads and writes affect the rate of IOPS I get from my Provisioned IOPS SSD (io2 and io1) volumes?

Yes, it does. When you provision IOPS for io2 or io1 volumes, the IOPS rate you get depends on the I/O size of your application reads and writes. Provisioned IOPS volumes have a base I/O size of 16KB. So, if you have provisioned a volume with 40,000 IOPS for an I/O size of 16KB, it will achieve up to 40,000 IOPS at that size. If the I/O size is increased to 32 KB, then you will achieve up to 20,000 IOPS, and so on. For more details, please visit technical documentation on Provisioned IOPS volumes. You can use Amazon CloudWatch to monitor your throughput and I/O sizes.

Q: What factors can affect the performance consistency I see with Provisioned IOPS SSD (io2 and io1) volumes?

Provisioned IOPS SSD (io2 and io1) volumes attached to EBS-optimized instances are designed to offer consistent performance, delivering within 10% of the provisioned IOPS performance 99.9% of the time over a given year. For maximum performance consistency with new volumes created from a snapshot, we recommend enabling Fast Snapshot Restore (FSR) on your snapshots. EBS volumes restored from FSR-enabled snapshots instantly receive their full performance.

Another factor that can impact your performance is if your application isn't sending enough I/O requests. This can be monitored by looking at your volume's queue depth. The queue depth is the number of pending I/O requests from your application to your volume. For maximum consistency, a Provisioned IOPS volume must maintain an average queue depth (rounded to the nearest whole number) of one for every 1000 provisioned IOPS in a minute. For example, for a volume provisioned with 3000 IOPS, the queue depth average must be 3. For more information about ensuring consistent performance of your volumes, see Increasing EBS Performance.

Q: What level of performance consistency can I expect to see from my HDD-backed volumes?

When attached to EBS-optimized instances, Throughput Optimized HDD (st1) and Cold HDD (sc1) volumes are designed to deliver within 10% of the expected throughput performance 99% of the time in a given year. Your exact performance depends on your application's I/O requirements and the performance of your EC2 instance.

Q: Does the I/O size of my application reads and writes affect the rate of throughput I get from my HDD-backed volumes?

Yes. The throughput rate you get depends on the I/O size of your application reads and writes. HDD-backed volumes process reads and writes in I/O sizes of 1MB. Sequential I/Os are merged and processed as 1 MB units while each non-sequential I/O is processed as 1MB even if the actual I/O size is smaller. Thus, while a transactional workload with small, random IOs, such as a database, won't perform well on HDD-backed volumes, sequential I/Os and large I/O sizes will achieve the advertised performance of st1 and sc1 for a longer period of time.

Q: What factors can affect the performance consistency of my HDD-backed volumes?

Throughput Optimized HDD (st1) and Cold HDD (sc1) volumes attached to EBS-optimized instances are designed to offer consistent performance, delivering within 10% of the expected throughput performance 99% of the time in a given year. There are several factors that could affect the level of consistency you see. For example, the relative balance between random and sequential I/O operations on the volume can impact your performance. Too many random small I/O operations will quickly deplete your I/O credits and lower your performance down to the baseline rate. Your throughput rate may also be lower depending on the instance selected. Although st1 can drive throughput up to 500 MB/s, performance will be limited by the separate instance-level limit for EBS traffic. Another factor is taking a snapshot which will decrease expected write performance down to the baseline rate, until the snapshot completes. This is specific to st1 and sc1.

Your performance can also be impacted if your application isn’t sending enough I/O requests. This can be monitored by looking at your volume’s queue depth and I/O size. The queue depth is the number of pending I/O requests from your application to your volume. For maximum consistency, HDD-backed volumes must maintain an average queue depth (rounded to the nearest whole number) of four or more for every 1 MB sequential I/O. For more information about ensuring consistent performance of your volumes, see Increasing EBS Performance.

Q: Can I stripe multiple volumes together to get better performance?

Yes. You can stripe multiple volumes together to achieve up to 260,000 IOPS or 60,000 Mbps (or 7500 MB/s) when attached to larger EC2 instances. However, performance for st1 and sc1 scales linearly with volume size so there may not be as much of a benefit to stripe these volumes together.

Q: How does Amazon EBS handle issues like storage contention?

EBS is a multi-tenant block storage service. We employ rate limiting as a mechanism to avoid resource contention. This starts with having defined performance criteria for the volumes – our volume types (gp2, PIOPS, st1, and sc1) all have defined performance characteristics in terms of IOPS and throughput. The next step is defining performance at the instance level. Each EBS Optimized instance has defined performance (both throughput and IOPS) for the set of EBS volumes attached to the instance. A customer can, therefore, size instances and volumes to get the desired level of performance. In addition, customers can use our reported metrics to observe instance level and volume level performance. They can set alarms to determine if what they are seeing does not match the expected performance – the metrics can also help determine if customers are configured at the right type of instance with the right amount of performance at the volume level or not. On the EBS end, we use the configured performance to inform how we allocate the appropriate instance and EBS infrastructure to support the volumes. By appropriately allocating infrastructure, we avoid resource contention. Additionally, we constantly monitor our infrastructure. This monitoring allows us to detect infrastructure failure (or imminent infrastructure failure) and therefore, move the volumes pro-actively to functioning hardware while the underlying infrastructure is either repaired or replaced (as appropriate).

Q: What level of performance consistency can I expect to see from my General Purpose SSD (gp3 and gp2) volumes?

When attached to EBS-optimized instances, General Purpose SSD (gp3 and gp2) volumes are designed to deliver within 10% of the provisioned IOPS performance 99% of the time in a given year. Your exact performance depends on your application's I/O requirements.

Q: What level of performance latency can I expect to see from my General Purpose SSD (gp3 and gp2) volumes?

When attached to EBS-optimized instances, General Purpose SSD (gp3 and gp2) volumes can achieve single digit millisecond latencies. Your exact performance depends on your application's I/O requirements.

Q: Do General Purpose SSD (gp3) volumes have burst?

No. All General Purpose SSD (gp3) volumes include 3,000 IOPS and 125 MB/s of consistent performance at no additional cost. Volumes can sustain the full 3,000 IOPS and 125 MB/s indefinitely.

Q: How does burst work on General Purpose SSD (gp2) volumes?

General Purpose SSD (gp2) volumes that are under 1,000 GB receive burst IOPS performance up to 3,000 IOPS for at least 30 min of sustained performance. Additionally, gp2 volumes deliver consistent performance of 3 IOPS per provisioned GB. For example, a 500 GB volume is capable of driving 1,500 IOPS consistently, and bursting to 3,000 IOPS for 60 minutes (3,000 IOPS * 60 seconds * 30 minutes / 1,500 IOPS / 60 seconds).

Q. What is the difference between io2 and io2 Block Express?

io2 volumes offer high performance block storage for all EC2 instances. For applications that require even higher performance, you can attach io2 volumes to X2idn, X2iedn, R5b, C7g, or Trn1 instance type which run on Block Express and provides 4x higher performance than io2. This will enable you to achieve up to 64 TiB capacity, 256,000 IOPS and 4,000 MB/s of throughput from a single io2 volume along with sub-millisecond average IO latency.

Q. What is EBS Block Express?

EBS Block Express is the next generation of Amazon EBS storage server architecture purpose-built to deliver the highest levels of performance with sub-millisecond latency for block storage at cloud scale. Block Express does this by using Scalable Reliable Datagrams (SRD), a high-performance lower-latency network protocol, to communicate with Nitro System-based EC2 instances. This is the same high performance and low latency network interface that is used for inter-instance communication in Elastic Fabric Adapter (EFA) for High Performance Computing (HPC) and Machine Learning (ML) workloads. Additionally, Block Express offers modular software and hardware building blocks that can be assembled in many different ways, giving us the flexibility to design and deliver improved performance and new features at a faster rate.

Q. What workloads are suited for io2 Block Express?

io2 Block Express is suited for performance and capacity intensive workloads that benefit from lower latency, higher IOPS, higher throughput, or larger capacity in a single volume. These workloads include relational and NoSQL databases such as SAP HANA, Oracle, MS SQL, PostgreSQL, MySQL, MongoDB, Cassandra, and critical business operation workloads such as SAP Business Suite, NetWeaver, Oracle eBusiness, PeopleSoft, Siebel, and ERP workloads such as Infor LN and Infor M3.

Q. How do I know if an io2 volume is running on Block Express?

If an io2 volume is attached to X2idn, X2iedn, R5b, C7g, or Trn1 instance then it runs on Block Express, which offers sub-millisecond latency and capability to drive up to 256,000 IOPS and 4,000 MB/s throughput, and up to 64 TiB in size for a single volume. io2 volumes attached to all other instances do not run on Block Express and offer single-digit millisecond latency and capability to drive up to 64K IOPS and 1 GB/s throughput, and up to 16 TiB in size for a single volume.

Snapshots

Q: How can I use EBS direct APIs for Snapshots?

This feature can be used via the following APIs that can be called using AWS CLI or via AWS SDK.

- List Snapshot Blocks: The ListSnapshotBlocks API operation returns the block indexes and block tokens for blocks in the specified snapshot.
- List Changed Blocks: The ListChangedBlocks API operation returns the block indexes and block tokens for blocks that are different between two specified snapshots of the same volume/snapshot lineage.
- Get Snapshot Blocks: The GetSnapshotBlock API operation returns the data in a block for the specified snapshot ID, block index, and block token.
- Start Snapshot: The StartSnapshot operation starts a snapshot, either as an incremental snapshot of an existing one or as a new snapshot. The started snapshot remains in a pending state until it is completed using the CompleteSnapshot action.
- Put Snapshot Block: The PutSnapshot operation adds data in the form of individual blocks to a started snapshot that is in a pending state. You must specify a Base64-encoded SHA256 checksum for the block of data transmitted. The service validates the checksum after the transmission is completed. The request fails if the checksum computed by service doesn't match what you specified.
- Complete Snapshot: The CompleteSnapshot operation completes a started snapshot that is in a pending state. The snapshot is then changed to a completed state.

For more information, please refer to technical documentation.

Q: What block sizes are supported by GetSnapshotBlock and PutSnapshotBlock APIs?

GetSnapshotBlock and PutSnapshotBlock APIs support 512KiB block size.

Q: Will I be able to access my snapshots using the regular Amazon S3 API?

No, snapshots are only available through the Amazon EC2 API.

Q: Do volumes need to be un-mounted to take a snapshot?

No, snapshots can be done in real time while the volume is attached and in use. However, snapshots only capture data that has been written to your Amazon EBS volume, which might exclude any data that has been locally cached by your application or OS. To ensure consistent snapshots on volumes attached to an instance, we recommend detaching the volume cleanly, issuing the snapshot command, and then reattaching the volume. For Amazon EBS volumes that serve as root devices, we recommend shutting down the machine to take a clean snapshot.

Q: Does it take longer to snapshot an entire 16 TB volume as compared to an entire 1 TB volume?

By design, an EBS Snapshot of an entire 16 TB volume should take no longer than the time it takes to snapshot an entire 1 TB volume. However, the actual time taken to create a snapshot depends on several factors including the amount of data that has changed since the last snapshot of the EBS volume.

Q: Are snapshots versioned? Can I read an older snapshot to do a point-in-time recovery?

Each snapshot is given a unique identifier, and customers can create volumes based on any of their existing snapshots.

Q: How can I discover Amazon EBS snapshots that are shared with me?

You can find snapshots that are shared with you by selecting Private Snapshots from the list in the Snapshots section of the AWS Management Console. This section lists both snapshots that you own and snapshots that are shared with you.

Q: How can I find which Amazon EBS snapshots are shared globally?

You can find snapshots that are shared globally by selecting Public Snapshots from the list in the Snapshots section of the AWS Management Console.

Q: How can I find a list of Amazon public datasets stored in Amazon EBS Snapshots?

You can use the AWS Management Console to find public datasets stored as Amazon Snapshots. Log into the console, select the Amazon EC2 Service, select Snapshots and then filter on Public Snapshots. All information on public datasets is available in our AWS Public Datasets resource center.

Q: When would I use Fast Snapshot Restore (FSR)?

You should enable FSR on snapshots if you are concerned about latency of data access when you restore data from a snapshot to a volume and want to avoid the initial performance hit during initialization. FSR is intended to help with use cases such as virtual desktop infrastructure (VDI), backup & restore, test/dev volume copies, and booting from custom AMIs. By enabling FSR on your snapshot, you will see improved and predictable performance whenever you need to restore data from that snapshot.

Q: Does enabling FSR for my snapshot speed up snapshot creation?

No. FSR-enabled snapshots improve restoring backup data from your snapshot to your volumes. FSR-enabled snapshots do not speed up snapshot creation time.

Q: How do I enable Fast Snapshot Restore (FSR)?

To use the feature, invoke the new enable-fast-snapshot-restores API on a snapshot within the availability zone (AZ) where initialized volumes are to be restored.

The FSR-enabled snapshot may be in any one of the following states: enabling, optimizing, enabled, disabling, disabled. State transitions are published as CloudWatch events and the FSR state can be checked via the describe-fast-snapshot-restores API.

Enabling FSR on a snapshot does not change any existing snapshot API interactions, and existing workflows will not need to change. FSR can be enabled or disabled on account-owned snapshots only. FSR cannot be applied to shared snapshots. You can view the list of your FSR-enabled snapshots via API or the console.

Q: How do I use Fast Snapshot Restore (FSR)?

Volumes created from an FSR-enabled snapshot are fully initialized. However, there are limits on the number of volumes that can be created with immediate full performance. These limits are expressed in the form of a credit bucket that is associated with an FSR-enabled snapshot in a given AZ. The important things to know regarding credits:

1. A single volume create operation consumes a single credit
2. The number of credits is a function of the FSR-enabled snapshot size
3. Credits refill over time
4. Maximum credit bucket size is 10

To estimate your credit bucket size and fill rate, divide 1,024 by your snapshot size. For example, a 100 GiB FSR-enabled snapshot will have the maximum balance of 10 credits with a fill rate of 10 credits every hour. A 4 TiB snapshot will have a maximum balance of 1 with a fill rate of 1 credit every 4 hours.

It's important to note that the credit bucket size is a function of the FSR-enabled snapshot size, not the size of the volumes that are created. For example, it is possible to create up to ten 1TiB volumes from a 100GiB snapshot at once.

Lastly, each AZ in which the snapshot is FSR-enabled gets its own credit bucket independent of other AZs.

Q: How many concurrent volumes can I create and what happens when I surpass this limit?

The size of the create credit bucket represents the maximum number and the balance of the credit bucket represents the number of creates available. When filled, up to 10 initialized volumes can be created from an FSR-enabled snapshot at once. Both the maximum size of the credit bucket and the credit bucket balance are published as CloudWatch metrics. Volume creations beyond the limit will proceed as if FSR is not enabled on the snapshot.

Q: How do I know when a volume was created from an FSR-enabled snapshot?

When using FSR, a new EBS-specific attribute (fastRestored) is added in the DescribeVolumes API to denote the status at create time. When a volume is created from an FSR-enabled snapshot without sufficient volume-create credits, the create will succeed but the volume will not be initialized.

Q: What happens to FSR when I delete a snapshot?

When you delete a snapshot, the FSR for your snapshot is automatically disabled and FSR billing for the snapshot will be terminated.

Q: Can I enable FSR for public and private snapshots shared with me?

Yes, you can enable FSR for public snapshots as well as all private snapshots shared with your account. To enable FSR for shared snapshots, you can use the same set of API calls that you use for enabling FSR on snapshots you own.

Q: How am I billed for enabling FSR on a snapshot shared with me?

When you enable FSR on your shared snapshot, you will be billed at standard FSR rates. Note that only your account will be billed for the FSR of the shared snapshot. The owner of the snapshot will not get billed when you enable FSR on the shared snapshot.

Q: What happens to the FSR for a shared snapshot when the owner of the snapshot stops sharing the snapshot or deletes it?

When the owner of your shared snapshot deletes the snapshot, or stops sharing the snapshot with you by revoking your permissions to create volumes from this snapshot, the FSR for your shared snapshot is automatically disabled and FSR billing for the snapshot will be terminated.

Encryption

Q: What is Amazon EBS encryption?

Amazon EBS encryption offers seamless encryption of EBS data volumes, boot volumes and snapshots, eliminating the need to build and maintain a secure key management infrastructure. EBS encryption enables data at rest security by encrypting your data using Amazon-managed keys, or keys you create and manage using the AWS Key Management Service (KMS). The encryption occurs on the servers that host EC2 instances, providing encryption of data as it moves between EC2 instances and EBS storage. For more details, see Amazon EBS encryption in the Amazon EC2 User Guide.

Q: What is the AWS Key Management Service (KMS)?

AWS KMS is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data. AWS Key Management Service is integrated with other AWS services including Amazon EBS, Amazon S3, and Amazon Redshift, to make it simple to encrypt your data with encryption keys that you manage. AWS Key Management Service is also integrated with AWS CloudTrail to provide you with logs of all key usage to help meet your regulatory and compliance needs. To learn more about KMS, visit the AWS Key Management Service product page.

Q: Why should I use EBS encryption?

You can use Amazon EBS encryption to meet security and encryption compliance requirements for data at rest encryption in the cloud. Pairing encryption with existing IAM access control policies improves your company's defense-in-depth strategy.

Q: How are my Amazon EBS encryption keys managed?

Amazon EBS encryption handles key management for you. Each newly created volume gets a unique 256-bit AES key; Volumes created from the encrypted snapshots share the key. These keys are protected by our own key management infrastructure, which implements strong logical and physical security controls to prevent unauthorized access. Your data and associated keys are encrypted using the industry-standard AES-256 algorithm.

Q: Does EBS encryption support boot volumes?

Yes.

Q: Can I create an encrypted data volume at the time of instance launch?

Yes, using customer master keys (CMKs) that are either AWS-managed or customer-managed. You can specify the volume details and encryption through a RunInstances API call with the BlockDeviceMapping parameter or through the Launch Wizard in the EC2 Console.

Q: Can I create additional encrypted data volumes at the time of instance launch that are not part of the AMI?

Yes, you can create encrypted data volume with either default or custom CMK encryption at the time of instances launch. You can specify the volume details and encryption through BlockDeviceMapping object in RunInstances API call or through Launch Wizard in EC2 Console.

Q: Can I launch an encrypted EBS instance from an unencrypted AMI?

Yes. See technical documentation for details.

Q: Can I share encrypted snapshots and AMIs with other accounts?

Yes. You can share encrypted snapshots and AMIs using a customer-managed customer master key (CMK) with other AWS accounts. See technical documentation for details.

Q: Can I ensure that all new volumes created are always encrypted?

Yes, you can enable EBS encryption by default with a single setting per region. This ensures that all new volumes are always encrypted. Refer to technical documentation for more details.

Multi-Attach

Q: Is there an additional fee to enable Multi-Attach?

No. Multi-Attach can be enabled on an EBS Provisioned IOPS io1 volume and there will be charges for the storage (GB-Mo) and IOPS (IOPS-Mo) provisioned.

Q: Can I boot an EC2 instance using a Multi-Attach enabled volume?

No.

Q: What happens if all of my attached instances do not have the ‘deleteOnTermination’ flag set?

The volume's deleteOnTermination behavior is determined by the configuration of the last attached instance that is terminated. To ensure predictable delete on termination behavior, enable or disable 'deleteOnTermination' for all of the instances to which the volume is attached.

If you want the volume to be deleted when the attached instances are terminated, enable ‘deleteOnTermination’ for all of instances to which the volume is attached. If you want to retain the volume after the attached instances have been terminated, disable ‘deleteOnTermination’ for all attached instances. For more information, see Multi-Attach technical documentation.

Q: Can my application use Multi-Attach?

If your application does not require storage layer coordination of write operations, such as a read-only application or it enforces application level IO fencing, then your application can use Multi-Attach.

Amazon Elastic File System (EFS)

Simple, serverless, set-and-forget, elastic file system.

How it works

Amazon Elastic File System (Amazon EFS) automatically grows and shrinks as you add and remove files with no need for management or provisioning.

Use cases

Simplify DevOps
Share code and other files in a secure, organized way to increase DevOps agility and respond faster to customer feedback.

Modernize application development
Persist and share data from your AWS containers and serverless applications with zero management required.

Enhance content management systems

Simplify persistent storage for modern content management system (CMS) workloads. Get your products and services to market faster, more reliably, and securely at a lower cost.

Accelerate data science
Easy to use and scale, Amazon EFS offers the performance and consistency needed for machine learning (ML) and big data analytics workloads.

What is Amazon Elastic File System?

Amazon Elastic File System (Amazon EFS) provides a simple, serverless, set-and-forget elastic file system for use with AWS Cloud services and on-premises resources. It is built to scale on demand to petabytes without disrupting applications, growing and shrinking automatically as you add and remove files, eliminating the need to provision and manage capacity to accommodate growth. Amazon EFS has a simple web services interface that allows you to create and configure file systems quickly and easily. The service manages all the file storage infrastructure for you, meaning that you can avoid the complexity of deploying, patching, and maintaining complex file system configurations.

Amazon EFS supports the Network File System version 4 (NFSv4.1 and NFSv4.0) protocol, so the applications and tools that you use today work seamlessly with Amazon EFS. Multiple compute instances, including Amazon EC2, Amazon ECS, and AWS Lambda, can access an Amazon EFS file system at the same time, providing a common data source for workloads and applications running on more than one compute instance or server.

With Amazon EFS, you pay only for the storage used by your file system and there is no minimum fee or setup cost. Amazon EFS offers a range of storage classes designed for different use cases. These include:

- **Standard storage classes** – EFS Standard and EFS Standard–Infrequent Access (Standard–IA), which offer multi-AZ resilience and the highest levels of durability and availability.
- **One Zone storage classes** – EFS One Zone and EFS One Zone–Infrequent Access (EFS One Zone–IA), which offer customers the choice of additional savings by choosing to save their data in a single Availability Zone.

For more information, see EFS storage classes. Costs related to Provisioned Throughput are determined by the throughput values you specify.

Amazon EFS is designed to provide the throughput, IOPS, and low latency needed for a broad range of workloads. With Amazon EFS, you can choose from two performance modes and two throughput modes:

- The default *General Purpose performance mode* is ideal for latency-sensitive use cases, like web serving environments, content management systems, home directories, and general file serving.
- File systems in the *Max I/O mode* can scale to higher levels of aggregate throughput and operations per second with a tradeoff of higher latencies for file system operations. For more information, see Performance modes.
- Using the default *Bursting Throughput mode*, throughput scales as your file system grows.
- Using *Provisioned Throughput mode*, you can specify the throughput of your file system independent of the amount of data stored. For more information, see Throughput modes.

The service is designed to be highly scalable, highly available, and highly durable. Amazon EFS file systems using Standard storage classes store data and metadata across multiple Availability Zones in an AWS Region. EFS file systems can grow to petabyte scale, drive high levels of throughput, and allow massively parallel access from compute instances to your data.

Amazon EFS provides file system access semantics, such as strong data consistency and file locking. For more information, see Data consistency in Amazon EFS. Amazon EFS also enables you to control access to your file systems through Portable Operating System Interface (POSIX) permissions. For more information, see Security in Amazon EFS.

Amazon EFS supports authentication, authorization, and encryption capabilities to help you meet your security and compliance requirements. Amazon EFS supports two forms of encryption for file systems, encryption in transit and encryption at rest. You can enable encryption at rest when creating an Amazon EFS file system. If you do, all your data and metadata is encrypted. You can enable encryption in transit when you mount the file system. NFS client access to EFS is controlled by both AWS Identity and Access Management (IAM) policies and network security policies like security groups. For more information, see Data encryption in Amazon EFS, Identity and access management for Amazon EFS, and Controlling network access to Amazon EFS file systems for NFS clients.

Note

Using Amazon EFS with Microsoft Windows–based Amazon EC2 instances is not supported.

Amazon Elastic File System (EFS) FAQs

General

Q. What is Amazon Elastic File System?

Amazon Elastic File System (Amazon EFS) is a simple, serverless, set-and-forget elastic file system that makes it easy to set up, scale, and cost-optimize file storage in AWS. With a few clicks in the AWS Management Console, you can create file systems that are accessible to Amazon Elastic Compute Cloud (EC2) instances, Amazon container services (Amazon Elastic Container Service [ECS], Amazon Elastic Kubernetes Service [EKS], and AWS Fargate), and AWS Lambda functions through a file system interface (using standard operating system file I/O APIs). They also support full file system access semantics, such as strong consistency and file locking.

Amazon EFS file systems can automatically scale from gigabytes to petabytes of data without needing to provision storage. Tens, hundreds, or even thousands of compute instances can access an Amazon EFS file system at the same time, and Amazon EFS provides consistent performance to each compute instance. Amazon EFS is designed to be highly durable and highly available. With Amazon EFS, there is no minimum fee or setup costs, and you pay only for what you use.

Q. What use cases does Amazon EFS support?

Amazon EFS is designed to provide performance for a broad spectrum of workloads and applications, including big data and analytics, media processing workflows, content management, web serving, and home directories.

Amazon EFS Standard storage classes are ideal for workloads that require the highest levels of durability and availability.

Amazon EFS One Zone storage classes are ideal for workloads such as development, build, and staging environments, as well as applications such as analytics, simulation, and media transcoding, and for backups or replicas of on-premises data that don't require multi-AZ resilience.

Q. When should I use Amazon EFS vs. Amazon EBS vs. Amazon S3?

AWS offers cloud storage services to support a wide range of storage workloads.

Amazon EFS is a file storage service for use with Amazon compute (EC2, containers, serverless) and on-premises servers. EFS provides a file system interface, file system access semantics (such as strong consistency and file locking), and concurrently accessible storage for up to thousands of EC2 instances.

Amazon Elastic Block Store (EBS) is a block-level storage service for use with EC2. Amazon EBS can deliver performance for workloads that require the lowest-latency access to data from a single EC2 instance.

Amazon Simple Storage Service (S3) is an object storage service. Amazon S3 makes data available through an internet API that can be accessed anywhere.

Q. What Regions is Amazon EFS currently available in?

Please refer to Regional Products and Services for details of Amazon EFS service availability by Region.

Q. How do I get started using Amazon EFS?

To use Amazon EFS, you must have an AWS account. If you don't already have one, you can sign up for an AWS account and instantly get access to the AWS Free Tier.

Once you have created an AWS account, refer to the Amazon EFS Getting Started guide to begin using EFS. You can create a file system through the AWS Management Console, the AWS Command Line Interface (AWS CLI), and the Amazon EFS API (and various language-specific SDKs).

Q. How do I access a file system from an Amazon EC2 instance?

To access your file system, mount the file system on an Amazon EC2 Linux-based instance using the standard Linux mount command and the file system's DNS name. To simplify accessing your Amazon EFS file systems, we recommend using the Amazon EFS mount helper utility. Once mounted, you can work with the files and directories in your file system just like you would with a local file system.

EFS uses the Network File System version 4 (NFS v4) protocol. For a step-by-step example of how to access a file system from an EC2 instance, see the guide here.

Q. What Amazon EC2 instance types and AMIs work with Amazon EFS?

Amazon EFS is compatible with all Linux-based AMIs for Amazon EC2. You can mix and match the instance types connected to a single file system. For a step-by-step example of how to access a file system from an EC2 instance, see the instance type guide here.

Q. How do I manage a file system?

Amazon EFS is a fully managed service, so all of the file storage infrastructure is managed for you. When you use Amazon EFS, you avoid the complexity of deploying and maintaining complex file system infrastructure. An Amazon EFS file system grows and shrinks automatically as you add and remove files, so you don't need to manage storage procurement or provisioning.

You can administer a file system through the AWS Management Console, the AWS CLI, or the Amazon EFS API (and various language-specific SDKs). The console, API, and SDK provide the ability to create and delete file systems, configure how file systems are accessed, create and edit file system tags, enable features such as Provisioned Throughput and Lifecycle Management, and display detailed information about file systems.

Q. How do I load data into a file system?

AWS DataSync provides a fast and simple way to securely sync existing file systems with Amazon EFS. DataSync works over any network connection, including with AWS Direct Connect or AWS VPN. AWS Direct Connect provides a high-bandwidth and lower-latency dedicated network connection, over which you can mount your Amazon EFS file systems. You can use AWS DataSync to copy files between two Amazon EFS file systems, including those in different AWS Regions and those belonging to different AWS accounts. You can also use standard Linux copy tools to move data files to Amazon EFS.

Storage classes and lifecycle management

Q. What storage classes does Amazon EFS offer?

Amazon EFS offers you the choice of creating file systems using Standard or One Zone storage classes. Standard storage classes store data with and across multiple AZs. One Zone storage classes store data redundantly within a single AZ, at a 47% lower price compared to file systems using Standard storage classes, for workloads that don't require multi-AZ resilience.

EFS offers four storage classes: two Standard storage classes, Amazon EFS Standard and Amazon EFS Standard-Infrequent Access (EFS Standard-IA), and two One Zone storage classes, Amazon EFS One Zone and Amazon EFS One Zone-Infrequent Access (EFS One Zone-IA).

Q. Is the EFS Infrequent Access storage class still available?

The EFS Infrequent Access (EFS IA) storage class has been renamed to EFS Standard-Infrequent Access (EFS Standard-IA). There are no functional differences in the EFS Standard-IA storage class compared to EFS IA. The name was changed to avoid confusion between the Standard and One Zone versions of the Infrequent Access storage classes.

Q. How do I move files to EFS Standard-IA and EFS One Zone-IA?

Moving files to EFS Standard-IA and EFS One Zone-IA starts by enabling Amazon EFS Lifecycle Management and choosing an age-off policy for your files. Lifecycle Management automatically moves your data from the EFS Standard to the EFS Standard-IA storage class or from the EFS One Zone to the EFS One Zone-IA storage class according to the lifecycle policy you choose. For example, you can automatically move files from EFS Standard to EFS Standard-IA if they aren't accessed after seven days.

Q. What is EFS Intelligent-tiering?

EFS Intelligent-tiering delivers automatic cost savings for workloads with changing access patterns. EFS Intelligent-Tiering uses EFS Lifecycle Management to monitor the access patterns of your workload and is designed to automatically move files that aren't accessed for the duration of the Lifecycle policy (for example, 30 days) from performance-optimized storage classes (EFS Standard or EFS One Zone) to their corresponding cost-optimized Infrequent Access (IA) storage class (EFS Standard-Infrequent Access or EFS One Zone-Infrequent Access). This helps you take advantage of IA storage pricing that is up to 92% lower than EFS Standard or EFS One Zone storage pricing. If access patterns change and that data is accessed again, Lifecycle Management automatically moves the files back to EFS Standard or EFS One Zone, eliminating the risk of unbounded access charges. If the files become infrequently accessed again, Lifecycle Management transitions the files back to the appropriate IA storage class based on your Lifecycle policy.

Q. When should I use Lifecycle Management to move files to the IA storage classes without a policy to move files back to EFS Standard or EFS One Zone, if accessed?

Use EFS Lifecycle Management to automatically move files to EFS Standard-IA or EFS One Zone-IA if your file system contains files that you're certain will be accessed infrequently or not at all. Enable Lifecycle Management by choosing a policy to move files to EFS Standard-IA or EFS One Zone-IA, depending on whether your file system uses EFS Standard or EFS One Zone storage classes. Both EFS Standard-IA and EFS One Zone-IA storage classes are ideal if you need your full dataset readily accessible, and you want to automatically save on storage costs as your files are accessed less frequently. Examples include satisfying audits, performing historical analysis, or backup and recovery.

Q. When should I use EFS Intelligent-Tiering?

Use EFS Intelligent-Tiering to automatically move files between performance-optimized and cost-optimized storage classes when data access patterns are unknown. Enable EFS Lifecycle Management by choosing a policy to automatically move files to EFS Standard-IA or EFS One Zone-IA. Additionally, choose a policy to automatically move files back to EFS Standard or EFS One Zone when they're accessed. With EFS Intelligent-Tiering, you can save on storage costs even if your application access patterns are unknown or access patterns change over time. With these two Lifecycle Management policies set, you pay only for data transition charges between storage classes, and not for repeated data access. Examples of workloads that may have unknown access patterns include web assets and blogs stored by content management systems, logs, machine learning (ML) inference files, and genomic data.

Q. What happens when I disable the policy to move files to the IA storage classes using Amazon EFS Lifecycle Management?

When you disable the policy to move files to the IA storage classes, files will no longer be moved to either EFS Standard-IA or EFS One Zone-IA storage classes (depending on whether your file systems use Standard or One Zone storage classes), and any files that have already moved to an IA storage class will remain there.

Q. What happens when I disable EFS Intelligent-Tiering?

When you disable both Lifecycle Management policies, files will no longer move between the performance-optimized and IA storage classes, and files will remain in the storage classes they resided in when you disabled the lifecycle policies. To disable EFS Intelligent-Tiering, you must disable both the policy that moves files to the EFS Standard-IA or EFS One Zone-IA storage classes, and the policy that moves files to the EFS Standard or EFS One Zone storage class on first access.

Q. What happens if I enable the policy to move files to EFS Standard or EFS One Zone on first access and disable the policy to move files to EFS Standard or EFS One Zone?

Any remaining files in the IA storage classes will move to EFS Standard or EFS One Zone if accessed.

Q. What Amazon EFS features are supported when using EFS Standard-IA and EFS One Zone-IA storage classes?

All Amazon EFS features are supported when using the EFS Standard-IA and EFS One Zone-IA storage classes. Files smaller than 128 KiB are not eligible for Lifecycle Management and will always be stored on either the EFS Standard storage class or the EFS One Zone storage class, depending on whether your file system uses Standard or One Zone storage classes.

Q. What is the latency difference between the performance-optimized storage classes (EFS Standard, EFS One Zone) and the cost-optimized IA storage classes (EFS Standard-IA, EFS One Zone-IA)?

When reading from or writing to the EFS Standard-IA storage class or EFS One Zone-IA storage class, your first-byte latency is higher than that of the EFS Standard or EFS One Zone storage classes. The EFS Standard and EFS One Zone storage classes are designed to provide sub-millisecond read latencies and single-digit millisecond write latencies on average. The EFS Standard-IA and EFS One Zone-IA storage classes are designed to provide double-digit millisecond latencies on average.

Q. What throughput can I drive against files stored in the EFS Standard-IA or EFS One Zone-IA storage class?

The throughput you can drive against an Amazon EFS file system scales linearly with the amount of data stored on the EFS Standard or EFS One Zone storage classes. All Amazon EFS file systems, regardless of size, can burst to 100 MiB/second of throughput. File systems with more than 1 TiB of data stored on EFS Standard or EFS One Zone storage classes can burst to 100 MiB/second per TiB of data stored on EFS Standard or EFS One Zone storage classes. If you require higher amounts of throughput to EFS Standard-IA or EFS One Zone-IA storage classes than your file system allows, use Amazon EFS Provisioned Throughput.

Data protection and availability

Q: How is Amazon EFS designed to provide high durability and availability?

By default, every EFS file system object (such as directory, file, and link) is redundantly stored across multiple AZs for file systems using Standard storage classes. If you select Amazon EFS One Zone storage classes, your data is redundantly stored within a single AZ. Amazon EFS is designed to sustain concurrent device failures by quickly detecting and repairing any lost redundancy. In addition, using Standard storage classes, a file system can be accessed concurrently from all AZs in the Region where it's located, which means that you can architect your application to failover from one AZ to other AZs in the Region to ensure the highest level of application availability. Mount targets are designed to be highly available within an AZ for all EFS storage classes. For more information on availability, see the Amazon EFS Service Level Agreement.

Q: How durable is Amazon EFS?

Amazon EFS is designed to provide 99.999999999% (11 9's) of durability over a given year. In addition, EFS Standard and EFS Standard-IA storage classes are designed to sustain data in the event of an entire Availability Zone loss. Because EFS One Zone storage classes store data in a single AWS Availability Zone, data stored in these storage classes may be lost in the event of a disaster or other fault within the Availability Zone that affects all copies of the data, or in the event of Availability Zone destruction. As with any environment, best practice is to have a backup and to put in place safeguards against accidental deletion. For Amazon EFS data, that best practice includes replicating your file system across Regions using Amazon EFS Replication, and a functioning, regularly tested backup using AWS Backup. File systems using EFS One Zone storage classes are configured to automatically back up files by default at file system creation, unless you choose to disable this functionality.

Q: What failure modes do I have to consider when using Amazon EFS One Zone compared to Standard storage classes?

File systems using Amazon EFS One Zone storage classes are not resilient to a complete AZ outage. In the event of an AZ outage, you will experience a loss of availability, because your file

system data is not replicated to a different AZ. In the event of disaster or fault within an AZ affecting all copies of your data, or a permanent AZ loss, you may experience loss of data that has not been replicated using Amazon EFS Replication to keep an up-to-date copy of your file system in a second AWS Region or an AZ. EFS Replication is designed to meet a recovery point objective (RPO) and recovery time objective (RTO) of minutes. You can also use AWS Backup to store additional copies of your file system data and restore them to a new file system in an AZ or Region of your choice. Amazon EFS file system backup data created and managed by AWS Backup is replicated to 3 AZs and is designed for 99.999999999% (11 9's) durability.

Q. How can I guard my EFS One Zone file system against the loss of an AZ?

You can use Amazon EFS Replication or AWS Backup to guard your EFS One Zone file system against the loss of an AZ. Amazon EFS Replication replicates your file system data to another AWS Region or within the same Region in a few clicks, without requiring additional infrastructure or a custom process to monitor and synchronize data changes. EFS replication is continuous and designed to provide a recovery point objective (RPO) and a recovery time objective (RTO) of minutes for most file systems.

Backups are enabled by default for all file systems using Amazon EFS One Zone storage classes. You can disable this setting when creating file systems. You are able to restore your file data from a recent backup to a newly created file system in any operating AZ in the event of an AZ loss. If Amazon EFS is impacted by an AZ loss, and your data is stored in One Zone storage classes, you may experience data loss for files that have changed since the last automatic backup.

Q: What is Amazon EFS Replication?

EFS Replication allows you to replicate your file system data to another AWS Region or within the same Region in a few clicks, without requiring additional infrastructure or a custom process to monitor and synchronize data changes. Amazon EFS Replication automatically and transparently replicates your data to a second file system in a Region or AZ of your choice. You can use the Amazon EFS console, AWS CLI, and APIs to enable replication on an existing file system. EFS Replication is continuous and designed to provide a recovery point objective (RPO) and a recovery time objective (RTO) of minutes, enabling you to meet your compliance and business continuity goals.

Q: Why should I use EFS Replication?

If you have requirements to maintain a copy of your file system hundreds of miles apart for purposes of disaster recovery, compliance, or business continuity planning, EFS Replication can help you meet those requirements. For applications that require a low network latency cross-region access, Amazon EFS Replication provides a read-only copy in the region of your choice. With Amazon EFS Replication, you can cost-optimize and save up to 75% on your disaster recovery storage costs by using low-cost EFS One Zone storage classes and a 7-day age-off lifecycle management policy for your destination file system. There is no need to build and maintain a custom process for data replication. EFS Replication also makes it easy to monitor and alarm on your RPO status using Amazon CloudWatch.

Q: How do I get started with EFS Replication?

Using the Amazon EFS console, simply enable Replication on the file system you want to replicate (source file system) and choose the Region or AZ where you want to store the replica (destination). You can also use the CreateReplicationConfiguration API from the AWS CLI or AWS SDK to enable EFS Replication. As part of configuring EFS Replication, you'll choose the Region in which to create your replica. If you choose to use EFS One Zone storage classes for your replica, you must also select your file system's AZ. Once EFS Replication is enabled,

Amazon EFS will automatically create a new destination file system in the destination Region or AZ you've selected. You can select the destination file system's lifecycle management policy, backup policies, provisioned throughput, mount targets, and access points independent of the source file system. For example, you can optimize the destination file system storage costs by enabling EFS Lifecycle Management with a shorter age-off policy (such as 7 days) when compared to the source file system's age-off policy (such as 7, 14, 30, 60, or 90 days). EFS Replication configurations such as the replication pair (source and destination), replication status, and last completed replication timestamp can be accessed using the DescribeReplicationConfigurations API.

Q: How does EFS Replication work?

When you enable EFS Replication on a file system, Amazon EFS automatically creates a new file system in the destination region and begins copying your data into it. Once the initial copy is completed, EFS Replication copies changes incrementally to deliver an RPO of minutes for most file systems. EFS Replication preserves all metadata, such as owners and permissions, when copying changes to files and folders. While EFS Replication is enabled, your destination file system is in read-only mode and can be updated only by EFS Replication. In the event that your source file system is unavailable, you can failover to the destination file system by deleting replication. Deleting the Replication makes the destination file system writeable for your applications to use.

Q: Can I change my destination file system's settings when EFS Replication is enabled?

Yes. When EFS Replication is enabled, you can modify your destination file system configuration settings, such as its lifecycle management policy including intelligent-tiering, backup policy, mount targets, access points, and provisioned throughput. All destination file systems are created with encryption of data at rest enabled irrespective of the source file system setting. You cannot change the performance mode of the destination file system. It always matches that of the source file system, except when you create a One Zone replica. In that case, General Purpose performance mode is used because Max I/O performance mode is not supported by EFS One Zone storage classes.

Q: Can I change which Region I'm replicating data to?

No. In order to change the Region of your destination, you first have to delete the replication configuration between your source and destination file system. You can then create a new replication configuration from the source by selecting the desired Region. Amazon EFS will create a new destination file system in the selected Region and begin to replicate the source file system's contents.

Q: Can I delete my source or destination file system if they're part of a replication pair?

You cannot delete either your source or your destination file system if it's part of a replication pair. In order to delete one of the file systems in the pair, you first need to delete the replication configuration.

Q: Is my replica file system point-in-time consistent?

No. EFS Replication doesn't provide point-in-time consistent replication. EFS Replication publishes a timestamp metric on Amazon CloudWatch called TimeSinceLastSync. All changes made to your source file system at least as of the published time will be copied over to the destination. Changes to your source file system after the recorded time may not have been replicated over. You can monitor the health of your EFS Replication using Amazon CloudWatch. If you interrupt the replication process due to a disaster recovery event, some files from the source file system may have transferred over but are not yet copied to their final locations on

your destination file system. These files and their contents can be found on your destination file system in a lost+found directory created by EFS Replication under the root directory.

Q: Can I select the VPC in which my mount targets are created?

Yes. When you enable EFS Replication for the first time, the replica file system will be automatically created for you. It's created in the Region of your choosing without mount targets. You can then create mount targets for your replica file system in the VPC of your choosing. You can also change the VPC for your replica file system by deleting any existing mount targets and creating new ones in a VPC of your choosing.

Q: How can I utilize my destination file system while replication is enabled and when replication is deleted?

When your replication is in Enabled state, only EFS Replication is allowed to make changes to your destination file system. You can access your replica in the read-only mode during this time. In the event of a disaster you can fail over to your destination file system by deleting your replication configuration from the Amazon EFS console or by using the DeleteReplicationConfiguration API. When you delete the Replication, Amazon EFS will stop replicating additional changes and make the destination file system writeable. You can then point your application to your destination file system to continue your operations. You can use the Amazon EFS console or the DescribeReplicationConfigurations API call to check your destination file system status after you've failed over.

Q: Is the data for my file system replica encrypted in transit and at rest?

For all file systems, Amazon EFS automatically and transparently encrypts all Amazon EFS network traffic using Transport Layer Security (TLS) version 1.2. Your destination file system is created with encryption at rest enabled. You can select an encryption key from those available in the destination Region AWS Key Management Service (KMS) or by using the default service "aws/elasticfilesystem" key in the Region where your destination file system is located.

Q: What permissions do I need to use EFS Replication?

To create and delete a replication, your AWS IAM or resource-based policy must have permission for the Amazon EFS API calls CreateFileSystem, CreateReplicationConfiguration, and DescribeReplicationConfigurations.

Q: Does my replication traffic go over the public internet?

No. EFS Replication traffic always stays on the global AWS backbone.

Q: Can I use EFS Replication to replicate my file system to more than one AWS Region or to multiple file systems within a second Region?

No. EFS Replication supports replication between exactly two file systems.

Q: Can I replicate Amazon EFS file systems across AWS accounts?

No. Amazon EFS does not support replicating file systems to a different AWS account.

Q: Does EFS Replication consume my file system burst credits, IOPS limit, and throughput limits?

No. EFS Replication activity does not consume burst credits or count against the file system IOPS and throughput limits for either file system in a replication pair.

Q: Can I expect my destination file system to be available as soon as I enable EFS Replication?

Yes. When you first enable EFS Replication, your replica file system will be created in read-only mode and your entire source file system will be copied to the destination you selected. The time to complete this operation depends on the size of your source file system. Although you can failover to your destination file system at any time, it is recommended that you wait until the copy is complete to minimize data loss. You can monitor the progress of your replication from the Amazon EFS console, which displays a timestamp that indicates the last time your source file system and destination file system were synchronized.

Scale and performance

Q. How much data can I store?

You can store petabytes of data with Amazon EFS. Amazon EFS file systems are elastic and automatically grow and shrink as you add and remove files, without any need to provision file system size up front - and you pay only for what you use.
Q. How many Amazon EC2 instances can connect to a file system?

Amazon EFS supports one to thousands of Amazon Elastic Compute Cloud (Amazon EC2) instances connecting to a file system concurrently.

Q. How many file systems can I create?

You can create up to one thousand file systems per Region. For information on Amazon EFS limits, visit the Amazon EFS Limits page.

Q. What is the difference between General Purpose performance mode and Max I/O performance mode? Which one should I choose?

General Purpose mode is the recommended performance mode for most file systems. It has the lowest latency and provides the best performance for the majority of applications. Max I/O performance mode scales to higher levels of aggregate throughput and operations per second with the tradeoff of higher latencies. File systems using One Zone storage classes always use the General Purpose performance mode. For more information, see the documentation on File System Performance.

Q. What latency can I expect for my Amazon EFS file system?

The expected latency for your Amazon EFS file system depends on the storage class, the performance mode (General Purpose or Max I/O), and the file system operation type (read or write). The table that follows displays the average expected latency for General Purpose file systems.

	Reads	**Writes**
EFS One Zone	As low as 600 microseconds	Low single-digit milliseconds
EFS One Zone-IA	Double-digit milliseconds	Double-digit milliseconds

EFS Standard	As low as 600 microseconds	Low single-digit milliseconds
EFS Standard-IA	Double-digit milliseconds	Double-digit milliseconds

Latency on Max I/O file systems is single-digit to double-digit milliseconds.

Q. How much throughput can a file system support?

Bursting throughput mode, which is the default throughput mode for Amazon EFS file systems, scales the throughput available to a file system as that file system grows. Because file-based workloads are typically spiky - requiring high levels of throughput at certain periods of time and lower levels of throughput otherwise - Amazon EFS is designed to "burst" which allows high throughput levels for periods of time. Also, because many workloads are read-heavy, read operations are metered at a 1:3 ratio to other NFS operations (such as write). All file systems deliver a consistent baseline performance of 50 MB/second per TB of EFS Standard or EFS One Zone storage, while all file systems (regardless of size) can burst to 100 MB/second, and file systems with more than 1 TB of EFS Standard or EFS One Zone storage can burst to 100 MB/second per TB. Since read operations are metered at a 1:3 ratio, you can drive up to 300 MiB/second per TiB of read throughput. As you add data to your file system, the maximum throughput available to the file system scales linearly and automatically with your storage in the Amazon EFS Standard storage class. If you need more throughput than you can achieve with the amount of data you have stored, you can configure Provisioned Throughput to the specific amount your workload requires.

File system throughput is shared across all EC2 instances connected to a file system. For example, a 1 TB file system that can burst to 100 MB/second of throughput can drive 100 MB/second from a single EC2 instance, or 10 EC2 instances can each drive 10 MB/second (100 MB/second collectively). For more information, see the documentation on File System Performance.

Q. What is Provisioned Throughput and when should I use it?

With EFS Provisioned Throughput, you can provision your file systems' throughput independent of the amount of data stored, optimizing your file system throughput performance to match your application's needs.

Provisioned Throughput is available for applications with a high throughput to storage (MB/second per TB) ratio. For example, if you're using Amazon EFS for development tools, web serving, or content management applications - where the amount of data in your file system is low relative to throughput demands - you can instantly get the high levels of throughput that your applications require.

You can select your file system's throughput mode via the AWS Management Console, AWS Command Line Interface (CLI), or Amazon EFS API. For more details, see the documentation on Provisioned Throughput.
Q. How does Amazon EFS Provisioned Throughput work?

When you select Provisioned Throughput for your file system, you can provision the throughput of that file system independently from the amount of data stored, paying for the storage and Provisioned Throughput separately—for example, $0.30 per GB-month for Standard storage and $6.00 per MB/second-month for Provisioned Throughput in US-East (N. Virginia). Read operations are metered at a 1:3 ratio, so you can drive up to 3 MiB/second of read throughput for each 1 MiB/second of throughput provisioned.

Provisioned Throughput also includes 50 KB/second per GB (or 1 MB/second per 20 GB) of throughput in the price of Amazon EFS Standard storage. For example, if you store 20 GB for one month on Amazon EFS Standard storage class and configure a throughput of 5 MB/second for that month, you're billed for 20 GB-month of storage and 4 (5 minus 1) MB/second-month of throughput.

Q: How do I monitor my read and write throughput usage?

You can monitor your throughput using Amazon CloudWatch. The TotalIOBytes, ReadIOBytes, WriteIOBytes, and MetadataIOBytes metrics reflect the actual throughput your applications are driving. PermittedThroughput and MeteredIOBytes reflect your metered throughput limit and usage, respectively, after metering read requests at a 1:3 ratio to other requests. With the Amazon EFS console, you can use the Percent Throughput Limit graph to monitor your throughput utilization. If you use custom CloudWatch dashboards or another monitoring tool, you can also create a CloudWatch metric math expression that compares MeteredIOBytes to PermittedThroughput. If these values are equal, you're consuming your entire amount of throughput and should consider configuring Provisioned Throughput or increasing the amount of throughput configured. For Bursting Throughput mode file systems, monitor the BurstCreditBalance metric and alert on a balance approaching zero to operate your file system at its burst rate rather than its base rate.
Q. How will I be billed in Provisioned Throughput mode?

In Provisioned Throughput mode, you're billed independently for storage you used and throughput you provisioned. You're billed hourly in the following dimensions:

- Storage (per GB-month): You're billed for the amount of storage you use in GB-month.

Throughput (per MB/second-month): You're billed for throughput you provision in MB/second-month.

Q. How often can I change my file system's Provisioned Throughput?

In Provisioned Throughput mode, you can increase its provisioned throughput as often as you want. Decrease your file system throughput in Provisioned Throughput mode or change between Provisioned Throughput and the default Bursting Throughput modes as long as it's been more than 24 hours since the last decrease or throughput mode change.
Q. What is the throughput of my file system if the Provisioned Throughput mode is set to less than the Baseline Throughput I am entitled to in Bursting Throughput mode?

In the default Bursting Throughput mode, the throughput of your file system scales with the amount of data stored. If your file system in the Provisioned Throughput mode grows in size after the initial configuration, your file system could potentially have a higher baseline rate in Bursting Throughput mode than in the Provisioned Throughput mode.

In that case, your file system throughput will be the throughput it's entitled to in the default Bursting Throughput mode, and you won't incur any additional charge for the throughput beyond the bursting storage cost. You can also burst according to the Amazon EFS throughput bursting model.

Security

Access Control

Q. How do I control which Amazon EC2 instances can access my file system?

You control which EC2 instances can access your file system using VPC security group rules and AWS IAM policies. Use VPC security groups to control the network traffic to and from your file system. Attach an IAM policy to your file system to control which clients can mount your file system and with what permissions, and use EFS Access Points to manage application access. Control access to files and directories with POSIX-compliant user and group-level permissions.

Q. How can I use IAM policies to manage file system access?

Using the Amazon EFS console, you can apply common policies to your file system, such as disabling root access, enforcing read-only access, or enforcing that all connections to your file system are encrypted. You can also apply more advanced policies, such as granting access to specific IAM roles, including those in other AWS accounts.

Access Points

Q. What is an Amazon EFS Access Point?

Amazon EFS Access Points simplify providing applications with access to shared datasets in an Amazon EFS file system. Amazon EFS Access Points work together with AWS IAM and enforce an operating system user and group, and a directory for every file system request made through the access point. You can create multiple access points per file system and use them to provide access to specific applications.

Q. Why should I use Amazon EFS Access Points?

Amazon EFS Access Points represent a flexible way to manage application access in NFS environments with increased scalability, security, and ease of use. Use cases that can benefit from Amazon EFS Access Points include container-based environments where developers build and deploy their own containers, data science applications that require access to production data, and sharing a specific directory in your file system with other AWS accounts.

Q. How do Amazon EFS Access Points work?

When you create an Amazon EFS Access Point, you can configure an operating system user and group, and a root directory for all connections that use it. If you specify the root directory's owner, EFS will automatically create it with the permissions you provide the first time a client connects to the access point. You can also update your file system's IAM policy to apply to your access points. For example, you can apply a policy that requires a specific IAM identity in order to connect to a given access point. For more information, see the Amazon EFS user guide.

Encryption

Q. What is Amazon EFS Encryption?

Amazon EFS offers the ability to encrypt data at rest and in transit.

Data encrypted at rest is transparently encrypted while being written, and transparently decrypted while being read, so you don't have to modify your applications. Encryption keys are managed by the AWS KMS, eliminating the need to build and maintain a secure key management infrastructure.

Data encryption in transit uses industry-standard Transport Layer Security (TLS) 1.2 to encrypt data sent between your clients and EFS file systems.

Encryption of data at rest and data in transit can be configured together or separately to help meet your unique security requirements.

Q. What is the AWS Key Management Service (KMS)?

AWS KMS manages the encryption keys for encrypted data at rest on EFS file systems. KMS is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data. KMS is integrated with AWS services, including Amazon EFS, Amazon EBS, and Amazon S3, making it simple to encrypt your data with encryption keys that you manage. AWS KMS is also integrated with AWS CloudTrail to provide you with logs of all key usage to help meet your regulatory and compliance needs.
Q. How do I enable encryption for my Amazon EFS file system?

You can enable encryption at rest in the Amazon EFS console or by using the AWS CLI or SDKs. When creating a new file system in the EFS console, click “Create File System” and click the checkbox to enable encryption.

Data can be encrypted in transit between your Amazon EFS file system and its clients by using the Amazon EFS mount helper.

Encryption of data at rest and data in transit can be configured together or separately to help meet your unique security requirements.

Q. Does encryption impact Amazon EFS performance?

Encrypting your data has a minimal effect on I/O latency and throughput.

On-premises access

Q. How do I access an Amazon EFS file system from servers in my on-premises datacenter?

To access Amazon EFS file systems from on premises, you must have an AWS Direct Connect or AWS VPN connection between your on-premises datacenter and your Amazon VPC.

You mount an Amazon EFS file system on your on-premises Linux server using the standard Linux mount command for mounting a file system using the NFS v4.1 protocol.

For more information about accessing Amazon EFS file systems from on-premises servers, see the documentation.
Q. What can I do by enabling access to my Amazon EFS file systems from my on-premises servers?

You can mount your Amazon EFS file systems on your on-premises servers, and move file data to and from Amazon EFS using standard Linux tools and scripts or AWS DataSync. The ability to move file data to and from Amazon EFS file systems allows for three use cases.

First, you can migrate data from on-premises datacenters to permanently reside in EFS file systems.

Second, you can support cloud bursting workloads to off-load your application processing to the cloud. You can move data from your on-premises servers into your Amazon EFS file systems, analyze it on a cluster of Amazon EC2 instances in your Amazon VPC, and store the results permanently in your Amazon EFS file systems or move the results back to your on-premises servers.

Third, you can periodically copy your on-premises file data to Amazon EFS to support backup and disaster recovery scenarios.

Q. Can I access my Amazon EFS file system concurrently from my on-premises datacenter servers as well as Amazon EC2 instances?

Yes. You can access your Amazon EFS file system concurrently from servers in your on-premises datacenter as well as Amazon EC2 instances in your Amazon VPC. Amazon EFS provides the same file system access semantics, such as strong data consistency and file locking, across all EC2 instances and on-premises servers accessing a file system.
Q. What is the recommended best practice when moving file data to and from on-premises servers?

Because of the propagation delay tied to data traveling over long distances, the network latency of the network connection between your on-premises datacenter and your Amazon VPC can be tens of milliseconds. If your file operations are serialized, the latency of the network connection directly impacts your read and write throughput; in essence, the volume of data you can read or write during a period of time is bounded by the amount of time it takes for each read and write operation to complete. To maximize your throughput, parallelize your file operations so that multiple reads and writes are processed by Amazon EFS concurrently. Standard tools like GNU parallel allow you to parallelize the copying of file data. For more information, see the online documentation.

Q. How do I copy existing data from on-premises file storage to Amazon EFS?

There are a number of methods to copy existing on-premises data into Amazon EFS. AWS DataSync provides a fast and simple way to securely sync existing file systems into EFS and works over any network, including AWS Direct Connect.

AWS Direct Connect provides a high-bandwidth and lower-latency dedicated network connection over which you can mount your EFS file systems. Once mounted, you can use DataSync to copy data into EFS up to 10 times faster than standard Linux copy tools.

Data transfer

Q. What AWS-native options do I have to transfer data into my file system?

AWS DataSync is an online data transfer service that makes it faster and simpler to move data between on-premises storage and Amazon EFS. DataSync uses a purpose-built protocol to accelerate and secure transfer over the internet or AWS Direct Connect, at speeds up to 10 times faster than open-source tools. Using DataSync, you can perform one-time data migrations, transfer on-premises data for timely in-cloud analysis, and automate replication to AWS for data protection and recovery.

AWS Transfer Family is a fully managed file transfer service that provides support for Secure File Transfer Protocol (SFTP), File Transfer Protocol over SSL (FTPS), and File Transfer Protocol (FTP). The AWS Transfer Family provides you with a fully managed, highly available file transfer service with auto scaling capabilities, eliminating the need for you to manage file transfer–related infrastructure. Your end users' workflows remain unchanged, while data uploaded and downloaded over the chosen protocols is stored in your Amazon EFS file system.
Q. How do I transfer data into or out of my Amazon EFS file system?

To get started with AWS DataSync, you can use the AWS Management Console or AWS CLI to connect the agent to your on-premises or in-cloud file systems using the Network File System (NFS) protocol, select your Amazon EFS file system, and start copying data. You must first deploy a software agent that is available for download from the console, except when copying files between two Amazon EFS file systems.

To get started with AWS Transfer Family, first ensure that your file system's directories are accessible by the POSIX users that you plan to assign to AWS Transfer. Then you can use the console, CLI, or API to create a Transfer Family endpoint and user(s). Once complete, your end users can use their SFTP, FTP, or FTPS clients to access data stored in your Amazon EFS file system.
Q. Can Amazon EFS data be transferred between Regions?

You can use AWS DataSync to transfer files between two Amazon EFS file systems, including ones in different AWS Regions. AWS Transfer Family endpoints must be in the same Region as your Amazon EFS file system.
Q. Can I access my file system with another AWS account?

Yes. You can use AWS DataSync to copy files to an Amazon EFS file system in another AWS account.

You can also configure your Amazon EFS file system to be accessed by AWS Transfer Family using another account as long as the account has been granted permissions to do so. To learn more about granting Transfer Family permissions to external AWS accounts via file system policies, see the documentation.

Compatibility

Q. What interoperability and compatibility is there between existing AWS services and Amazon EFS?

Amazon EFS is integrated with a number of other AWS services, including Amazon CloudWatch, AWS CloudFormation, AWS CloudTrail, AWS IAM, and AWS Tagging services.

CloudWatch allows you to monitor file system activity using metrics. CloudFormation allows you to create and manage file systems using templates.

CloudTrail allows you to record all Amazon EFS API calls in log files.

IAM allows you to control who can administer your file system. AWS Tagging services allows you to label your file systems with metadata that you define.

You can plan and manage your Amazon EFS file system costs by using AWS Budgets. You can work with AWS Budgets from the AWS Billing and Cost Management console. To use AWS Budgets, you create a monthly cost budget for your Amazon EFS file systems.
Q. What type of locking does Amazon EFS support?

Locking in Amazon EFS follows the NFS v4.1 protocol for advisory locking and allows your applications to use both whole file and byte range locks.
Q. Are file system names global (like Amazon S3 bucket names)?

Every file system has an automatically generated ID number that is globally unique. You can tag your file system with a name, and these names don't need to be unique.

Access from AWS services

Q. Can I access Amazon EFS from Amazon EC2 instances?

Yes. You can access Amazon EFS from tens of thousands of EC2 instances launched using any supported launch method. When you use the EC2 launch instance wizard, you can create new Amazon EFS file systems, and add existing file systems to the instance you're launching. All the file systems you added to the instance are automatically mounted on the instance when it's

launched. Learn more about using Amazon EFS with EC2 instances in the Amazon EC2 documentation.

Q. Can I access Amazon EFS from Amazon ECS containers?

Yes. You can access Amazon EFS from containerized applications launched by Amazon ECS using both Amazon EC2 and AWS Fargate launch types by referencing an EFS file system in your task definition. Find instructions for getting started in the ECS documentation.

Q. Can I access Amazon EFS from Amazon Elastic Kubernetes Service (EKS) pods?

Yes. You can access Amazon EFS from containerized applications launched by Amazon EKS, with either Amazon EC2 or AWS Fargate launch types, using the Amazon EFS CSI driver. Find instructions for getting started in the EKS documentation.

Q. Can I access Amazon EFS from AWS Lambda functions?

Yes. You can access Amazon EFS from functions running in AWS Lambda by referencing an Amazon EFS file system in your function settings. Find instructions for getting started in the AWS Lambda documentation.

Q. Can I access Amazon EFS from Amazon SageMaker?

Yes. You can access training data in Amazon EFS from Amazon SageMaker training jobs by referencing an Amazon EFS file system in your CreateTrainingJob request. EFS is also automatically used for home directories created by Amazon SageMaker Studio.

Amazon DynamoDB

Fast, flexible NoSQL database service for single-digit millisecond performance at any scale.

How it works

Amazon DynamoDB is a fully managed, serverless, key-value NoSQL database designed to run high-performance applications at any scale. DynamoDB offers built-in security, continuous backups, automated multi-Region replication, in-memory caching, and data import and export tools.

Use cases

Develop software applications
Build internet-scale applications supporting user-content metadata and caches that require high concurrency and connections for millions of users and millions of requests per second.

Create media metadata stores

Scale throughput and concurrency for media and entertainment workloads such as real-time video streaming and interactive content, and deliver lower latency with multi-Region replication across AWS Regions.

Deliver seamless retail experiences
Use design patterns for deploying shopping carts, workflow engines, inventory tracking, and customer profiles. DynamoDB supports high-traffic, extreme-scaled events and can handle millions of queries per second.

Scale gaming platforms
Focus on driving innovation with no operational overhead. Build out your game platform with player data, session history, and leaderboards for millions of concurrent users.

What is Amazon DynamoDB?

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. DynamoDB lets you offload the administrative burdens of operating and scaling a distributed database so that you don't have to worry about hardware provisioning, setup and configuration, replication, software patching, or cluster scaling. DynamoDB also offers encryption at rest, which eliminates the operational burden and complexity involved in protecting sensitive data. For more information, see DynamoDB encryption at rest.

With DynamoDB, you can create database tables that can store and retrieve any amount of data and serve any level of request traffic. You can scale up or scale down your tables' throughput capacity without downtime or performance degradation. You can use the AWS Management Console to monitor resource utilization and performance metrics.

DynamoDB provides on-demand backup capability. It allows you to create full backups of your tables for long-term retention and archival for regulatory compliance needs. For more information, see Using On-Demand backup and restore for DynamoDB.

You can create on-demand backups and enable point-in-time recovery for your Amazon DynamoDB tables. Point-in-time recovery helps protect your tables from accidental write or delete operations. With point-in-time recovery, you can restore a table to any point in time during the last 35 days. For more information, see Point-in-time recovery: How it works.

DynamoDB allows you to delete expired items from tables automatically to help you reduce storage usage and the cost of storing data that is no longer relevant. For more information, see Expiring items by using DynamoDB Time to Live (TTL).

High availability and durability

DynamoDB automatically spreads the data and traffic for your tables over a sufficient number of servers to handle your throughput and storage requirements, while maintaining consistent and fast performance. All of your data is stored on solid-state disks (SSDs) and is automatically replicated across multiple Availability Zones in an AWS Region, providing built-in high availability and data durability. You can use global tables to keep DynamoDB tables in sync across AWS Regions. For more information, see Global tables - multi-Region replication for DynamoDB.

Amazon DynamoDB FAQs

About Amazon DynamoDB

What is Amazon DynamoDB?

DynamoDB is a fast and flexible nonrelational database service for any scale. DynamoDB enables customers to offload the administrative burdens of operating and scaling distributed databases to AWS so that they don't have to worry about hardware provisioning, setup and configuration, throughput capacity planning, replication, software patching, or cluster scaling.

What does DynamoDB manage on my behalf?

DynamoDB takes away one of the main stumbling blocks of scaling databases: the management of database software and the provisioning of the hardware needed to run it. You can deploy a nonrelational database in a matter of minutes. DynamoDB automatically scales throughput capacity to meet workload demands, and partitions and repartitions your data as your table size grows. Also, DynamoDB synchronously replicates data across three facilities in an AWS Region, giving you high availability and data durability.

What is the consistency model of DynamoDB?

When reading data from DynamoDB, users can specify whether they want the read to be eventually consistent or strongly consistent:

- Eventually consistent reads (the default) – The eventual consistency option maximizes your read throughput. However, an eventually consistent read might not reflect the results of a recently completed write. All copies of data usually reach consistency within a second. Repeating a read after a short time should return the updated data.

- Strongly consistent reads — In addition to eventual consistency, DynamoDB also gives you the flexibility and control to request a strongly consistent read if your application, or an element of your application, requires it. A strongly consistent read returns a result that reflects all writes that received a successful response before the read.

- ACID transactions – DynamoDB transactions provide developers atomicity, consistency, isolation, and durability (ACID) across one or more tables within a single AWS account and region. You can use transactions when building applications that require coordinated inserts, deletes, or updates to multiple items as part of a single logical business operation.

Getting started

What kind of query functionality does DynamoDB support?

DynamoDB supports GET/PUT operations by using a user-defined primary key. The primary key is the only required attribute for items in a table. You specify the primary key when you create a table, and it uniquely identifies each item. DynamoDB also provides flexible querying by letting you query on nonprimary key attributes using global secondary indexes and local secondary indexes.

A primary key can be either a single-attribute partition key or a composite partition-sort key. A single-attribute partition key could be, for example, UserID. Such a single attribute partition key would allow you to quickly read and write data for an item associated with a given user ID.

DynamoDB indexes a composite partition-sort key as a partition key element and a sort key element. This multipart key maintains a hierarchy between the first and second element values. For example, a composite partition-sort key could be a combination of UserID (partition) and Timestamp (sort). Holding the partition key element constant, you can search across the sort key element to retrieve items. Such searching would allow you to use the Query API to, for example, retrieve all items for a single UserID across a range of time stamps.

How do I update and query data items with DynamoDB?

After you have created a table using the DynamoDB console or CreateTable API, you can use the PutItem or BatchWriteItem APIs to insert items. Then, you can use the GetItem, BatchGetItem, or, if composite primary keys are enabled and in use in your table, the Query API to retrieve the items you added to the table.

Can DynamoDB be used by applications running on any operating system?

Yes. DynamoDB is a fully managed cloud service that you access via API. Applications running on any operating system (such as Linux, Windows, iOS, Android, Solaris, AIX, and HP-UX) can use DynamoDB. We recommend using the AWS SDKs to get started with DynamoDB.

Planning

How am I charged for my use of DynamoDB?

Each DynamoDB table has provisioned read-throughput and write-throughput associated with it. You are billed by the hour for that throughput capacity. Note that you are charged by the hour for the throughput capacity, whether or not you are sending requests to your table. If you would like to change your table's provisioned throughput capacity, you can do so using the AWS Management Console, the UpdateTable API, or the PutScalingPolicy API for auto scaling. Also, DynamoDB charges for data storage as well as the standard internet data transfer fees.

What is the maximum throughput I can provision for a single DynamoDB table?

Maximum throughput per DynamoDB table is practically unlimited. For information about the limits in place, see Limits in DynamoDB. If you want to request a limit increase, contact Amazon.

What is the minimum throughput I can provision for a single DynamoDB table?
The smallest provisioned throughput you can request is 1 write capacity unit and 1 read capacity unit for both auto scaling and manual throughput provisioning. Such provisioning falls

within the free tier which allows for 25 units of write capacity and 25 units of read capacity. The free tier applies at the account level, not the table level. In other words, if you add up the provisioned capacity of all your tables, and if the total capacity is no more than 25 units of write capacity and 25 units of read capacity, your provisioned capacity would fall into the free tier.

What are DynamoDB table classes?

DynamoDB offers two table classes designed to help you optimize for cost. The DynamoDB Standard table class is the default, and recommended for the vast majority of workloads. The DynamoDB Standard-Infrequent Access (DynamoDB Standard-IA) table class is optimized for tables that store data that is accessed infrequently, where storage is the dominant cost. Each table is associated with a table class and each table class offers a different pricing for data storage as well as read and write requests. You can select the most cost-effective table class based on your table's storage requirements and data access patterns.

When should I use DynamoDB Standard-IA?

DynamoDB Standard-IA helps you reduce your DynamoDB total costs for tables that store infrequently accessed data such as applications' logs, old social media posts, e-commerce order history, and past gaming achievements. If storage is your dominant table cost— storage cost exceeds 50 percent of the cost of throughput (reads and writes) consistently—then the DynamoDB Standard-IA table class is the most economical choice for you. Learn more about DynamoDB Standard-IA pricing in the DynamoDB pricing page.

How do DynamoDB Standard-IA tables work with existing DynamoDB features and integrate with other AWS services?

DynamoDB Standard-IA tables are no different than DynamoDB Standard tables in supporting all existing DynamoDB features including global tables, secondary indexes, on-demand backups, point-in-time recovery (PITR), and Amazon DynamoDB Accelerator (DAX). DynamoDB Standard-IA tables also have built-in integration with other AWS services in the same way as DynamoDB Standard tables. For example, you can monitor the performance of your DynamoDB Standard-IA tables using Amazon CloudWatch, use AWS CloudFormation templates to provision and manage your DynamoDB Standard-IA tables, stream your change data records to Amazon Kinesis Data Streams, and export your DynamoDB Standard-IA tables data to Amazon Simple Storage Service (Amazon S3).

How it works

Data models and APIs
For more information about data models and APIs, see Amazon DynamoDB: How It Works.

Scalability, availability, and durability
For information about scalability, availability, and durability, see Amazon DynamoDB Product Details.

Auto scaling
For information about DynamoDB auto scaling, see Managing Throughput Capacity Automatically with DynamoDB Auto Scaling.

Security and control
For information about DynamoDB security and control, see Identity and Access Management in Amazon DynamoDB.

Amazon DocumentDB (with MongoDB compatibility)

Scale JSON workloads with ease using a fully managed document database service.

How it works

Amazon DocumentDB is a scalable, highly durable, and fully managed database service for operating mission-critical MongoDB workloads.

Migrate self-managed MongoDB workloads

Set up, secure, and scale MongoDB-compatible databases without worrying about running cluster management software, configuring backups, or monitoring production workloads.

JSON Management

As a type of NoSQL database, Amazon DocumentDB makes it easy to insert, query, index, and perform aggregations over JSON data.

Use cases

Store and query CMS data
Improve the customer experience through fast, reliable access to reviews, images, and other content stored in your content management system (CMS).

Manage user profiles, preferences, and requests
Generate customer recommendations and enable online transactions. Manage millions of user profiles and preferences.

Scale mobile and web applications
Build applications that scale to process millions of user requests per second with millisecond latency.

What Is Amazon DocumentDB (with MongoDB Compatibility)

Amazon DocumentDB (with MongoDB compatibility) is a fast, reliable, and fully managed database service. Amazon DocumentDB makes it easy to set up, operate, and scale MongoDB-compatible databases in the cloud. With Amazon DocumentDB, you can run the same application code and use the same drivers and tools that you use with MongoDB.

Before using Amazon DocumentDB, you should review the concepts and features described in How It Works. After that, complete the steps in Get Started Guide.

Overview of Amazon DocumentDB

The following are some high-level features of Amazon DocumentDB:

- Amazon DocumentDB automatically grows the size of your storage volume as your database storage needs grow. Your storage volume grows in increments of 10 GB, up to a maximum of 64 TiB. You don't need to provision any excess storage for your cluster to handle future growth.
- With Amazon DocumentDB, you can increase read throughput to support high-volume application requests by creating up to 15 replica instances. Amazon DocumentDB replicas share the same underlying storage, lowering costs and avoiding the need to perform writes at the replica nodes. This capability frees up more processing power to serve read requests and reduces the replica lag time—often down to single digit milliseconds. You can add replicas in minutes regardless of the storage volume size. Amazon DocumentDB also provides a reader endpoint, so the application can connect without having to track replicas as they are added and removed.
- Amazon DocumentDB lets you scale the compute and memory resources for each of your instances up or down. Compute scaling operations typically complete in a few minutes.
- Amazon DocumentDB runs in Amazon Virtual Private Cloud (Amazon VPC), so you can isolate your database in your own virtual network. You can also configure firewall settings to control network access to your cluster.
- Amazon DocumentDB continuously monitors the health of your cluster. On an instance failure, Amazon DocumentDB automatically restarts the instance and associated processes. Amazon DocumentDB doesn't require a crash recovery replay of database redo logs, which greatly reduces restart times. Amazon DocumentDB also isolates the database cache from the database process, enabling the cache to survive an instance restart.
- On instance failure, Amazon DocumentDB automates failover to one of up to 15 Amazon DocumentDB replicas that you create in other Availability Zones. If no replicas have been provisioned and a failure occurs, Amazon DocumentDB tries to create a new Amazon DocumentDB instance automatically.
- The backup capability in Amazon DocumentDB enables point-in-time recovery for your cluster. This feature allows you to restore your cluster to any second during your retention period, up to the last 5 minutes. You can configure your automatic backup retention period up to 35 days. Automated backups are stored in Amazon Simple Storage Service (Amazon S3), which is designed for 99.999999999% durability. Amazon DocumentDB backups are

automatic, incremental, and continuous, and they have no impact on your cluster performance.

- With Amazon DocumentDB, you can encrypt your databases using keys that you create and control through AWS Key Management Service (AWS KMS). On a database cluster running with Amazon DocumentDB encryption, data stored at rest in the underlying storage is encrypted. The automated backups, snapshots, and replicas in the same cluster are also encrypted.

If you are new to AWS services, use the following resources to learn more:

- AWS offers services for computing, databases, storage, analytics, and other functionality. For an overview of all AWS services, see Cloud Computing with Amazon Web Services.
- AWS provides a number of database services. For guidance on which service is best for your environment, see Databases on AWS.

Clusters

A *cluster* consists of 0 to 16 instances and a cluster storage volume that manages the data for those instances. All writes are done through the primary instance. All instances (primary and replicas) support reads. The cluster's data is stored in the cluster volume with copies in three different Availability Zones.

Instances

An Amazon DocumentDB instance is an isolated database environment in the cloud. An instance can contain multiple user-created databases. You can create and modify an instance using the AWS Management Console or the AWS CLI.

The computation and memory capacity of an instance are determined by its *instance class*. You can select the instance that best meets your needs. If your needs change over time, you can choose a different instance class. For instance class specifications, see Instance Class Specifications.

Amazon DocumentDB instances run only in the Amazon VPC environment. Amazon VPC gives you control of your virtual networking environment: You can choose your own IP address range, create subnets, and configure routing and access control lists (ACLs).

Before you can create Amazon DocumentDB instances, you must create a cluster to contain the instances.

Not all instance classes are supported in every region. The following table shows which instance classes are supported in each region.

Monitoring

There are several ways that you can track the performance and health of an instance. You can use the free Amazon CloudWatch service to monitor the performance and health of an instance. You can find performance charts on the Amazon DocumentDB console. You can subscribe to Amazon DocumentDB events to be notified when changes occur with an instance, snapshot, parameter group, or security group.

For more information, see the following:

- Monitoring Amazon DocumentDB with CloudWatch
- Logging Amazon DocumentDB API Calls with AWS CloudTrail

Amazon DocumentDB (with MongoDB compatibility) FAQs

General

What is Amazon DocumentDB (with MongoDB compatibility)?

Amazon DocumentDB (with MongoDB compatibility) is a fast, scalable, highly available, and fully managed enterprise document database service that supports MongoDB workloads. As a document database, Amazon DocumentDB makes it easy to store, query, and index JSON data. Developers can use the same MongoDB application code, drivers, and tools as they do today to run, manage, and scale workloads on Amazon DocumentDB. Enjoy improved performance, scalability, and availability without worrying about managing the underlying infrastructure.

Customers can use AWS Database Migration Service (DMS) to easily migrate their on-premises or Amazon Elastic Compute Cloud (EC2) MongoDB non-relational databases to Amazon DocumentDB with virtually no downtime. There are no upfront investments required to use Amazon DocumentDB, and customers only pay for the capacity they use.

What use cases are well-suited for a document database like Amazon DocumentDB?

Document-oriented databases are one of the fastest growing categories of noSQL databases, with the primary reason being that document databases offer both flexible schemas and extensive query capabilities. The document model is a great choice for use cases with dynamic datasets that require ad-hoc querying, indexing, and aggregations. With the scale that Amazon DocumentDB provides, it is used by a wide variety of customers for use cases such as content management, personalization, catalogs, mobile and web applications, IoT, and profile management.

What does "MongoDB-compatible" mean?

"MongoDB compatible" means that Amazon DocumentDB interacts with the Apache 2.0 open source MongoDB 3.6 and 4.0 APIs. As a result, you can use the same MongoDB drivers, applications, and tools with Amazon DocumentDB with little or no changes. While Amazon DocumentDB supports a vast majority of the MongoDB APIs that customers actually use, it does not support every MongoDB API. Our focus has been to deliver the capabilities that customer actually use and need.

Since launch, we have continued to work backwards from customers and have delivered an additional 80+ capabilities, including MongoDB 4.0 compatibility and transactions. To learn more about the supported MongoDB APIs, see the compatibility documentation. To learn more about recent Amazon DocumentDB launches, see "Amazon DocumentDB Announcements" on the Amazon DocumentDB resources page.

Is **Amazon DocumentDB restricted by the MongoDB SSPL license?**

No. Amazon DocumentDB does not utilize any MongoDB SSPL code and thus is not restricted by this license. Instead, Amazon DocumentDB interacts with the Apache 2.0 open-source MongoDB 3.6 and 4.0 APIs. We will continue to listen and work backward from our customers to deliver the capabilities that they need. To learn more about the supported MongoDB APIs, see the compatibility documentation. To learn more about recent Amazon DocumentDB launches, see "Amazon DocumentDB Announcements" on the Amazon DocumentDB resources page.

How can I migrate data from an existing MongoDB database to Amazon DocumentDB?

Customers can use AWS Database Migration Service (DMS) to easily migrate their on-premises or Amazon Elastic Compute Cloud (EC2) MongoDB databases to Amazon DocumentDB with virtually no downtime. With DMS, you can migrate from a MongoDB replica set or from a

sharded cluster to Amazon DocumentDB. Additionally, you can use most existing tools to migrate data from a MongoDB database to Amazon DocumentDB, including mongodump/mongorestore, mongoexport/mongoimport, and third-party tools that support Change Data Capture (CDC) via the oplog. For more information, see Migrating to Amazon DocumentDB.

Do I need to change client drivers to use Amazon DocumentDB?

No, Amazon DocumentDB works with a vast majority of MongoDB drivers compatible with MongoDB 3.4+.

Does Amazon DocumentDB support ACID transactions?

Yes. With the launch of support for MongoDB 4.0 compatibility, Amazon DocumentDB supports the ability to perform atomicity, consistency, isolation, durability (ACID) transactions across multiple documents, statements, collections, and databases.

Is Amazon DocumentDB subject to MongoDB's end of life (EOL) schedule?

No, Amazon DocumentDB does not follow the same support lifecycles as MongoDB and MongoDB's EOL schedule does not apply to Amazon DocumentDB.

How do I access my Amazon DocumentDB cluster?

Amazon DocumentDB clusters are deployed within a customer's Amazon VPC (VPC) and can be accessed directly by Amazon Elastic Compute Cloud (EC2) instances or other AWS services that are deployed in the same VPC. Additionally, Amazon DocumentDB can be accessed by Amazon EC2 instances or other AWS services in different VPCs in the same region or other regions via VPC peering. Access to Amazon DocumentDB clusters must be done through the mongo shell or with MongoDB drivers. Amazon DocumentDB requires that you authenticate when connecting to a cluster. For additional options, see Connecting to an Amazon DocumentDB Cluster from Outside an Amazon VPC.

Why are Amazon RDS permissions and resources required to use Amazon DocumentDB?

For certain management features such as instance lifecycle management, encryption-at-rest with Amazon Key Management Service (KMS) keys and security groups management, Amazon DocumentDB leverages operational technology that is shared with Amazon Relational Database Service (RDS) and Amazon Neptune. When using the describe-db-instances and describe-db-clusters AWS CLI APIs, we recommend filtering for Amazon DocumentDB resources using the following parameter: "--filter Name=engine,Values=docdb".

What instances types does Amazon DocumentDB offer?

Please see the Amazon DocumentDB pricing page for current information on available instance types per region.

How do I try Amazon DocumentDB?

To try Amazon DocumentDB, please see the Getting Started guide.

Does Amazon DocumentDB have an SLA?

Yes. For more information, please see Amazon DocumentDB (with MongoDB compatibility) Service Level Agreement.

Performance

What type of performance can I expect from Amazon DocumentDB?

When writing to storage, Amazon DocumentDB only persists a write-ahead logs, and does not need to write full buffer page syncs. As a result of this optimization, which does not compromise durability, Amazon DocumentDB writes are typically faster than traditional databases. Amazon DocumentDB clusters can scale out to millions of reads per second with up to 15-read replicas.

Does Amazon DocumentDB have a free tier and can you get started for free?

Yes, you can try Amazon DocumentDB for free using the 1-month free trial. If you have not used Amazon DocumentDB before, you are eligible for a one month free trial. Your organization gets 750 hours per month of t3.medium instance usage, 30 million IOs, 5 GB of storage, and 5 GB of backup storage for free for 30 days. Once your one month free trial expires or your usage exceeds the free allowance, you can shut down your cluster to avoid any charges, or keep it running at our standard on-demand rates. To learn more, refer to the DocumentDB free trial page.

Hardware, Scaling, and Storage

What are the minimum and maximum storage limits of an Amazon DocumentDB cluster?

The minimum storage is 10 GB. Based on your cluster usage, your Amazon DocumentDB storage will automatically grow, up to 64 TB in 10 GB increments with no impact on performance. There is no need to provision storage in advance.

How does Amazon DocumentDB scale?

Amazon DocumentDB scales in two dimensions: storage and compute. Amazon DocumentDB's storage automatically scales from 10 GB to 64 TB in increments of 10 GB. Amazon DocumentDB's compute capacity can be scaled up by creating larger instances and horizontally (for greater read throughput) by adding additional replica instances (up to 15) to the cluster.

How do I scale the compute resources associated with my Amazon DocumentDB cluster?

You can scale the compute resources allocated to your instance in the AWS Management Console by selecting the desired instance and clicking the “modify” button. Memory and CPU resources are modified by changing your instance class.

When you modify your instance class, your requested changes will be applied during your specified maintenance window. Alternatively, you can use the "Apply Immediately" flag to apply your scaling requests immediately. Both of these options will have an availability impact for a few minutes as the scaling operation is performed. Bear in mind that any other pending system changes will also be applied.

Does Amazon DocumentDB support MongoDB sharding?

No. Amazon DocumentDB's distributed storage architecture is a different approach to scaling than MongoDB sharding.

Backup and Restore

How do I enable backups for my cluster?

Automated backups are always enabled on Amazon DocumentDB clusters. Amazon DocumentDB's simple database backup capability enables point-in-time recovery for your clusters. You can increase your backup window for point-in-time restores up to 35 days. Backups do not impact database performance.

Can I take cluster snapshots and keep them around as long as I want?

Yes. Manual snapshots can be retained beyond the backup window and there is no performance impact when taking snapshots. Note that restoring data from cluster snapshots requires creating a new cluster.

If my instance fails, what is my recovery path?

Amazon DocumentDB automatically maintains six copies of your data across three Availability Zones (AZs) and will automatically attempt to recover your instance in a healthy AZ with no data loss. In the unlikely event your data is unavailable within Amazon DocumentDB storage, you can restore from a cluster snapshot or perform a point-in-time restore operation to a new cluster. Note that the latest restorable time for a point-in-time restore operation can be up to five minutes in the past.

What happens to my automated backups and cluster snapshots if I delete my cluster?

You can choose to create a final snapshot when deleting your instance. If you do, you can use this snapshot to restore the deleted instance at a later date. Amazon DocumentDB retains this final user-created snapshot along with all other manually created snapshots after the instance is deleted. Only snapshots are retained after the instance is deleted (i.e., automated backups created for point-in-time restore are not kept).

What happens to my automated backups and cluster snapshots if I delete my account?

Deleting your AWS account will delete all automated backups and snapshot backups contained in the account.

Can I share my snapshots with another AWS account?

Yes. Amazon DocumentDB gives you the ability to create snapshots of your cluster, which you can use later to restore a cluster. You can share a snapshot with a different AWS account, and the owner of the recipient account can use your snapshot to restore a cluster that contains your data. You can even choose to make your snapshots public – that is, anybody can restore a cluster containing your (public) data. You can use this feature to share data between your various environments (production, dev/test, staging, etc.) that have different AWS accounts, as well as keep backups of all your data secure in a separate account in case your main AWS account is ever compromised.

Will I be billed for shared snapshots?

There is no charge for sharing snapshots between accounts. However, you may be charged for the snapshots themselves, as well as any clusters that you restore from shared snapshots.

Can I automatically share snapshots?

We do not support sharing automatic cluster snapshots. To share an automatic snapshot, you must manually create a copy of the snapshot, and then share the copy.

Can I share my Amazon DocumentDB snapshots across different regions?

No. Your shared Amazon DocumentDB snapshots will only be accessible by accounts in the same region as the account that shares them.

Can I share an encrypted Amazon DocumentDB snapshot?

Yes. You can share encrypted Amazon DocumentDB snapshots. The recipient of the shared snapshot must have access to the KMS key that was used to encrypt the snapshot.

Can I use Amazon DocumentDB snapshots outside of the service?

No. Amazon DocumentDB snapshots can only be used inside of the service.

What happens to my backups if I delete my cluster?

You can choose to create a final snapshot when deleting your cluster. If you do, you can use this snapshot to restore the deleted cluster at a later date. Amazon DocumentDB retains this final user-created snapshot along with all other manually created snapshots after the cluster is deleted.

High Availability and Replication

How does Amazon DocumentDB improve my cluster's fault tolerance to disk failures?

Amazon DocumentDB automatically divides your storage volume into 10 GB segments spread across many disks. Each 10 GB chunk of your storage volume is replicated six ways, across three Availability Zones (AZs). Amazon DocumentDB is designed to transparently handle the loss of up to two copies of data without affecting write availability and up to three copies without affecting read availability. Amazon DocumentDB's storage volume is also self-healing. Data blocks and disks are continuously scanned for errors and repaired automatically.

How does Amazon DocumentDB improve recovery time after a database crash?

Unlike other databases, after a database crash, Amazon DocumentDB does not need to replay the redo log from the last database checkpoint (typically five minutes) and confirm that all changes have been applied, before making the database available for operations. This reduces database restart times to less than 60 seconds in most cases. Amazon DocumentDB moves the cache out of the database process and makes it available immediately at restart time. This prevents you from having to throttle access until the cache is repopulated to avoid brownouts.

What kind of replicas does Amazon DocumentDB support?

Amazon DocumentDB supports read replicas, which share the same underlying storage volume as the primary instance. Updates made by the primary instance are visible to all Amazon DocumentDB replicas.

Feature: Amazon DocumentDB read replicas

Number of replicas: Up to 15

Replication Type: Asynchronous (typically milliseconds)

Performance impact on primary: Low

Act as failover target: Yes (no data loss)

Automated failover: Yes

Can I have cross-region replicas with Amazon DocumentDB?

Yes, you can replicate your data across regions using the Global Cluster feature. Global Clusters span across multiple AWS Regions. Global clusters replicate your data to clusters in up to five Regions with little to no impact on performance. Global clusters provide faster recovery from Region-wide outages and enable low-latency global reads. To learn more see our blog post.

Can I prioritize certain replicas as failover targets over others?

Yes. You can assign a promotion priority tier to each instance on your cluster. If the primary instance fails, Amazon DocumentDB will promote the replica with the highest priority to primary. If there are inconsistencies between two or more replicas in the same priority tier, then Amazon DocumentDB will promote the replica that is the same size as the primary instance.

Can I modify priority tiers for instances after they have been created?

You can modify the priority tier for an instance at any time. Simply modifying priority tiers will not trigger a failover.

Can I prevent certain replicas from being promoted to the primary instance?

You can assign lower priority tiers to replicas that you do not want promoted to the primary instance. However, if the higher priority replicas on the cluster are unhealthy or unavailable for some reason, then Amazon DocumentDB will promote the lower priority replica.

How does Amazon DocumentDB assure high availability of my cluster?

Amazon DocumentDB can be deployed in a high-availability configuration by using replica instances in multiple AWS Availability Zones as failover targets. In the event of a primary instance failure, a replica instance is automatically promoted to be the new primary with minimal service interruption.

How can I improve upon the availability of a single Amazon DocumentDB instance?

You can add additional Amazon DocumentDB replicas. Amazon DocumentDB replicas share the same underlying storage as the primary instance. Any Amazon DocumentDB replica can be promoted to become primary without any data loss and therefore can be used for enhancing fault tolerance in the event of a primary instance failure. To increase cluster availability, simply

create one to 15 replicas, in multiple AZs, and Amazon DocumentDB will automatically include them in failover primary selection in the event of an instance outage.

What happens during failover and how long does it take?

Failover is automatically handled by Amazon DocumentDB so that your applications can resume database operations as quickly as possible without manual administrative intervention.

- If you have an Amazon DocumentDB replica instance in the same or a different Availability Zone, when failing over, Amazon DocumentDB flips the canonical name record (CNAME) for your instance to point at the healthy replica, which is in turn promoted to become the new primary. Start-to-finish, failover typically completes within 30 seconds.
- If you do not have an Amazon DocumentDB replica instance (i.e. a single instance cluster), Amazon DocumentDB will attempt to create a new instance in the same Availability Zone as the original instance. This replacement of the original instance is done on a best-effort basis and may not succeed, for example, if there is an issue that is broadly affecting the Availability Zone.

Your application should retry database connections in the event of connection loss.

If I have a primary instance and an Amazon DocumentDB replica instance actively taking read traffic and a failover occurs, what happens?

Amazon DocumentDB will automatically detect a problem with your primary instance and begin routing your read/write traffic to an Amazon DocumentDB replica instance. On average, this failover will complete within 30 seconds. In addition, the read traffic that your Amazon DocumentDB replicas instances were serving will be briefly interrupted.

How far behind the primary will my replicas be?

Since Amazon DocumentDB replicas share the same data volume as the primary instance, there is virtually no replication lag. We typically observe lag times in the 10s of milliseconds.

Security and Compliance

Can I use Amazon DocumentDB in Amazon Virtual Private Cloud (Amazon VPC)?

Yes. All Amazon DocumentDB clusters must be created in a VPC. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon DocumentDB clusters.

Does Amazon DocumentDB support role-based access control (RBAC)?

Amazon DocumentDB supports RBAC with built-in roles. RBAC enables you to enforce least privilege as a best practice by restricting the actions that users are authorized to perform. For more information, see Amazon DocumentDB role-based access control.

How do the existing MongoDB authentication modes work with Amazon DocumentDB?

Amazon DocumentDB utilizes VPC's strict network and authorization boundary. Authentication and authorization for Amazon DocumentDB management APIs is provided by IAM users, roles, and policies. Authentication to an Amazon DocumentDB database are done via standard MongoDB tools and drivers with Salted Challenge Response Authentication Mechanism (SCRAM), the default authentication mechanism for MongoDB.

Does Amazon DocumentDB support encrypting my data-at-rest?

Yes. Amazon DocumentDB allows you to encrypt your clusters using keys you manage through AWS Key Management Service (KMS). On a cluster running with Amazon DocumentDB encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, snapshots, and replicas in the same cluster. Encryption and decryption are handled seamlessly. For more information about the use of KMS with Amazon DocumentDB, see the Encrypting Amazon DocumentDB Data at Rest.

Can I encrypt an existing unencrypted cluster?

Currently, encrypting an existing unencrypted Amazon DocumentDB cluster is not supported. To use Amazon DocumentDB encryption for an existing unencrypted cluster, create a new cluster with encryption enabled and migrate your data into it.

What compliance certifications does Amazon DocumentDB meet?

Amazon DocumentDB was designed to meet the highest security standards and to make it easy for you to verify our security and meet your own regulatory and compliance obligations. Amazon DocumentDB has been assessed to comply with PCI DSS, ISO 9001, 27001, 27017, and 27018, SOC 1, 2 and 3, and Health Information Trust Alliance (HITRUST) Common Security Framework (CSF) certification, in addition to being HIPAA eligible. AWS compliance reports are available for download in AWS Artifact.

Comparing Amazon DocumentDB and MongoDB

It may seem that there are multiple options for deploying MongoDB in the cloud. Amazon DocumentDB, for example, claims to be an AWS-native database with full support for the MongoDB API. While they support some MongoDB features, it is important to note that DocumentDB is not fully compatible. The only place to access fully-featured MongoDB as a service on AWS is through MongoDB Atlas.

In this comparison we aim to explain the differences between Amazon DocumentDB and MongoDB Atlas, focusing around the following key areas:

- **Compatibility:** Amazon claims that migrating an application from MongoDB to DocumentDB is "as easy as changing the database endpoint to the new Amazon DocumentDB cluster". We have debunked this by running a suite of compatibility tests, which show that DocumentDB is only ~35% compatible with the features and functionality available in MongoDB Atlas.
- **Architecture:** Amazon DocumentDB is built on top of AWS's Aurora platform, which has historically been used to host relational databases. This

has implications on scalability and the potential for adding more native MongoDB features in the future..

- **Deployment:** Amazon's DocumentDB relies on AWS's proprietary technology, and there is no way to run the database outside of AWS.
- **Developer Productivity:** DocumentDB lacks native integration with features or tools to support mobile, time series, search, and analytical use cases. Users must take on the operational burden of moving data out of DocumentDB and into other services in order to access features to support those workloads.
- **Operational Maturity:** Database management and maintenance in DocumentDB is held back by limited tooling, a complex upgrade process, few backup options, and gaps in security features.

What is DocumentDB?

Amazon DocumentDB is a NoSQL JSON document database service with a limited degree of compatibility with MongoDB.

DocumentDB is **not based on the MongoDB server**. Rather it emulates the MongoDB API, and runs on top of Amazon's Aurora backend platform. This creates significant architectural constraints, functionality limitations and broken compatibility.

DocumentDB claims to support the MongoDB 4.0 API, which implies that it is at parity with MongoDB v4.0, released back in June 2018. In actual fact, the DocumentDB 4.0 feature set more closely resembles MongoDB 3.0 and 3.2, released in 2015. In addition, compatibility testing reveals it **fails over 64% of the MongoDB API correctness tests**. Applications written for MongoDB will need to be re-written to work with Amazon DocumentDB.

Interested in up-to-date results on DocumentDB's compatibility with the MongoDB API? Get the latest results at Is DocumentDB Really MongoDB?

The key differences between DocumentDB and MongoDB's on-demand, elastic, and fully managed Atlas service are summarized below.

	Amazon DocumentDB	MongoDB Atlas
Fully compatible with MongoDB	**No, incomplete** Imitation API fails 64% of correctness tests	**Yes**
Support for latest MongoDB version	**No** Feature set resembles MongoDB 3.0/3.2, released in 2015.	**Yes** MongoDB 6.0

	Amazon DocumentDB	MongoDB Atlas
Scale writes and partition data beyond a single node / Sharding support	**Limited** No sharding, single primary for write operations. Largest instance supports 30,000 concurrent connections.	**Yes** Largest Atlas instance supports 128,000 concurrent connections. Full sharding support.
Replicate and scale beyond a single region / Comply with data locality regulations and survive regional outages	**No** Single primary constrained to a single region, with up to 15 replicas	**Yes** Global clusters, with up to 50 replicas per shard across multiple regions
High resilience, rapid failure recovery, fast failover, retryable writes, multi-region	**No** ~120 second failover, no retryable writes, no multi-region in a single cluster	**Yes** Typical failover sub-5 seconds, retryable reads and writes, multi-region & multi-cloud clusters
Multi-statement distributed ACID transactions	**Limited** Ambiguous commits, poor error handling, small data sizes. Transactions unavailable as sharding is unsupported	**Yes**

	Amazon DocumentDB	MongoDB Atlas
Integrated text search, geospatial processing, graph traversals	**Limited** Data replicated to multiple prerequisite bolt-on AWS services for text search, and only basic geospatial operators, adding cost and complexity	**Yes** All available from a single API and platform
Native support for time series data	**No**	**Yes**
Hedged Reads Queries submitted to multiple replicas for consistent low latency	**No**	**Yes**
Online Archive Automatically tier data out from database to cloud object storage (Amazon S3)	**No**	**Yes** Online Archive
Integrated querying of data in Amazon S3	**No** Data must be replicated to multiple adjacent AWS	**Yes** Atlas Query Federation

	Amazon DocumentDB	MongoDB Atlas
	services, driving up cost and complexity	
On-demand Materialized Views	**No**	**Yes** $merge aggregation stage
Schema governance	**No** Schema controls must be enforced in the app	**Yes** JSON schema
Rich data types	**Limited** DocumentDB supports storing decimal128 values, but none of the powerful aggregation features for working with them.	**Yes**
Reactive, event-driven data pipelines	**Limited** Change streams run against primary only & incur additional cost.	**Yes** MongoDB Change Streams & Atlas Triggers
Support for role-based access control and authentication restrictions	**Limited** Coarse-grained roles only	**Yes**

	Amazon DocumentDB	MongoDB Atlas
Client-side field level encryption for fine-grained separation of duties in the cloud	**No**	**Yes**
Availability of advanced developer and analysis tools	**Limited**	**Yes** MongoDB Compass, Charts, SQL Connector, Tableau Connector, Spark Connector
Fine-grained monitoring telemetry & prescriptive performance recommendations	**No** <50 metrics	**Yes** 100+ metrics. Performance Advisor for index and schema recommendations
Freedom from vendor lock-in	**No** AWS only	**Yes** AWS, Azure, and Google Cloud. Available on ~100 regions
Develop & run anywhere On your mobile	**No** AWS only	**Yes** Atlas App Services / MongoDB Server /

	Amazon DocumentDB	MongoDB Atlas
device, on your laptop, in your own data center, and in hybrid configurations		multi-cloud clusters in Atlas
Mobile & Edge Sync	**No** Data must be replicated to multiple adjacent AWS services, driving up cost and complexity	**Yes** Atlas Device Sync
Access to MongoDB expertise	**No**	**Yes** Hundreds of engineers with multi-year MongoDB development, support, and consulting experience.

What is MongoDB Atlas?

MongoDB offers the fully managed, on-demand and global MongoDB Atlas service, in the public cloud. Atlas enables customers to deploy, operate, and scale MongoDB databases on AWS, Azure, or Google Cloud. MongoDB Atlas is available through a pay-as-you-go model and billed on an hourly basis. It's easy to get started – use a simple GUI or programmatic API calls to select the public cloud provider, region, instance size, and features you need. MongoDB Atlas provides:

- Automated database and infrastructure provisioning along with auto-scaling so teams can get the database resources they need, when they need them, and can elastically scale in response to application demands.
- Security features to protect your data, with network isolation, fine-grained access control, auditing, and end-to-end encryption down to the level of individual fields. enabling you to comply with regulations such as ISO, HIPAA and GDPR. Built in replication both within and across regions for always-on availability.
- Global Clusters for a fully managed, globally distributed database that provides low latency, responsive reads and writes to users anywhere, with strong data placement controls for regulatory compliance.
- Combined transactional and analytical capabilities with Atlas Analytics Nodes to isolate analytics queries from operational workloads while providing real-time insight. Native MongoDB analytics tools, such as MongoDB Charts and MongoDB Connectors (for SQL, Tableau and Spark) are configured to utilize Analytics Nodes by default. MongoDB's rich aggregation pipeline engine allows users to run expressive queries against their data.
- Fully integrated native MongoDB data visualization tools – MongoDB Charts, which supports the full richness of the document model, including nested, hierarchical and geospatial data and provide embedding and sharing capabilities. Quickly build visualizations of your data without needing to deploy or manage any software or infrastructure.
- Fully integrated MongoDB Atlas Data Lake which allows users to quickly and easily query data in any format on Amazon S3 using the MongoDB Query Language (MQL) and tools and get value from data faster.
- Fully managed backups with point in time recovery to protect against data corruption, and the ability to query backups in-place without full restores.
- Automatic data tiering which helps lower costs by moving data to lower cost storage such as Amazon S3.
- Fine-grained monitoring and customizable alerts for comprehensive performance visibility. Automated patching and single-click upgrades for new major versions of the database, enabling you to take advantage of the latest MongoDB features.

- Access to the Atlas App Services, with Data Access Controls, Functions, and Triggers delivered in a completely serverless model.
- Native time-series support optimized for both highly performant data ingestion and querying, along with reduced I/O and storage overhead
- The full text Atlas Search service providing rich search capabilities against your fully managed databases with no additional infrastructure or systems to provision, manage, or scale.
- Live migration to move your self-managed MongoDB clusters into the Atlas service or to move Atlas clusters between cloud providers.
- Stable API to make upgrades risk-free, future-proofing your development
- Widespread coverage on the major cloud platforms with availability in ~100 cloud regions across Amazon Web Services, Microsoft Azure, and Google Cloud. MongoDB Atlas delivers a consistent experience across each of the cloud platforms, ensuring developers can deploy wherever they need to, without compromising critical functionality or risking lock-in.

MongoDB Atlas is serving a vast range of workloads for startups, Fortune 500 companies, and government agencies, including mission-critical applications handling highly sensitive data in regulated industries. The developer experience across MongoDB Atlas and self-managed MongoDB is consistent, ensuring that you easily move from on-premises to the public cloud, and between providers as your needs evolve.

Beyond the database, the MongoDB Atlas Data Lake allows you to quickly and easily query data in any format on Amazon S3 using the MongoDB Query Language (MQL) and tools. You can spin up your data lake right alongside your Atlas OLTP clusters from a common UI. With Atlas Data Lake you can realize the value of your data lake faster: you don't have to move data anywhere, you can work with complex data immediately in its native form, and with its fully-managed, serverless architecture, you control costs and remove the operational burden. DocumentDB offers no equivalent capability, and so users must spin up an entirely separate service with a different query language to access S3 data.

Built and run by the same team that develops the data platform, MongoDB Atlas is the best way to run MongoDB apps.

Is DocumentDB Compatible with MongoDB?

Amazon DocumentDB claims to support the MongoDB 4.0 API, which implies that it is at parity with MongoDB v4.0, released back in June 2018. In actual fact, the DocumentDB 4.0 feature set still closely resembles early MongoDB 3.0 and 3.2, released in 2015, and compatibility testing reveals it fails 64% of the MongoDB API correctness tests. Applications written for MongoDB will need to be rewritten to work with Amazon DocumentDB. However, since DocumentDB emulates a MongoDB API, applications written for DocumentDB can be easily migrated into MongoDB Atlas.

FAQs

When should I use a document database?

Document databases store data in JSON-like documents which allows them to be inherently easy for developers to work with. MongoDB is a modern document database which contains many of the powerful features found in popular relational database. These include ACID-compliant transactions, schema governance, enterprise-grade security, and more.

Can MongoDB read data from Amazon S3?

MongoDB offers the Atlas Data Federation engine, which allows users to quickly and easily query data in any format on Amazon S3 using the MongoDB Query API.

What are the key differences between MongoDB and DocumentDB?

DocumentDB does not contain any MongoDB code, supports a limited subset of MongoDB features, and is more costly and difficult to manage.

Is MongoDB a good fit for mobile and web applications?

As part of its developer data platform, MongoDB Atlas Application Services offers a suite of tools for building mobile and web applications. This includes the Realm database.

Amazon OpenSearch Service

Securely unlock real-time search, monitoring, and analysis of business and operational data.

How it works

Amazon OpenSearch Service makes it easy for you to perform interactive log analytics, real-time application monitoring, website search, and more. OpenSearch is an open source, distributed search and analytics suite derived from Elasticsearch. Amazon OpenSearch Service is the successor to Amazon Elasticsearch Service, and offers the latest versions of OpenSearch, support for 19 versions of Elasticsearch (1.5 to 7.10 versions), as well as visualization capabilities powered by OpenSearch Dashboards and Kibana (1.5 to 7.10 versions). Amazon OpenSearch Service currently has tens of thousands of active customers with hundreds of thousands of clusters under management processing hundreds of trillions of requests per month.

Introduction to Amazon OpenSearch Service

Amazon OpenSearch Service securely unlocks real-time search, monitoring, and analysis of business and operational data for use cases like application monitoring, log analytics, observability, and website search.

Features

AWS manages the software installation, upgrades, patching, scaling (up to 3 PB), and cross-region replication with no downtime. Amazon OpenSearch Service is also bundled with a dashboard visualization tool, OpenSearch Dashboards, which helps visualize not only log and trace data, but also machine-learning powered results for anomaly detection and search relevance ranking.

Managed OpenSearch
Focus on analysis instead of spending time managing your deployment, and adjusting deployment configurations as requirements change—while using the power of open source search.

Secure
Meet and maintain high security for authentication, authorization, encryption, audit, and regulatory compliance.

Observability
Deliver log and trace analytics solutions while developing interactive queries and visualizing results with high adaptability and speed.

Cost-conscious
Integrate fast, scalable full-text search capabilities. Manage growing analytics costs for hot, UltraWarm, and cold tiers. All features are included without upsell.

Use cases

Monitor and debug applications and infrastructure
Easily store and analyze data for comprehensive visibility into your system performance with observability logs, metrics, and traces. Set up automated alerts when your system underperforms and find root cause for availability issues.

Manage security and event information (SIEM)
Centralize and analyze logs from disparate applications and systems across your network for real-time threat detection and incident management.

Enable seamless, personalized search
Help users quickly find relevant data with a fast, personalized search experience within your applications, websites, and data lake catalogs.

Observability
Efficiently find and fix problems, improve application health, and deliver better customer experiences.

What is Amazon OpenSearch Service?

Amazon OpenSearch Service is a managed service that makes it easy to deploy, operate, and scale OpenSearch clusters in the AWS Cloud. Amazon OpenSearch Service supports OpenSearch and legacy Elasticsearch OSS (up to 7.10, the final open source version of the software). When you create a cluster, you have the option of which search engine to use.

OpenSearch is a fully open-source search and analytics engine for use cases such as log analytics, real-time application monitoring, and clickstream analysis. For more information, see the OpenSearch documentation.

Amazon OpenSearch Service provisions all the resources for your cluster and launches it. It also automatically detects and replaces failed OpenSearch Service nodes, reducing the overhead associated with self-managed infrastructures. You can scale your cluster with a single API call or a few clicks in the console.

To get started using OpenSearch Service, you create an OpenSearch Service *domain*, which is equivalent to an OpenSearch *cluster*. Each EC2 instance in the cluster acts as one OpenSearch Service node.

You can use the OpenSearch Service console to set up and configure a domain in minutes. If you prefer programmatic access, you can use the AWS CLI or the AWS SDKs.

Features of Amazon OpenSearch Service

OpenSearch Service includes the following features:

Scale

- Numerous configurations of CPU, memory, and storage capacity known as *instance types*, including cost-effective Graviton instances
- Up to 3 PB of attached storage
- Cost-effective UltraWarm and cold storage for read-only data

Security

- AWS Identity and Access Management (IAM) access control
- Easy integration with Amazon VPC and VPC security groups
- Encryption of data at rest and node-to-node encryption

- Amazon Cognito, HTTP basic, or SAML authentication for OpenSearch Dashboards
- Index-level, document-level, and field-level security
- Audit logs
- Dashboards multi-tenancy

Stability

- Numerous geographical locations for your resources, known as *Regions* and *Availability Zones*
- Node allocation across two or three Availability Zones in the same AWS Region, known as *Multi-AZ*
- Dedicated master nodes to offload cluster management tasks
- Automated snapshots to back up and restore OpenSearch Service domains

Flexibility

- SQL support for integration with business intelligence (BI) applications
- Custom packages to improve search results

Integration with popular services

- Data visualization using OpenSearch Dashboards
- Integration with Amazon CloudWatch for monitoring OpenSearch Service domain metrics and setting alarms
- Integration with AWS CloudTrail for auditing configuration API calls to OpenSearch Service domains
- Integration with Amazon S3, Amazon Kinesis, and Amazon DynamoDB for loading streaming data into OpenSearch Service
- Alerts from Amazon SNS when your data exceeds certain thresholds

Related services

OpenSearch Service commonly is used with the following services:

Amazon CloudWatch

OpenSearch Service domains automatically send metrics to CloudWatch so that you can monitor domain health and performance. For more information, see Monitoring OpenSearch cluster metrics with Amazon CloudWatch.

CloudWatch Logs can also go the other direction. You might configure CloudWatch Logs to stream data to OpenSearch Service for analysis. To learn more, see Loading streaming data from Amazon CloudWatch.

AWS CloudTrail

Use AWS CloudTrail to get a history of the OpenSearch Service configuration API calls and related events for your account. For more information, see Monitoring Amazon OpenSearch Service API calls with AWS CloudTrail.

Amazon Kinesis

Kinesis is a managed service for real-time processing of streaming data at a massive scale. For more information, see Loading streaming data from Amazon Kinesis Data Streams and Loading streaming data from Amazon Kinesis Data Firehose.

Amazon S3

Amazon Simple Storage Service (Amazon S3) provides storage for the internet. This guide provides Lambda sample code for integration with Amazon S3. For more information, see Loading streaming data from Amazon S3.

AWS IAM

AWS Identity and Access Management (IAM) is a web service that you can use to manage access to your OpenSearch Service domains. For more information, see Identity and Access Management in Amazon OpenSearch Service.

AWS Lambda

AWS Lambda is a compute service that lets you run code without provisioning or managing servers. This guide provides Lambda sample code to stream data from DynamoDB, Amazon S3, and Kinesis. For more information, see Loading streaming data into Amazon OpenSearch Service.

Amazon DynamoDB

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. To learn more about streaming data to OpenSearch Service, see Loading streaming data from Amazon DynamoDB.

Amazon QuickSight

You can visualize data from OpenSearch Service using Amazon QuickSight dashboards. For more information, see Using Amazon OpenSearch Service with Amazon QuickSight in the *Amazon QuickSight User Guide*.

Note

OpenSearch includes certain Apache-licensed Elasticsearch code from Elasticsearch B.V. and other source code. Elasticsearch B.V. is not the source of that other source code. ELASTICSEARCH is a registered trademark of Elasticsearch B.V.

What Is Elasticsearch?

Elasticsearch is a distributed search and analytics engine built on Apache Lucene. Since its release in 2010, Elasticsearch has quickly become the most popular search engine and is commonly used for log analytics, full-text search, security intelligence, business analytics, and operational intelligence use cases.

On January 21, 2021, Elastic NV announced that they would change their software licensing strategy and not release new versions of Elasticsearch and Kibana under the permissive Apache License, Version 2.0 (ALv2) license. Instead, new versions of the software will be offered under the Elastic license, with source code available under the Elastic License or SSPL. These licenses are not open source and do not offer users the same freedoms. To ensure that the open source community and our customers continue to have a secure, high-quality, fully open source search and analytics suite, we introduced the OpenSearch project, a community-driven, ALv2 licensed fork of open source Elasticsearch and Kibana.

How does Elasticsearch work?

You can send data in the form of JSON documents to Elasticsearch using the API or ingestion tools such as Logstash and Amazon Kinesis Firehose. Elasticsearch automatically stores the original document and adds a searchable reference to the document in the cluster's index. You can then search and retrieve the document using the Elasticsearch API. You can also use Kibana, a visualization tool, with Elasticsearch to visualize your data and build interactive dashboards.

You can run Apache 2.0 licensed Elasticsearch versions (up until version 7.10.2 & Kibana 7.10.2) on-premises, on Amazon EC2, or on Amazon OpenSearch Service. With on-premises or Amazon EC2 deployments, you are responsible for installing Elasticsearch and other necessary software, provisioning infrastructure, and managing the cluster. Amazon OpenSearch Service, on the other hand, is a fully managed service, so you don't have to worry about time-consuming cluster management tasks such as hardware provisioning, software patching, failure recovery, backups, and monitoring.

Elasticsearch benefits

Fast time-to-value

Elasticsearch offers simple REST based APIs, a simple HTTP interface, and uses schema-free JSON documents, making it easy to get started and quickly build applications for a variety of use-cases.

High performance

The distributed nature of Elasticsearch enables it to process large volumes of data in parallel, quickly finding the best matches for your queries.

Complimentary tooling and plugins

Elasticsearch comes integrated with Kibana, a popular visualization and reporting tool. It also offers integration with Beats and Logstash, while enable you to easily transform source data and load it into your Elasticsearch cluster. You can also use a number of open-source Elasticsearch plugins such as language analyzers and suggesters to add rich functionality to your applications.

Near real-time operations

Elasticsearch operations such as reading or writing data usually take less than a second to complete. This lets you use Elasticsearch for near real-time use cases such as application monitoring and anomaly detection.

Easy application development

Elasticsearch provides support for various languages including Java, Python, PHP, JavaScript, Node.js, Ruby, and many more.

Getting started with Elasticsearch on AWS

Managing and scaling Elasticsearch can be difficult and requires expertise in Elasticsearch setup and configuration. To make it easy for customers to run open-source Elasticsearch, AWS offers Amazon OpenSearch Service to perform interactive log analytics, real-time application monitoring, website search, and more.

To learn more about OpenSearch and the ways to operationalize it, please click here.

OpenSearch includes certain Apache-licensed Elasticsearch code from Elasticsearch B.V. and other source code. Elasticsearch B.V. is not the source of that other source code. ELASTICSEARCH is a registered trademark of Elasticsearch B.V.

What is OpenSearch?

OpenSearch is a distributed, community-driven, Apache 2.0-licensed, 100% open-source search and analytics suite used for a broad set of use cases like real-time application monitoring, log analytics, and website search. OpenSearch provides a highly scalable system for providing fast access and response to large volumes of data with an integrated visualization tool, OpenSearch Dashboards, that makes it easy for users to explore their data. OpenSearch is powered by the Apache Lucene search library, and it supports a number of search and analytics capabilities such as k-nearest neighbors (KNN) search, SQL, Anomaly Detection, Machine Learning Commons, Trace Analytics, full-text search, and more.

Why should I use OpenSearch?

OpenSearch enables you to easily ingest, secure, search, aggregate, view, and analyze data for a number of use cases such as log analytics, application search, enterprise search, and more. With OpenSearch, you benefit from having a 100% open source product you can use, modify,

extend, monetize, and resell however you want. There are a growing number of OpenSearch Project partners that offer a variety of services such as professional support, enhanced features, and managed OpenSearch services. The OpenSearch Project continues to provide a secure, high-quality search and analytics suite with a rich roadmap of new and innovative functionality.

Why was OpenSearch created?

Developers embrace open source software for many reasons, one of the most important being the freedom to use that software where and how they wish. On January 21, 2021, Elastic NV announced that they would change their software licensing strategy and not release new versions of Elasticsearch and Kibana under the permissive Apache License, Version 2.0 (ALv2). Instead, Elastic is releasing Elasticsearch and Kibana with source code available under the Elastic License or Server Side Public License (SSPL). These licenses are not open source and do not offer users the same freedoms. Because some developers want their software to be open source and because they want it to avoid single vendor lock-in, we made the decision to create and maintain a fork from the last ALv2 version of Elasticsearch and Kibana. The fork is called OpenSearch and is available under ALv2.

How does OpenSearch relate to Amazon OpenSearch Service?

Amazon OpenSearch Service is an AWS-managed service that lets you run and scale OpenSearch clusters without having to worry about managing, monitoring, and maintaining your infrastructure, or having to build in-depth expertise in operating OpenSearch clusters. We launched support for OpenSearch in Amazon OpenSearch Service in Sep 2021, starting with version 1.0, and renamed the service from Amazon Elasticsearch Service to Amazon OpenSearch Service. Since then, the service has supported several new features that were added to OpenSearch across multiple versions. Some examples are support for cross-cluster replication, trace analytics, data streams, transforms, a new observability user interface, and notebooks in OpenSearch Dashboards. In addition to this, there have been significant improvements to k-NN, anomaly detection, PPL, SQL, and alerting. While Amazon OpenSearch Service continues to support legacy open-source Elasticsearch versions (until 7.10), we strongly recommend that you use OpenSearch on the service to benefit from new features that are being built as part of open-source OpenSearch versions.

Will Amazon OpenSearch Service support new Elasticsearch versions beyond 7.10?

No. Elasticsearch versions beyond 7.10 are not open source and are not released under the permissive ALv2 license. While Amazon OpenSearch Service will continue to support open source Elasticsearch versions until 7.10, moving forward, we will leverage new versions of open source OpenSearch to deliver features and innovations for customers using Amazon OpenSearch Service for their various use cases such as log analytics, search, and observability.

Will OpenSearch maintain compatibility and feature parity with new Elasticsearch versions?

The roadmap for OpenSearch is community-driven, and various organizations including AWS contribute new features to the open source code based. OpenSearch is used for a broad set of

use cases like real-time application monitoring, log analytics, and website search. The roadmap for OpenSearch should be viewed as independent to that of Elasticsearch, and the focus for OpenSearch will be to provide new features and innovations that the community and customers ask for. While OpenSearch may include similar features to new features available in Elasticsearch (and vice versa), the implementation of all features are unique between the two projects. The primary goal for OpenSearch will be to build what best suits the needs of the OpenSearch community and our customers.

What are some features that OpenSearch provides?

Feature	Benefit
Advanced Security	Offers encryption, authentication, authorization, and auditing features. They include integrations with Active Directory, LDAP, SAML, Kerberos, JSON web tokens, and more. OpenSearch also provides fine-grained role-based access control to indices, documents and fields.
Built-in Search Capabilities	Offers a number of features to help you customize your search experience such as Full-text querying, Autocomplete, Scroll Search, customizable scoring and ranking, and more.
SQL Query Syntax	Provides the familiar SQL query syntax. Use aggregations, group by, and where clauses to investigate your data. Read data as JSON documents or CSV tables so you have the flexibility to use the format that works best for you.
Search Support in SQL	Enables you to use the familiar SQL query syntax while getting access to the rich set of search capabilities such as fuzzy matching, boosting, phrase matching, and more.
Data Prepper	Data Prepper is a server side data collector capable of filtering, enriching, transforming, normalizing and aggregating data for downstream analytics and visualization. Data Prepper lets users build custom pipelines to improve the operational view of applications.
Trace Analytics	Trace Analytics provides a way to ingest and visualize OpenTelemetry data in OpenSearch. This data can help you find and fix performance problems in distributed applications.
Application Analytics	Use application analytics to create custom observability applications to view the availability status of your systems, where you can combine log events with trace and metric data into a single view of overall system health. This lets you quickly pivot between logs, traces, and metrics to dig into the source of any issues.
Piped Processing Language	The Piped Processing Language provides a familiar query syntax with a comprehensive set of commands delimited by pipes (|) to query data.
Operational Panels	Build operational panels to organize Observability visualizations generated using the Piped Processing Language (PPL).

Event Analytics	Use the Piped Processing Language (PPL) queries to interactively build and view different visualizations of your data including trace log correlations.
ML Commons Library	Use a range of machine learning algorithms like kmeans and anomaly detection to train models and predict trends in your data. ML Commons integrates directly with PPL and the REST API.
Reporting	Schedule, export, and share reports from dashboards, saved searches, alerts and visualizations.
Anomaly Detection	Leverage Machine Learning anomaly detection based on the Random Cut Forest (RCF) algorithm to automatically detect anomalies as your data is ingested. Combine with Alerting to monitor data in near real time and automatically send alert notifications.
Index Management	Define custom policies to automate routine index management tasks, such as rollover and delete, and apply them to indices and index patterns.
Index Transforms	Create a summarized view of your data centered around certain fields, so you can visualize or analyze the data in different ways. For example, suppose that you have airline data that's scattered across multiple fields and categories, and you want to view a summary of the data that's organized by airline, quarter, and then price. You can use a transform job to create a new, summarized index that's organized by those specific categories.
Index Rollups	Pick the fields that interest you and use index rollup to create a new index with only those fields aggregated into coarser time buckets. You can store months or years of historical data at a fraction of the cost with the same query performance.
Performance Analyzer and RCA Framework	Query numerous cluster performance metrics and aggregations. Use PerfTop, the command line interface (CLI) to quickly display and analyze those metrics. Use the root cause analysis (RCA) framework to investigate performance and reliability issues in clusters.
Asynchronous Search	Run complex queries without worrying about the query timing out and with Asynchronous Search queries run in the background. Track query progress and retrieve partial results as they become available.
Trace Analytics	Ingest and visualize OpenTelemetry data for distributed applications. Visualize the flow of events between these applications to identify performance problems.
Alerting	Automatically monitor data and send alert notifications automatically to stakeholders. With an intuitive interface and a powerful API, easily set up, manage, and monitor alerts. Craft highly specific alert conditions using OpenSearch's full query language and scripting capabilities.

Bucket Level Alerting	Create alerting policies that alert on grouped trends in your data. For example, you can alerting for each host that has an average CPU above your desired threshold.
Cross Cluster Replication	Replicate indexes, mappings, and metadata from one OpenSearch cluster to another in order to create cross cluster redundancy or offload reporting querying to a secondary cluster.
k-NN search	Using Machine Learning, run the nearest neighbor search algorithm on billions of documents across thousands of dimensions with the same ease as running any regular OpenSearch query. Use aggregations and filter clauses to further refine similarity search operations. k-NN similarity search powers use cases such as product recommendations, fraud detection, image and video search, related document search, and more.
Dashboard Notebooks	Combine dashboards, visualizations, text, and more to provide context and detailed explanations when analyzing data.
OpenSearch Clients	OpenSearch supports a range of language clients such as Go, JavaScript, Python, Java, and more. Use these clients to build applications that integrate directly with OpenSearch.

Who sponsors and maintains OpenSearch?

Many organizations including AWS, SAP, CapitalOne, RedHat, Logz.io, Aiven.io, Bonsai, Logit.io, InstaCluster, and BAInsight have publicly backed OpenSearch.

We encourage and accept contributions from the community and do not require a Contributor License Agreement (CLA) in order to do so. Additionally, if you are an active contributor and would like to take on more responsibility in the project, we have defined a process for non-AWS employees to gain maintainer rights in OpenSearch project repos here.

What is the future of OpenSearch?

We launched the first Generally Available OpenSearch version 1.0 in July 2021, and since then have launched many new versions, detailed here. With great interest and support from the community, we introduced several new features across key use cases around log analytics, search, and observability. These areas continue to be key themes for future improvements for both OpenSearch and our visualization layer OpenSearch Dashboards. Please see the roadmap for a quick overview of the key areas OpenSearch is focusing on.

How is OpenSearch licensed?

All of the software in the OpenSearch project is released under the Apache License, Version 2.0 (ALv2). ALv2 grants well-understood and permissive usage rights that match the freedoms people expect with open source software: freedoms such as being able use, modify, extend, monetize, and resell the open source software where and how they want. For OpenSearch, we believe this license will enable broad adoption and contributions benefiting all members of the

community. We have also published permissive usage guidelines for the OpenSearch trademark, so you can use the name to promote your offerings.

Amazon OpenSearch Service FAQs

General

Q: What is Amazon OpenSearch Service?

Amazon OpenSearch Service is a managed service that makes it easy for you to perform interactive log analytics, real-time application monitoring, website search, and more. OpenSearch is an open source, distributed search and analytics suite derived from Elasticsearch. Amazon OpenSearch Service offers the latest versions of OpenSearch, support for 19 versions of Elasticsearch (1.5 to 7.10 versions), as well as visualization capabilities powered by OpenSearch Dashboards and Kibana (1.5 to 7.10 versions). Amazon OpenSearch Service currently has tens of thousands of active customers with hundreds of thousands of clusters under management processing trillions of requests per month. See the Amazon OpenSearch Service FAQ for more information.

Q: Which OpenSearch and Elasticsearch versions does Amazon OpenSearch Service support?

Amazon OpenSearch Service offers the latest versions of OpenSearch and support for 19 versions of Elasticsearch (1.5 to 7.10 versions). For more details refer our documentation.

Q: What is an Amazon OpenSearch Service domain?

Amazon OpenSearch Service domains are Elasticsearch (1.5 to 7.10) or OpenSearch clusters created using the Amazon OpenSearch Service console, CLI, or API. Each domain is an OpenSearch or Elasticsearch cluster in the cloud with the compute and storage resources you specify. You can create and delete domains, define infrastructure attributes, and control access and security. You can run one or more Amazon OpenSearch Service domains.

Q: What does Amazon OpenSearch Service manage on my behalf?

Amazon OpenSearch Service manages the work involved in setting up a domain, from provisioning infrastructure capacity in the network environment you request to installing the OpenSearch or Elasticsearch software. Once your domain is running, Amazon OpenSearch Service automates common administrative tasks, such as performing backups, monitoring instances and patching software. Amazon OpenSearch Service integrates with Amazon CloudWatch to produce metrics that provide information about the state of the domains. Amazon OpenSearch Service also offers options to modify your domain instance and storage settings to simplify the task of tailoring your domain based to your application needs.

Q: Does Amazon OpenSearch Service support the open-source Elasticsearch and OpenSearch APIs?

Amazon OpenSearch Service supports most of the commonly used OpenSearch and Elasticsearch APIs, so the code, applications, and popular tools that you're already using with your Elasticsearch (up to version 7.10) or OpenSearch environments work seamlessly. For a full list of supported operations, see our documentation.

Q: What are the Availability Zone (AZ) deployment options available on Amazon OpenSearch Service?

Amazon OpenSearch Service offers customers the option to deploy their instances across one, two, or three AZs. Customers running development or test workloads can pick the single AZ option. Those running production-grade workloads should use two or three AZs. Three AZ deployments are strongly recommended for workloads with higher availability requirements.

Note: The three AZ option is only available in regions where there are three or more AZs.

Q: In which regions does Amazon OpenSearch Service offer three AZ deployments?

Amazon OpenSearch Service supports three AZ deployments in all regions in which the service is available, except US West (N. California), where we support two AZs only.

Setup and configuration

Q: Can I create and modify my Amazon OpenSearch Service domain through the Amazon OpenSearch Service console?

Yes. You can create a new Amazon OpenSearch Service domain with the Domain Creation Wizard in the console with just a few clicks. While creating a new domain you can specify the number of instances, instance types, and EBS volumes you want allocated to your domain. You can also modify or delete existing Amazon OpenSearch Service domains using the console.

Q: Does Amazon OpenSearch Service support Amazon VPC?

Yes, Amazon OpenSearch Service is integrated with Amazon VPC. When choosing VPC access, IP addresses from your VPC are attached to your Amazon OpenSearch Service domain and all network traffic stays within the AWS network and is not accessible to the Internet. Moreover, you can use security groups and IAM policies to restrict access to your Amazon OpenSearch Service domains.

Q: Can I use CloudFormation Templates to provision Amazon OpenSearch Service domains?

Yes. AWS CloudFormation supports Amazon OpenSearch Service. For more information, see the CloudFormation Template Reference documentation.

Q: Does Amazon OpenSearch Service support configuring dedicated master nodes?

Yes. You can configure dedicated master nodes for your domains. When choosing a dedicated master configuration, you can specify the instance type and instance count.

Q: Can I create multiple Elasticsearch or OpenSearch indices within a single Amazon OpenSearch Service domain?

Yes. You can create multiple Elasticsearch or OpenSearch indices within the same Amazon OpenSearch Service domain. Elasticsearch and OpenSearch automatically distribute the indices and any associated replicas between the instances allocated to the domain.

Q: How do I ingest data into my Amazon OpenSearch Service domain?

Amazon OpenSearch Service supports three options for data ingestion:

- For large data volumes, we recommend Amazon Kinesis Data Firehose, a fully managed service that automatically scales to match the throughput of your data and requires no ongoing administration. It can also transform, batch and compress the data before loading it.
- Amazon OpenSearch Service supports integration with Logstash. You can configure your Amazon OpenSearch Service domain as the data store for all logs arriving from your Logstash implementation.
- You can use native Elasticsearch (up to version 7.10) or OpenSearch APIs, such as the index and bulk APIs, to load data into your domain.

Q: Does Amazon OpenSearch Service support integration with Logstash?

Yes. Amazon OpenSearch Service supports integration with Logstash. You can set up your Amazon OpenSearch Service domain as the backend store for all logs coming through your Logstash implementation. You can set up access control on your Amazon OpenSearch Service domain to either use request signing to authenticate calls from your Logstash implementation, or use resource based IAM policies to include IP addresses of instances running your Logstash implementation.

Q: Does Amazon OpenSearch Service support integration with Kibana?

Yes. Amazon OpenSearch Service offers visualization capabilities powered by OpenSearch Dashboards and Kibana (1.5 to 7.10 versions).

Q: What storage options are available with Amazon OpenSearch Service?

You can choose between local on-instance storage or EBS volumes. During domain creation, if you select EBS storage, you can increase and decrease the size of the storage volume as necessary.

Q: What types of EBS volumes does Amazon OpenSearch Service support?

You can choose between Magnetic, General Purpose, and Provisioned IOPS EBS volumes.

Q: Is there a limit on the amount of storage that can be allocated to an Amazon OpenSearch Service domain?

Yes. Amazon OpenSearch Service deploys storage based the choice of instance and/or the size of the associated EBS volume. The maximum storage per node is 24 TB with R6g.12xlarge instances with EBS gp3 storage. With the default maximum of 80 data nodes allowed per Amazon OpenSearch Service domain, you can allocate about 1920 TB of storage to a single domain. You can request a service limit increase up to 200 instances per domain by creating a case with the AWS Support Center. With 200 instances, you can allocate about 3 PB of storage to a single domain.

Q: How are dedicated master instances distributed across AZs?

If you deploy your data instances in a single AZ, your dedicated master instances are also deployed in the same AZ. However, if you deploy your data instances across two or three AZs, Amazon OpenSearch Service automatically distributes the dedicated master instances across three AZs. The exception to this rule occurs if a region only has two AZs or if you select an older-generation instance type for the master instances that is not available in all AZs. For more details, refer our documentation.
Q: What is the recommended AZ configuration for production workloads?

For production workloads, we recommend deploying your data instances across three AZs since it offers better availability. Also, we recommend provisioning instances in multiples of three for equal distribution across AZs. In regions where three AZs are not available, we recommend using a two AZ deployment with an even number of data instances. In all cases, we recommend provisioning three dedicated master instances.
Q: How can I configure my domain for three AZ deployment?

You can enable three AZ deployment for both existing and new domains using the AWS console, CLI or SDKs. For more details, refer our documentation.
Q: Is there a fee for enabling three AZ deployment?

No. Amazon OpenSearch Service does not charge anything for enabling three AZ deployment. You only pay for the number of instances in your domain, not the number of AZs to which they are deployed.
Q: I no longer see the “zone awareness” option in my console. Is my domain no longer zone aware?

All domains configured for multiple AZs will have zone awareness enabled to ensure shards are distributed across Availability Zones. In the console, you can now explicitly choose two or three AZ deployments. Domains previously configured with “Zone Awareness” will continue to be deployed across two AZs unless they are reconfigured. For more details, refer our documentation.
Q: How does Amazon OpenSearch Service handle instance failures and AZ disruptions?

If one or more instances in an AZ are unreachable or not functional, Amazon OpenSearch Service automatically tries to bring up new instances in the same AZ to replace the affected instances. In the rare event that new instances cannot be brought up in the AZ, Amazon OpenSearch Service brings up new instances in the other available AZs if the domain has been configured to deploy instances across multiple AZs. Once the AZ issue resolves, Amazon OpenSearch Service rebalances the instances such that they are equally distributed across the AZs configured for the domain. For more details refer our documentation.
Q: If I have only one replica for the indices in my domain, should I use two or three AZs?

Even when you configure one replica, we recommend three AZs. If an AZ disruption occurs in a three AZ domain, you only lose one-third of your capacity but if the disruption occurs in a two AZ domain, you lose half your capacity, which can be more disruptive. Also, in a three AZ domain, when an AZ is disrupted, Amazon OpenSearch Service can fall back to the two remaining AZs, and still support cross-AZ replication . In a two AZ domain, you lose cross-AZ replication if one AZ is disrupted, which can further reduce availability. For more details refer our documentation.
Q: How do I leverage three AZ deployment for my VPC domain?

The number of AZs your domain is deployed to corresponds to the number of subnets you have configured for your VPC domain. You need to configure at least three subnets in your VPC domain to enable three AZ deployment. For more details on configuring VPC, refer our documentation.

Administration

Q: Can programs running on servers in my own data center access my Amazon OpenSearch Service domains?

Yes. The programs with public Internet access can access Amazon OpenSearch Service domains through a public endpoint. If your data center is already connected to Amazon VPC through Direct Connect or SSH tunneling, you can also use VPC access. In both cases, you can configure IAM policies and security groups to allow programs running on servers outside of AWS to access your Amazon OpenSearch Service domains. Click here for more information about signed requests.
Q: How can I migrate data from my existing OpenSearch/Elasticsearch cluster to a new Amazon OpenSearch Service domain?

To migrate data from an existing Elasticsearch or OpenSearch cluster, you should create a snapshot of an existing cluster, and store the snapshot in your Amazon S3 bucket. Then you can create a new Amazon OpenSearch Service domain and load data from the snapshot into the newly created Amazon OpenSearch Service domain using the restore API.
Q: How can I scale an Amazon OpenSearch Service domain?

Amazon OpenSearch Service allows you to control the scaling of your Amazon OpenSearch Service domains using the console, API, and CLI. You can scale your Amazon OpenSearch Service domain by adding, removing, or modifying instances or storage volumes depending on your application needs. Amazon OpenSearch Service is integrated with Amazon CloudWatch to provide metrics about the state of your Amazon OpenSearch Service domains to enable you to make appropriate scaling decisions for your domains.
Q: Does scaling my Amazon OpenSearch Service domain require downtime?

No. Scaling your Amazon OpenSearch Service domain by adding or modifying instances, and storage volumes is an online operation that does not require any downtime.
Q: Does Amazon OpenSearch Service support cross-zone replication?

Yes. If you enable replicas for your OpenSearch/Elasticsearch indices and use multiple Availability Zones, Amazon OpenSearch Service automatically distributes your primary and replica shards across instances in different AZs.
Q: Does Amazon OpenSearch Service expose any performance metrics through Amazon CloudWatch?

Yes. Amazon OpenSearch Service exposes several performance metrics through Amazon CloudWatch including number of nodes, cluster health, searchable documents, EBS metrics (if applicable), CPU, memory and disk utilization for data and master nodes. Please refer to the service documentation for a full listing of available CloudWatch metrics.
Q: I wish to perform security analysis or operational troubleshooting of my Amazon OpenSearch Service deployment. Can I get a history of all the Amazon OpenSearch Service API calls made on my account?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by AWS CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about AWS CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console home page.
Q: What is a snapshot?

A snapshot is a copy of your Amazon OpenSearch Service domain at a moment in time.
Q: Why would I need snapshots?

Creating snapshots can be useful in case of data loss caused by node failure, as well as the unlikely event of a hardware failure. You can use snapshots to recover your Amazon OpenSearch Service domain with preloaded data or to create a new Amazon OpenSearch Service domain with preloaded data. Another common reason to use backups is for archiving purposes. Snapshots are stored in Amazon S3.
Q: Does Amazon OpenSearch Service provide automated snapshots?

Yes. By default, Amazon OpenSearch Service automatically creates hourly snapshots of each Amazon OpenSearch Service domain and retains them for 14 days.
Q: How long are the automated daily hourly snapshots stored by Amazon OpenSearch Service?

Amazon OpenSearch Service will retain the last 14 days' worth of automated hourly snapshots.
Q: Is there a charge for the automated daily hourly snapshots?

There is no additional charge for the automated hourly snapshots. The snapshots are stored for free in an Amazon OpenSearch Service S3 bucket and will be made available for node recovery purposes.
Q: Can I create additional snapshots of my Amazon OpenSearch Service domains as needed?

Yes. You can use the snapshot API to create additional manual snapshots in addition to the daily-automated snapshots created by Amazon OpenSearch Service. The manual snapshots are stored in your S3 bucket and will incur relevant Amazon S3 usage charges.
Q: Can snapshots created by the manual snapshot process be used to recover a domain in the event of a failure?

Yes. Customers can create a new Amazon OpenSearch Service domain and load data from the snapshot into the newly created Amazon OpenSearch Service domain using the OpenSearch/Elasticsearch restore API.
Q: What happens to my snapshots when I delete my Amazon OpenSearch Service domain?

The daily snapshots retained by Amazon OpenSearch Service will be deleted as part of domain deletion. Before deleting a domain, you should consider creating a snapshot of the domain in your own S3 buckets using the manual snapshot process. The snapshots stored in your S3 bucket will not be affected if you delete your Amazon OpenSearch Service domain.
Q: What types of OpenSearch/Elasticsearch logs are exposed by Amazon OpenSearch Service?

Amazon OpenSearch Service exposes three Elasticsearch or OpenSearch logs through Amazon CloudWatch Logs: error logs, search slow logs, and index slow logs. These logs are useful for troubleshooting performance and stability issues with one's domain.
Q: What are slow logs?

Slow logs are log files that help track the performance of various stages in an operation. OpenSearch and Elasticsearch exposes two kinds of slow logs:

- Index Slow Logs – These logs provide insights into the indexing process and can be used to fine-tune the index setup.
- Search Slow Logs – These logs provide insights into how fast or slow queries and fetches are performing. These logs help fine tune the performance of any kind of search operation on OpenSearch or Elasticsearch.

For complete details on slow logs, please refer to OpenSearch documentation.
Q: How can I enable slow logs on Amazon OpenSearch Service?

Slows logs can be enabled via the click of a button from the Console or via our CLI and APIs. For more details please refer to our documentation.
Q: Can I only enable slow logs for specific indices?

Yes. You can update the settings for a specific index to enable or disable slow logs for it. For more details refer to our documentation.
Q: Does turning on slow logs in Amazon OpenSearch Service automatically enable logging for all indexes?

No. Turning on slow logs in Amazon OpenSearch Service enables the option to publish the generated logs to Amazon CloudWatch Logs for indices in the given domain. However, in order to generate the logs you have to update the settings for one or more indices to start the logging process. For more details on setting the index configuration for enabling slow logs, please refer to our documentation.
Q: If I turn off the slow logs in Amazon OpenSearch Service, does it mean that log files are no longer being generated?

No. The generation of log files are dependent on the index settings. To turn off generation of the log files you have to update the index configuration. For more details on setting the index configuration for enabling slow logs, see our documentation.
Q: Can I change the granularity of logging?

You can only change the granularity of logging for slow logs. OpenSearch and Elasticsearch expose multiple levels of logging for slow logs. You need to set the appropriate level in the configuration of your index. For more details on setting the index configuration for enabling slow logs, please refer to OpenSearch documentation.
Q: Will enabling slow logs or error logs cost me anything?

When slow logs or error logs are enabled, Amazon OpenSearch Service starts publishing the generated logs to CloudWatch Logs. Amazon OpenSearch Service does not charge anything for enabling the logs. However, standard CloudWatch charges will apply.
Q: What kinds of error logs are exposed by Amazon OpenSearch Service?

OpenSearch uses Apache Log4j 2 and its built-in log levels (from least to most severe) of TRACE, DEBUG, INFO, WARN, ERROR, and FATAL. If you enable error logs, Amazon OpenSearch Service publishes log lines of WARN, ERROR, and FATAL, and select errors from the DEBUG level to CloudWatch. For more details, refer our documentation.
Q: How can I enable error logs on Amazon OpenSearch Service?

Error logs can be enabled via the click of a button from the AWS Console or via our CLI and APIs. For more details please refer to our documentation.
Q: Can I enable error logs for only specific indices?

No, error logs are exposed for the entire domain. That is, once enabled, log entries from all indices in the domain will be made available.
Q: Are error logs available for all versions of Elasticsearch supported by Amazon OpenSearch Service?

No, error logs are available only for Elasticsearch versions 5.x and above.
Q: Is there any limit on the size of each log entry?

Yes. Each log entry made into CloudWatch will be limited to 255,000 characters. If your log entry is bigger than that, it will be truncated to 255,000 characters.
Q: What is the recommended best practice for using slow logs?

Slow logs are only needed when you want to troubleshoot your indexes or fine-tune performance. The recommended approach is to only enable logging for those indexes for which you need additional performance insights. Also, once the investigation is done, you should turn off logging so that you don't incur any additional costs on account of it. For more details, see our documentation.

Q: How can I consume logs from CloudWatch Logs?

CloudWatch offers multiple ways to consume logs. You can view log data, export it to S3, or process it in real time. To learn more, see the CloudWatch Logs developer guide.
Q: Are slow logs available for all versions of OpenSearch and Elasticsearch supported by Amazon OpenSearch Service?

Yes, slow logs can be enabled for all versions of OpenSearch and Elasticsearch supported by Amazon OpenSearch Service. However, there are slight differences in the way log settings can be specified for each version of Elasticsearch. Please refer to our documentation for more details.
Q: Will the cluster have any down time when logging is turned on or off?

No. There will not be any down-time. Every time the log status is updated, we will deploy a new cluster in the background and replace the existing cluster with the new one. This process will not cause any down time. However, since a new cluster is deployed the update to the log status will not be instantaneous.
Q: Which Elasticsearch and OpenSearch versions does the in-place upgrade feature support?

Amazon OpenSearch Service currently supports in-place version upgrade for domains with any OpenSearch version or Elasticsearch versions 5.x and above. The target versions that we support for the upgrade are 5.6, 6.3, 6.4, 6.5, 6.7, 6.8, 7.1, 7.4, 7.7, 7.8, 7.9, and 7.10. For more details refer our documentation.
Q: My domain runs a version of Elasticsearch older than 5.x. How do I upgrade those domains?

Please refer to our documentation for details on migrating from various Elasticsearch versions.
Q: Will my domain be offline while the in-place upgrade is in progress?

No. Your domain remains available throughout the upgrade process. However, part of the upgrade process involves relocating shards, which can impact domain performance. We recommend upgrading when the load on your domain is low.
Q: How can I check if my domain's Elasticsearch version can be upgraded?

In-place version upgrade is available only for domains running Elasticsearch 5.x and above. If your domain is of version 5.x or above, you can run the upgrade eligibility check to validate whether your domain can be upgraded to the desired version. Please refer to our documentation to learn more.
Q: What are the tests done by Amazon OpenSearch Service to validate my domains upgrade eligibility?

For detailed list of the tests we run to validate upgrade eligibility, please refer to our documentation.
Q: Can I update my domain configuration while the version upgrade is in progress?

No. Once the in-place version upgrade has been triggered, you cannot make changes to your domain configuration until the upgrade completes or fails. You can continue reading and writing data while the upgrade is in progress. Also, you can delete the domain, in which case the upgrade is terminated and the domain deleted.
Q: What happens to the automated system snapshot when the in-place version upgrade is in progress?

The version upgrade process automatically takes a snapshot of the system and only starts the actual upgrade if the snapshot succeeds. If the upgrade is in progress when the automated snapshot's start time is reached, the automated snapshot is skipped for that day and continued on the next day.

Q: How does Amazon OpenSearch Service safeguard against issues that can crop up during version upgrades?

Amazon OpenSearch Service runs a set of tests before triggering the upgrade to check for known issues that can block the upgrade. If no issues are encountered, the service takes a snapshot of the domain and starts the upgrade process if the snapshot is successful. The upgrade is not triggered if there are any issues encountered with any of the steps.
Q: What happens if the system encounters issues while performing the in-place version upgrade?

If encountered issues are minor and fixable, Amazon OpenSearch Service automatically tries to address them and unblock the upgrade. However, if an issue blocks the upgrade, the service reverts back to the snapshot that was taken before the upgrade and logs the error. For more details on viewing the logs from the upgrade progress, please refer to our documentation.
Q: Can I view the history of upgrades on my domain?

Yes. You can view the upgrade logs from the AWS console or request them using the CLI or SDKs. Please refer to our documentation for more details.
Q: Can I pause or cancel the version upgrade after it has been triggered?

No. After the upgrade has been triggered, it cannot be paused or cancelled until it either completes or fails.
Q: Can I run in-place version upgrade on multiple domains in parallel?

Yes. However, if you want to keep all of your domains on the same version, we recommend running the upgrade eligibility check on all domains before upgrading them. This extra step can help catch issues with one domain that might not be present on others.
Q: How long does the in-place version upgrade take?

Depending on the amount of data and the size of the cluster, upgrades can take anywhere from a few minutes to a few hours to complete.
Q: Can I just upgrade the domain quickly without retaining any of the data?

No. With in-place version upgrade, all the data in your cluster is also restored as part of the upgrade process. If you only wish to upgrade the domain alone, you can take a snapshot of your data, delete all your indexes from the domain and then trigger an in-place version upgrade. Alternatively, you can create a separate domain with the newer version and then restore your data to that domain.
Q: Can I downgrade to previous version if I'm not comfortable with the new version?

No. If you need to downgrade to an older version, contact AWS Support to restore the automatic, pre-upgrade snapshot on a new domain. If you took a manual snapshot of the original domain, you can perform this step yourself.

Security

Q: How can I secure my Amazon OpenSearch Service domain?

Amazon OpenSearch Service provides multiple security features and is HIPAA eligible and compliant with PCI DSS, SOC, ISO, and FedRamp standards, so that you can meet your security and compliance needs. Access to Amazon OpenSearch Service management APIs for operations such as creating and scaling domains are controlled with AWS Identity and Access Management (IAM) policies.

Amazon OpenSearch Service domains can be configured to be accessible with an endpoint within your VPC or a public endpoint accessible to the internet. Network access for VPC

endpoints is controlled with security groups and for public endpoints access can be granted or restricted by IP address.

In addition to network-based access control, Amazon OpenSearch Service provides user authentication via IAM and basic authentication using username and password. Authorization can be granted at the domain level (via Domain Access Policies) as well as at the index, document, and field level (via the fine-grained access control feature powered by OpenSearch). Additionally the fine-grained access control feature extends OpenSearch Dashboards and Kibana with read-only views and secure multi-tenant support.

Amazon OpenSearch Service also supports an integration with Amazon Cognito, to allow your end-users to log-in to OpenSearch Dashboards and Kibana through enterprise identity providers such as Microsoft Active Directory using SAML 2.0, Amazon Cognito User Pools, and more. Once you sign-in, Amazon Cognito establishes a session using the appropriate IAM principal, which provides access to the Amazon OpenSearch Service domain. These IAM principals are then available to be used with the fine-grained access control feature powered by OpenSearch.

Q: How does security authentication and authorization work in Amazon OpenSearch Service?

Amazon OpenSearch Service security has three main layers: Network, Domain access policies, and fine-grained access control. The first security layer is the network, which determines whether requests reach a domain. We support public access via the internet or VPC access limited to specific security groups in your VPC. The domain access policy is the second security layer. After a request reaches a domain endpoint, the Domain Access Policy allows or denies the request access to a given URL. The Domain Access Policy accepts or rejects requests at the edge of the domain, before they reach OpenSearch/Elasticsearch itself. The third and final security layer is fine-grained access control. After a Domain Access Policy allows a request to reach a domain endpoint, fine-grained access control evaluates the user credentials and either authenticates the user or denies the request. If fine-grained access control authenticates the user, it fetches all roles mapped to that user and uses the complete set of permissions to determine what data the user has access to.

Q: Does Amazon OpenSearch Service support encryption?

Yes, Amazon OpenSearch Service supports encryption at rest through AWS Key Management Service (KMS), node-to-node encryption over TLS, and the ability to require clients to communicate of HTTPS. Encryption at rest encrypts shards, log files, swap files, and automated S3 snapshots. You can use AWS-managed keys or choose one of your own. Node-to-node encryption enables TLS for all communications between nodes. Amazon OpenSearch Service automatically deploys and rotates certificates throughout the life of the domain. If you require you clients to communicate over HTTPS, you also have the ability to specify the minimum TLS version.

Q: If I set up VPC access for my Amazon OpenSearch Service domain, how can I access OpenSearch Dashboards and Kibana?

When VPC access is enabled, the endpoint for Amazon OpenSearch Service is only accessible within the customer VPC. To use your laptop to access OpenSearch Dashboards and Kibana from outside the VPC, you need to connect the laptop to the VPC using VPN or VPC Direct Connect.

Reserved Instance pricing

Q: What is a Reserved Instance (RI)?

Amazon OpenSearch Service Reserved Instances give you the option to reserve an instance for a one- or three-year term, and in turn receive significant savings compared to the On-Demand Instance pricing.

Q: How are Reserved Instances different from On-Demand Instances?

Functionally, Reserved Instances and On-Demand Instances are exactly the same. The only difference is how your instance(s) are billed. With Reserved Instances, you purchase a one- or three-year reservation and receive a lower effective hourly usage rate (compared to On-Demand Instances) for the duration of the term. Unless you purchase Reserved Instances in a Region, all instances in that Region are billed at On-Demand Instance hourly rates.

Q: What are the payment options for Reserved Instances?

Three options are available:

- No Upfront Reserved Instances (NURI) – NURIs offer significant savings compared to On-Demand Instance prices. You pay nothing upfront, but commit to paying for the Reserved Instance over the course of the one- or three-year term.
- Partial Upfront Reserved Instances (PURI) – PURIs offer higher savings than NURIs. You pay for a portion of the total cost upfront, and the remainder over the course of the term. This option balances payments between upfront and hourly.
- All Upfront Reserved Instances (AURI) – AURIs offer the highest savings of all of the Reserved Instance payment options. You pay for the entire reservation with one upfront payment, and pay nothing on an hourly basis.

Q: How do I purchase Reserved Instances?

You purchase Reserved Instances in the "Reserved Instance" section of the AWS Management Console for Amazon OpenSearch Service. Alternatively, you can use the Amazon OpenSearch Service API or AWS Command Line Interface to list and purchase Reserved Instances.

Once you purchase a Reserved Instance, you can use it just like an On-Demand Instance. As long as the purchased reservation is active, Amazon OpenSearch Service applies the reduced hourly rate to it.

Q: Are Reserved Instances specific to an Availability Zone?

Amazon OpenSearch Service Reserved Instances are purchased for a Region rather than for a specific Availability Zone. After you purchase a Reserved Instance for a Region, the discount applies to matching usage in any Availability Zone within that Region.

Q: How many Reserved Instances can I purchase?

You can procure up to 100 Reserved Instances in a single purchase. If you need more Reserved Instances, you need to place more purchase requests.

Q: Do Reserved Instances include a capacity reservation?

Amazon OpenSearch Service Reserved Instances are purchased for a Region rather than for a specific Availability Zone. Hence, they are not capacity reservations. Even if capacity is limited in one Availability Zone, Reserved Instances can still be purchased in the Region. The discount applies to matching usage in any Availability Zone within that Region.

Q: What if I have an existing On-Demand Instance that I'd like to convert to a Reserved Instance?

Simply purchase a Reserved Instance of the same type as the existing On-Demand Instance. If the Reserved Instance purchase succeeds, Amazon OpenSearch Service automatically applies the new hourly usage charge for the duration of your reservation.

Q: If I sign up for a Reserved Instance, when does the term begin? What happens to my Reserved Instance when the term ends?

Pricing changes and the reservation term associated with your Reserved Instance become active after your request is received and the payment authorization is processed. If the one-time payment (if applicable) or new hourly rate (if applicable) cannot be successfully authorized by the next billing period, the discounted price does not take effect and your term does not begin. You can follow the status of your reservation using the console, API, or CLI. For more details, refer our documentation.

When your Reserved Instance term expires, your Reserved Instance reverts to the appropriate On-Demand Instance hourly usage rate for your instance class and Region.

Q: How do I control which instances are billed at the Reserved Instance rate?

When computing your bill, our system automatically applies your reservation(s) such that all eligible instances are charged at the lower hourly Reserved Instance rate. Amazon OpenSearch Service does not distinguish between On-Demand and Reserved Instances while operating Amazon OpenSearch Service domains.

Q: If I scale my Reserved Instance up or down, what happens to my reservation?

Each Reserved Instance is associated with the instance type and Region that you picked for it. If you change the instance type in the Region where you have the Reserved Instance, you will not receive discounted pricing. You must ensure that your reservation matches the instance type you plan to use.

Q: Can I move a Reserved Instance from one Region or Availability Zone to another?

Each Reserved Instance is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each Reserved Instance can, however, be used in any of the Availability Zones within the associated Region.

Q: Are Reserved Instances applicable if use multiple Availability Zones?

A Reserved Instance is for an AWS Region and can be used in any of the Availability Zones in that Region.

Q: Are Reserved Instances available for both Master nodes and Data nodes?

Yes. Amazon OpenSearch Service does not differentiate between Master and Data nodes when applying Reserved Instance discounts.

Q: Can I cancel a Reserved Instance?

No, you cannot cancel your Reserved Instances, and the one-time payment (if applicable) and discounted hourly usage rate (if applicable) are not refundable. Also, you cannot transfer the Reserved Instance to another account. You must pay for every hour during your Reserved Instance's term, regardless of your usage.

Q: If I purchase a Reserved Instance from a payer (master) account, is it accessible to all the member accounts?

Yes. Reserved Instance pricing and application follows the policies defined for consolidated billing on AWS.

Q: If AWS reduces prices of On-Demand Instances for Amazon OpenSearch Service, will the amount I pay for my current Reserved Instances change?

No. The price you pay for already-purchased Reserved Instances does not change for the term of the reservation.

Q: Can I sell my Reserved Instances on the Reserved Instance Marketplace?

No. Reserved Instances purchased on Amazon OpenSearch Service cannot be sold on the Reserved Instance Marketplace.

Q: Are volume discounts available for Reserved Instance purchase?

No. We do not offer volume discounts for Amazon OpenSearch Service Reserved Instances.

UltraWarm

Q. What is UltraWarm?

UltraWarm is a fully-managed, low-cost, warm storage tier for Amazon OpenSearch Service. It is compatible with OpenSearch, Elasticsearch (until version 7.10), OpenSearch Dashboards, and Kibana (until version 7.10), enabling you to analyze data using the same tools that Amazon OpenSearch Service provides today. UltraWarm seamlessly integrates with Amazon OpenSearch Service's existing features such as integrated alerting, SQL querying, and more.

Q. Why should I use UltraWarm?

UltraWarm enables you to cost effectively expand the data you want to analyze on Amazon OpenSearch Service gaining valuable insights on data that previously may have been deleted or archived. With UltraWarm, you can now economically retain more of your data to interactively analyze it whenever you want.

Q. How does UltraWarm relate to/work with Amazon OpenSearch Service?

Amazon OpenSearch Service supports two integrated storage tiers, hot and UltraWarm. The hot tier is powered by data nodes which are used for indexing, updating, and providing the fastest

access to data. UltraWarm nodes complement the hot tier by providing low cost, read-only tier for older and less-frequently accessed data.

Q. Why does UltraWarm only need primary data for durability?

UltraWarm uses Amazon Simple Storage Service (Amazon S3) for storage, which is designed for 99.999999999 percent durability, and removes the need to configure an Elasticsearch replica for your warm data. Additionally, if you have more than one UltraWarm node, in the event of a node failure, the other UltraWarm nodes will automatically access the data as needed.

Q. How much data can I store in UltraWarm?

UltraWarm supports up to 3 PB of primary data. UltraWarm is designed to allow you to fully utilize 100% of this storage and because UltraWarm stores data on S3 for durability, you do not need to use additional storage for Elasticsearch replicas.

Q. What are the performance characteristics of UltraWarm?

UltraWarm delivers an interactive experience in OpenSearch Dashboards and Kibana by implementing granular I/O caching, prefetching, and query engine optimizations to provide similar performance to high-density instances using local storage.

Q. How can I start using UltraWarm?

To get started with UltraWarm, create a new Amazon OpenSearch Service domain with UltraWarm enabled via the console, CLI, or APIs. Once your domain is created you can move data from hot to UltraWarm using the OpenSearch/Elasticsearch APIs. Learn more.

Cold storage

Q. What is cold storage?

Cold storage is a fully-managed lowest cost storage tier for Amazon OpenSearch Service that makes it easy for you to securely store and analyze your historical logs on-demand. Cold storage enables you to fully detach storage from compute when they are not actively performing analysis of their data and allows you to keep your data readily available at low cost. Cold storage data is available within the Amazon OpenSearch Service domain via your UltraWarm nodes. Cold storage seamlessly integrates with OpenSearch and OpenSearch Dashboards, as well as Elasticsearch (version 7.9, 7.10) and Kibana (version 7.9, 7.10). It enables you to analyze data using the same tools that Amazon OpenSearch Service provides today.

Q. Why should I use cold storage?

Cold storage enables you to cost effectively expand the data you want to analyze on Amazon OpenSearch Service and gain valuable insights on data that previously may have been deleted or archived. Cold storage is a great fit if you have the need to do research or forensic analysis on your older data and you want to use all of the capabilities of Amazon OpenSearch Service to do so, at an affordable price. Cold storage is built for scale and is backed by Amazon S3. Find and discover the data you need, attach it to the UltraWarm nodes in your cluster, and make it available for analysis in seconds. Attached cold data is subject to the existing fine-grained access control policies that limit access at the index, document, and field level.

Q. How does cold storage relate to/work with Amazon OpenSearch Service?

With cold storage, Amazon OpenSearch Service supports three integrated storage tiers: hot, UltraWarm, and cold. The hot tier is used for indexing, updating, and providing the fastest access to data. UltraWarm provides a seamless extension of the hot tier by providing compute

nodes that provide a highly performant interactive experience for data that is durably stored in Amazon S3 and needs to be persistently available, currently supporting up to 3PB of data in a single domain. With cold storage, you can now detach indices from UltraWarm while not in use and free up compute to help lower costs. With the new cold storage APIs and OpenSearch Dashboards and Kibana interface, you can discover indices based on index patterns and data timestamps to easily find what you need for analysis. That data can then be attached to the domain and ready for analysis in seconds. When you are done with analysis, simply detaching the data then frees up your compute again.

Q. How much data can I store in cold storage?

Cold storage is built for scale. While the storage limits for hot and warm data remain at 3PB, you can store any amount of data in cold storage.

Q. What are the performance characteristics of cold storage?

Cold storage builds on UltraWarm, which provides specialized nodes that store data in Amazon S3 and uses a sophisticated caching solution to provide an interactive experience. Cold data must first be attached to the UltraWarm nodes of your Amazon OpenSearch Service domain. Once attached, queries on this data are powered by existing UltraWarm nodes offering the same performance as your warm data. Attaching cold indices to your domain takes seconds if there is sufficient UltraWarm capacity available for the requested data. If you need additional capacity, UltraWarm data nodes must be added, which can take up to a few minutes.

Cross-Cluster Search

Q: What is Cross-Cluster Search?

Cross-cluster-search is an Elasticsearch and OpenSearch feature that enables performing queries and aggregation across two connected clusters. It works by setting up a light weight uni-directional connection between participating clusters.

Q: What are the minimum requirements for a domain to participate in cross-cluster search?

Domains participating in a cross-cluster search needs to meet the following criteria:

- Participating domains should be on OpenSearch or Elasticsearch version 6.8 and above
- Participating domains need to have encryption in transit enabled
- Participating domains need to have Fine Grained Access Control (FGAC) enabled
- Participating domains versions should adhere to the same rules as rolling version upgrade

Q: What are the instance types that support cross-cluster search?

Cross-cluster search is currently supported on the following instance types

- i2, i3 family
- r3, r4, r5 family
- m4, m5 family
- c4, c5, family
- Graviton family

Q: What are the instance types that do not support cross-cluster search?

Cross-cluster search is not supported on the t2 and m3 family instances due to technical limitation.

Q: Can domains in two different AWS accounts participate in cross-cluster search?

Yes. Participating domains can belong to two different AWS accounts.

Q: Can domains in two different AWS regions participate in cross-cluster search?

No.

Q: How can I start using cross-cluster search?

To get started with cross-cluster search, follow the documentation here

Cross-Cluster replication

Q: What is Cross-cluster replication?

Cross-cluster replication, a new capability that allows Amazon OpenSearch Service customers to automate copying and synchronizing indices from one cluster to another at low latency in same or different AWS Regions.

Q: What are the minimum requirements for a domain to participate in cross-cluster replication?

Domains participating in a cross-cluster replications needs to meet the following criteria:

- Participating domains should be on Elasticsearch version 7.10
- Participating domains need to have encryption in transit enabled
- Participating domains need to have Fine Grained Access Control (FGAC) enabled
- Participating domains versions should adhere to the same rules as rolling version upgrade

Q: Can domains in two different AWS Regions participate in cross-cluster replication?

Yes. domains in two different AWS Regions can participate in cross-cluster replication.

Q: Does cross-cluster replication support Ultrawarm and Cold Storage?

No. Current implementation of cross-cluster replication does not support Ultrawarm or Cold Storage.

Q: What are the charges for cross-cluster replication?

Yes. You need to pay standard AWS data transfer charges for the data transferred in and out of Amazon OpenSearch Service.

Trace Analytics

Q: What is Trace Analytics?

Trace Analytics is a new feature of Amazon OpenSearch Service that enables developers and IT operators to find and fix performance problems in distributed applications, which leads to faster

problem resolution times. Trace Analytics is built using OpenTelemetry, a Cloud Native Computing Foundation (CNCF) project that provides a single set of APIs, libraries, agents, and collector services to capture distributed traces and metrics, which enables customers to leverage Trace Analytics without having to re-instrument their applications. Trace Analytics is powered by the OpenSearch, which is open source and freely available for everyone to download and use.

Q: Why should I use Trace Analytics?

Developers and IT Ops need Trace Analytics to find and fix performance problems in their distributed applications. By adding trace data to the existing log analytics capabilities of Amazon OpenSearch Service, customers can use the same service to both isolate the source of performance problems and diagnose their root cause. In addition, with the support for the OpenTelemetry standard, Trace Analytics supports integration with Jaeger and Zipkin SDKs, two popular open source distributed tracing systems, which allows developers to continue using these SDKs and not have to re-instrument their applications.

Q: How does Trace Analytics relate to/work with Amazon OpenSearch Service?

Trace Analytics is an integrated feature of Amazon OpenSearch Service. It is available to all customers at no extra charge. Trace Analytics has a user interface based on OpenSearch Dashboards and Kibana for visualizing and exploring trace data and is integrated with key features of Amazon OpenSearch Service such as anomaly detection, alerting, fine-grained access control, and enterprise security. Trace Analytics complements customers' usage of Amazon OpenSearch Service for search and analysis of log data when resolving application performance problems.

Q: What data sources does Trace Analytics support?

Trace Analytics today supports the collection of trace data from application libraries and SDKs that are compatible with the open source OpenTelemetry Collector, including Jaeger, Zipkin, and X-Ray SDKs. Trace Analytics also integrates with AWS Distro for OpenTelemetry, which is a distribution of OpenTelemetry APIs, SDKs, and agents/collectors. It is a performant and secure distribution of OpenTelemetry components that has been tested for production use and is supported by AWS. Customers can use AWS Distro for OpenTelemetry to collect traces and metrics for multiple monitoring solutions, including Amazon OpenSearch Service and AWS X-Ray for trace data and Amazon CloudWatch for metrics.

Q: How can I start using Trace Analytics?

To get started with Trace Analytics, follow the documentation here.

Name change

Q: Why did the name change to Amazon OpenSearch Service from Amazon Elasticsearch Service?

We announced the OpenSearch project, a community-driven, open source fork of Elasticsearch and Kibana, on April 12, 2021. We committed to making a long-term investment in OpenSearch to ensure users continue to have a secure, high-quality, fully open source search and analytics suite with a rich roadmap of new and innovative functionality. This project includes OpenSearch (derived from Elasticsearch 7.10.2) and OpenSearch Dashboards (derived from Kibana 7.10.2). We launched version 1.0 of OpenSearch on July 12, 2021. As part of our long-term commitment to OpenSearch, we added support for OpenSearch 1.0 on the managed service on September 7, 2021 and changed the name from Amazon Elasticsearch Service to Amazon OpenSearch Service. Along with OpenSearch 1.0, we continue to support legacy Elasticsearch versions until

7.10 on the service. Aside from the name change, you can rest assured that we will continue to deliver the same great experience without any impact to ongoing operations, development methodology, or business use. Learn more about OpenSearch here: https://opensearch.org/.

Q: Do I, as a customer, have to take any action as part of this name change?

We have strived to make this name change as seamless as possible for you. There are aspects, such as the new SDK/configuration APIs, that require your action to ensure you derive the best benefits from the service. While the existing SDK will continue to work from a compatibility perspective, any new functionality that requires new configuration APIs will only be implemented in the new SDK. Hence, we recommend that you move to the new SDK. In addition, irrespective of the new SDK, we strongly recommend that you move your existing IAM policies to use the renamed configuration APIs. As of now, your existing IAM policies will continue to work with the old API definition. However, we will move over to the new API-based permission validation and we will eventually require you to use the new APIs in your policies (specifically for the APIs where there is a name change; e.g. CreateElasticsearchDomain to CreateDomain). Please see the documentation for more details.

Q: Do I have to move to the new SDK to upgrade to OpenSearch 1.0?

No. From a backward compatibility perspective, we will ensure that your existing setup continues to work with OpenSearch 1.0. However, we recommend that you eventually move to the latest SDK for a cleaner and up-to-date experience, as mentioned above.

Q: Are there any changes to pricing with this name change?

No. There are no changes to pricing.

OpenSearch includes certain Apache-licensed Elasticsearch code from Elasticsearch B.V. and other source code. Elasticsearch B.V. is not the source of that other source code. ELASTICSEARCH is a registered trademark of Elasticsearch B.V.

Upgrades

Q: I'm using the Elasticsearch engine in Amazon OpenSearch Service. Why should I upgrade to the OpenSearch 1.x engine? What are the benefits for me?

Upgrading to OpenSearch 1.x ensures your search infrastructure is built on a growing and dynamic Apache-Licensed open source project and gives you access to a wealth of innovative improvements and features available in OpenSearch 1.2 (as of this writing). Features such as enterprise-grade security, alerting, data-lifecycle management, observability, ML-based anomaly detection, and more are all part of OpenSearch Service, with no additional licensing fees.

Q: Will I have downtime if I upgrade?

We use a blue/green (BG) deployment process during upgrade. During a BG, the service adds nodes to the OpenSearch Service cluster in the new configuration and version, migrates data from the old nodes, and drops the old nodes when the data migration is complete. During a BG, search and indexing APIs are available and function normally. Although BG is designed not to interfere with query and indexing requests, some changes (especially those involving changes to security-related settings) can cause dashboards to be unavailable during the period of change.

Q: Is AWS deprecating older versions of Elasticsearch Service?

AWS maintains 19 versions of Apache-2.0-licensed Elasticsearch. None of these versions are deprecated or planned for deprecation at this time.

Q: Will the upgrade trigger a BG? If not, what is the process for upgrading our nodes?

Yes, the upgrade will trigger a BG deployment process. Please review the upgrade preparation and steps here.

Q: I want to move to Amazon OpenSearch Service 1.x to take advantage of AWS Graviton2 instances, but I am locked in with my existing reserved instances (RIs). How can you help?

Please work with your AWS account team for information based on your specific situation with RIs.

Q: What should I plan for before initiating an upgrade to Amazon OpenSearch Service 1.x or greater?

The OpenSearch project 1.0 is a fork of open source Elasticsearch 7.10.2. It is wire-compatible with Elasticsearch 7.10--you don't need to change your usage. To migrate, you can upgrade your domain to version Elasticsearch 7.10 from any prior version in the 6.x and 7.x series, take a snapshot, and restore that snapshot to a domain running OpenSearch Service 1.x. Some clients or tools include version checks that may cause the client or tool to not work with OpenSearch Service. When you upgrade, enable compatibility mode to work around these version checks.

Q: Can I continue to use my existing clients, data collection, and data ingestion tools with Amazon OpenSearch Service 1.x?

In most cases, you can continue to use your existing clients. The APIs and core search functionality are compatible with Elasticsearch version 7.10.2. If you have older clients, the clients perform version checking, or the clients leverage functionality that is targeted to older versions of Elasticsearch such as major versions 5 or 6, then we suggest that you try to bring your clients up to a minimum standard of support on 7.10.2 to ensure a smooth transition.

The OpenSearch project supports a broad range of clients that are specifically built to address the OpenSearch versions of the engine on Amazon OpenSearch Service. Check your client against the list of the latest OpenSearch clients and supported programming languages for those clients.

You can enable the compatibility mode feature to interoperate with clients from other vendors, but remember to check the version that OpenSearch reports. Enable this setting to ensure the service responds with version 7.10.2 to clients developed prior to the introduction of the OpenSearch Service engine.

Q: I'm running Elasticsearch version 5.x or earlier. What's my best upgrade path?

Elasticsearch 5.x indices are not compatible with Elasticsearch 7.10 or OpenSearch 1.x. You must create a new index and load data from your source. If you are running a log analytics workload, you can evaluate whether your data retention strategy supports running in parallel while you build up a full data set on the new domain.

Q: Are there any partners that can help me with my upgrade?

Yes, please contact opensearchmigration-si-support@amazon.com to request a list of partners for your region, industry, and project complexity. AWS Partner Network (APN) partners are trained and have the experience to assist you with upgrades.

Q: Will Amazon OpenSearch Service remain compatible with Elasticsearch in the future? What's the plan for the future?

OpenSearch 1.0 is a fork of Elasticsearch 7.10.2. OpenSearch and Elasticsearch are compatible. If you enable compatibility mode, Elasticsearch clients are also compatible with OpenSearch 1.0.

Amazon OpenSearch Service does not and will not offer versions of the Elasticsearch engine after version 7.10.2.

As AWS announced when we forked Elasticsearch, we intended to build, and have built a thriving community around OpenSearch. We have published a roadmap for OpenSearch, complete with community input and consensus on the feature priorities. We will make every reasonable effort to maintain compatibility with Elasticsearch. Our goal is to grow with our community and Amazon OpenSearch Service customers.

You can directly upgrade to OpenSearch Service 1.0 from Elasticsearch and Kibana versions 6.8.0 to 7.10.2, and open distro for Elasticsearch (ODFE) 1.x. For rolling upgrades from ODFE to OpenSearch, we recommend first upgrading to ODFE 1.13, then upgrading to OpenSearch 1.0.

Amazon Relational Database Service (Amazon RDS)

Set up, operate, and scale a relational database in the cloud with just a few clicks.

Amazon RDS on VMware

Amazon Relational Database Service (RDS) on VMware is a service that delivers AWS-managed relational databases in on-premises VMware environments. RDS on VMware automates time-consuming administration tasks such as database provisioning, operating system and database patching, backups, point-in-time restore, and database instance health monitoring. It makes it easy to set up, operate, and scale databases in your on-premises VMware vSphere environments, freeing you to focus on your applications.

RDS on VMware is generally available and will support Microsoft SQL Server, MySQL, and PostgreSQL database engines in customer-owned data centers, including for workloads that

need to remain on-premises to be in compliance with security, privacy, regulatory, or data sovereignty policies.

Amazon RDS

Amazon Relational Database Service (Amazon RDS) is a web service that makes it easier to set up, operate, and scale a relational database in the cloud. It provides cost-efficient, resizeable capacity for an industry-standard relational database and manages common database administration tasks. Amazon Aurora is a fully-managed relational database engine that's built for the cloud and compatible with MySQL and PostgreSQL. Amazon Aurora is part of Amazon RDS.

How it works

Amazon Relational Database Service (Amazon RDS) is a collection of managed services that makes it simple to set up, operate, and scale databases in the cloud. Choose from seven popular engines — Amazon Aurora with MySQL compatibility, Amazon Aurora with PostgreSQL compatibility, MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server — and deploy on-premises with Amazon RDS on AWS Outposts.

Amazon RDS

Amazon RDS Custom

Amazon RDS on AWS Outposts

Use cases

Build web and mobile applications
Support growing apps with high availability, throughput, and storage scalability. Take advantage of flexible pay-per-use pricing to suit various application usage patterns.

Move to managed databases
Innovate and build new apps with Amazon RDS instead of worrying about self-managing your databases, which can be time consuming, complex, and expensive.

Break free from legacy databases
Free yourself from expensive, punitive, commercial databases by migrating to Amazon Aurora. When you migrate to Aurora, you get the scalability, performance, and availability of commercial databases at 1/10th the cost.

What is Amazon Relational Database Service (Amazon RDS)?

Amazon Relational Database Service (Amazon RDS) is a web service that makes it easier to set up, operate, and scale a relational database in the AWS Cloud. It provides cost-efficient, resizable capacity for an industry-standard relational database and manages common database administration tasks.

Note

This guide covers Amazon RDS database engines other than Amazon Aurora. For information about using Amazon Aurora, see the *Amazon Aurora User Guide*.

If you are new to AWS products and services, begin learning more with the following resources:

- For an overview of all AWS products, see What is cloud computing?
- Amazon Web Services provides a number of database services. For guidance on which service is best for your environment, see Running databases on AWS.

Amazon RDS and Amazon EC2

Amazon RDS is a managed database service. It's responsible for most management tasks. By eliminating tedious manual tasks, Amazon RDS frees you to focus on your application and your users. We recommend Amazon RDS over Amazon EC2 as your default choice for most database deployments.

Amazon RDS provides the following specific advantages over database deployments that aren't fully managed:

- You can use the database products you are already familiar with: MariaDB, Microsoft SQL Server, MySQL, Oracle, and PostgreSQL.
- Amazon RDS manages backups, software patching, automatic failure detection, and recovery.
- You can turn on automated backups, or manually create your own backup snapshots. You can use these backups to restore a database. The Amazon RDS restore process works reliably and efficiently.
- You can get high availability with a primary instance and a synchronous secondary instance that you can fail over to when problems occur. You can also use read replicas to increase read scaling.
- In addition to the security in your database package, you can help control who can access your RDS databases. To do so, you can use AWS Identity and Access Management (IAM) to define users and permissions. You can also help protect your databases by putting them in a virtual private cloud (VPC).

Amazon RDS Custom for Oracle and Microsoft SQL Server

Amazon RDS Custom is an RDS management type that gives you full access to your database and operating system.

You can use the control capabilities of RDS Custom to access and customize the database environment and operating system for legacy and packaged business

applications. Meanwhile, Amazon RDS automates database administration tasks and operations.

In this deployment model, you can install applications and change configuration settings to suit your applications. At the same time, you can offload database administration tasks such as provisioning, scaling, upgrading, and backup to AWS. You can take advantage of the database management benefits of Amazon RDS, with more control and flexibility.

For Oracle Database and Microsoft SQL Server, RDS Custom combines the automation of Amazon RDS with the flexibility of Amazon EC2. For more information on RDS Custom, see Working with Amazon RDS Custom.

With the shared responsibility model of RDS Custom, you get more control than in Amazon RDS, but also more responsibility. For more information, see Shared responsibility model.

Amazon RDS on AWS Outposts

Amazon RDS on AWS Outposts extends RDS for SQL Server, RDS for MySQL, and RDS for PostgreSQL databases to AWS Outposts environments. AWS Outposts uses the same hardware as in public AWS Regions to bring AWS services, infrastructure, and operation models on-premises. With RDS on Outposts, you can provision managed DB instances close to the business applications that must run on-premises. For more information, see Working with Amazon RDS on AWS Outposts.

DB instances

A *DB instance* is an isolated database environment in the AWS Cloud. The basic building block of Amazon RDS is the DB instance.

Your DB instance can contain one or more user-created databases. You can access your DB instance by using the same tools and applications that you use with a standalone database instance. You can create and modify a DB instance by using the AWS Command Line Interface (AWS CLI), the Amazon RDS API, or the AWS Management Console.

DB engines

A *DB engine* is the specific relational database software that runs on your DB instance. Amazon RDS currently supports the following engines:

- MariaDB
- Microsoft SQL Server
- MySQL
- Oracle
- PostgreSQL

Each DB engine has its own supported features, and each version of a DB engine can include specific features. Support for Amazon RDS features varies across AWS Regions and specific versions of each DB engine. To check feature support in different engine versions and Regions, see Supported features in Amazon RDS by AWS Region and DB engine.

Additionally, each DB engine has a set of parameters in a DB parameter group that control the behavior of the databases that it manages.

DB instance classes

A *DB instance class* determines the computation and memory capacity of a DB instance. A DB instance class consists of both the DB instance type and the size. Each instance type offers different compute, memory, and storage capabilities. For example, db.m6g is a general-purpose DB instance type powered by AWS Graviton2 processors. Within the db.m6g instance type, db.m6g.2xlarge is a DB instance class.

You can select the DB instance that best meets your needs. If your needs change over time, you can change DB instances. For information, see DB instance classes.

DB instance storage

Amazon EBS provides durable, block-level storage volumes that you can attach to a running instance. DB instance storage comes in the following types:

- General Purpose (SSD)
- Provisioned IOPS (PIOPS)
- Magnetic

The storage types differ in performance characteristics and price. You can tailor your storage performance and cost to the needs of your database.

Each DB instance has minimum and maximum storage requirements depending on the storage type and the database engine it supports. It's important to have sufficient storage so that your databases have room to grow. Also, sufficient storage makes sure that features for the DB engine have room to write content or log entries. For more information, see Amazon RDS DB instance storage.

Amazon Virtual Private Cloud (Amazon VPC)

You can run a DB instance on a virtual private cloud (VPC) using the Amazon Virtual Private Cloud (Amazon VPC) service. When you use a VPC, you have control over your virtual networking environment. You can choose your own IP address range, create subnets, and configure routing and access control lists. The basic functionality of Amazon RDS is the same whether it's running in a VPC or not. Amazon RDS manages backups, software patching, automatic failure detection, and recovery. There's no additional cost to run your DB instance in a VPC. For more information on using Amazon VPC with RDS, see Amazon VPC VPCs and Amazon RDS.

Amazon RDS uses Network Time Protocol (NTP) to synchronize the time on DB Instances.

Security

A *security group* controls the access to a DB instance. It does so by allowing access to IP address ranges or Amazon EC2 instances that you specify.

For more information about security groups, see Security in Amazon RDS.

Monitoring an Amazon RDS DB instance

There are several ways that you can track the performance and health of a DB instance. You can use the Amazon CloudWatch service to monitor the performance and health of a DB instance. CloudWatch performance charts are shown in the Amazon RDS console. You can also subscribe to Amazon RDS events to be notified about changes to a DB instance, DB snapshot, or DB parameter group. For more information, see Monitoring metrics in an Amazon RDS instance.

How to work with Amazon RDS

There are several ways that you can interact with Amazon RDS.

AWS Management Console

The AWS Management Console is a simple web-based user interface. You can manage your DB instances from the console with no programming required. To access the Amazon RDS console, sign in to the AWS Management Console and open the Amazon RDS console at https://console.aws.amazon.com/rds/.

Command line interface

You can use the AWS Command Line Interface (AWS CLI) to access the Amazon RDS API interactively. To install the AWS CLI, see Installing the AWS Command Line Interface. To begin using the AWS CLI for RDS, see AWS Command Line Interface reference for Amazon RDS.

Programming with Amazon RDS

If you are a developer, you can access the Amazon RDS programmatically. For more information, see Amazon RDS API reference.

For application development, we recommend that you use one of the AWS Software Development Kits (SDKs). The AWS SDKs handle low-level details such as authentication, retry logic, and error handling, so that you can focus on your application logic. AWS SDKs are available for a wide variety of languages. For more information, see Tools for Amazon web services .

AWS also provides libraries, sample code, tutorials, and other resources to help you get started more easily. For more information, see Sample code & libraries.

Amazon Relational Database Service (Amazon RDS) FAQs

General

What is Amazon RDS?

Amazon Relational Database Service (Amazon RDS) is a managed service that makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity, while managing time-consuming database administration tasks, freeing you to focus on your applications and business.

Amazon RDS gives you access to the capabilities of a familiar MySQL, MariaDB, Oracle, SQL Server, or PostgreSQL database. This means that the code, applications, and tools you already use today with your existing databases should work seamlessly with Amazon RDS. Amazon RDS can automatically back up your database and keep your database software up to date with the latest version. You benefit from the flexibility of being able to easily scale the compute

resources or storage capacity associated with your relational database instance. In addition, Amazon RDS makes it easy to use replication to enhance database availability, improve data durability, or scale beyond the capacity constraints of a single database instance for read-heavy database workloads. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use.

When would I use Amazon RDS vs. Amazon EC2 Relational Database AMIs?

Amazon Web Services provides a number of database alternatives for developers. Amazon RDS enables you to run a fully managed and fully featured relational database while offloading database administration. Using one of our many relational database AMIs on Amazon EC2 allows you to manage your own relational database in the cloud. There are important differences between these alternatives that may make one more appropriate for your use case. See Cloud Databases with AWS for guidance on which solution is best for you.

Are there hybrid or on-premises deployment options for Amazon RDS?

Yes, you can run Amazon RDS on premises using Amazon RDS on Outposts. Please see the Amazon RDS on Outposts FAQs for additional information.

Can I get help to learn more about and onboard to Amazon RDS?

Yes, Amazon RDS specialists are available to answer questions and provide support. Contact Us and you'll hear back from us in one business day to discuss how AWS can help your organization.

How do I set up a connection between an application or a SQL based client running on an Amazon EC2 compute instance and my Amazon RDS database instance/cluster?

You can set up a connection between an EC2 compute instance and a new Amazon RDS database using the Amazon RDS console. On the "Create database" page, select "Connect to an EC2 compute resource" option in the Connectivity Section. When you select this option, Amazon RDS automates the manual networking set up tasks such as creating a VPC, security groups, subnets, and ingress/egress rules to establish a connection between your application and database.

Additionally, you can set up a connection between an existing Amazon RDS database and an EC2 compute instance. To do so, open the RDS console, select an RDS database from the database list page, and choose "Set up EC2 connection" from the "Action" menu dropdown list. Amazon RDS automatically sets up your related network settings to enable a secure connection between the selected EC2 instance and the RDS database.

This connectivity automation improves productivity for new users and application developers. Users can now quickly and seamlessly connect an application or a client using SQL on an EC2 compute instance to an RDS database within minutes.

Database Instances

What is a database instance (DB instance)?

You can think of a DB instance as a database environment in the cloud with the compute and storage resources you specify. You can create and delete DB instances, define/refine infrastructure attributes of your DB instance(s), and control access and security via the AWS Management Console, Amazon RDS APIs, and AWS Command Line Interface. You can run one

or more DB instances and each DB instance can support one or more databases or database schemas, depending on engine type.

How do I create a DB instance?

DB instances are simple to create using either the AWS Management Console, Amazon RDS APIs, or AWS Command Line Interface. To launch a DB instance using the AWS Management Console, click "RDS" and then the “Launch DB Instance” button on the Instances tab. From there, you can specify the parameters for your DB instance, including DB engine and version, license model, instance type, storage type and amount, and primary user credentials.

You also have the ability to change your DB instance’s backup retention policy, preferred backup window, and scheduled maintenance window. Alternatively, you can create your DB instance using the CreateDBInstance API or create-db-instance command.

How do I access my running DB instance?

Once your DB instance is available, you can retrieve its endpoint via the DB instance description in the AWS Management Console, DescribeDBInstances API or describe-db-instances command. Using this endpoint, you can construct the connection string required to connect directly with your DB instance using your favorite database tool or programming language. In order to allow network requests to your running DB instance, you will need to authorize access. For a detailed explanation of how to construct your connection string and get started, please refer to our Getting Started Guide.

How many DB instances can I run with Amazon RDS?

By default, customers are allowed to have up to a total of 40 Amazon RDS DB instances. Of those 40, up to 10 can be Oracle or SQL Server DB instances under the "License Included" model. All 40 can be used for Amazon Aurora, MySQL, MariaDB, PostgreSQL, and Oracle under the "BYOL" model. Note that RDS for SQL Server has a limit of up to 100 databases on a single DB instance to learn more see the Amazon RDS for SQL Server User Guide.

If your application requires more DB instances, you can request additional DB instances via this request form.

How many databases or schemas can I run within a DB instance?

- RDS for Amazon Aurora: No limit imposed by software
- RDS for MySQL: No limit imposed by software
- RDS for MariaDB: No limit imposed by software
- RDS for Oracle: 1 database per instance; no limit on the number of schemas per database imposed by software
- RDS for SQL Server: Up to 100 databases per instance
- RDS for PostgreSQL: No limit imposed by software

How do I import data into an Amazon RDS DB instance?

There are a number of simple ways to import data into Amazon RDS, such as with the mysqldump or mysqlimport utilities for MySQL; Data Pump, import/export, or SQL Loader for Oracle; Import/Export wizard, full backup files (.bak files), or Bulk Copy Program (BCP) for SQL Server; or pg_dump for PostgreSQL. For more information on data import and

export, please refer to the Data Import Guide for MySQL, the Data Import Guide for Oracle, the Data Import Guide for SQL Server, or the Data Import Guide for PostgreSQL.

In addition, AWS Database Migration Service can help you migrate databases to AWS easily and securely.

What is a maintenance window? Will my DB instance be available during maintenance events?

The Amazon RDS maintenance window is your opportunity to control when DB instance modifications, database engine version upgrades, and software patching occurs, in the event they are requested or required. If a maintenance event is scheduled for a given week, it will be initiated during the maintenance window you identify.

Maintenance events that require Amazon RDS to take your DB instance offline are scale compute operations (which generally take only a few minutes from start-to-finish), database engine version upgrades, and required software patching. Required software patching is automatically scheduled only for patches that are security and durability related. Such patching occurs infrequently (typically once every few months) and should seldom require more than a fraction of your maintenance window.

If you do not specify a preferred weekly maintenance window when creating your DB instance, a 30-minute default value is assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your DB instance in the AWS Management Console, the ModifyDBInstance API, or the modify-db-instance command. Each of your DB instances can have different preferred maintenance windows, if you so choose.

Running your DB instance as a Multi-AZ deployment can further reduce the impact of a maintenance event. Please refer to the Amazon RDS User Guide for more information on maintenance operations.

What should I do if my queries seem to be running slowly?

For production databases, we encourage you to enable Enhanced Monitoring, which provides access to over 50 CPU, memory, file system, and disk I/O metrics. You can enable these features on a per-instance basis and you can choose the granularity (all the way down to 1 second). High levels of CPU utilization can reduce query performance and in this case, you may want to consider scaling your DB instance class. For more information on monitoring your DB instance, refer to the Amazon RDS User Guide.

If you are using RDS for MySQL or MariaDB, you can access the slow query logs for your database to determine if there are slow-running SQL queries and, if so, the performance characteristics of each. You could set the "slow_query_log" DB Parameter and query the mysql.slow_log table to review the slow-running SQL queries. Please refer to the Amazon RDS User Guide to learn more.

If you are using RDS for Oracle, you can use the Oracle trace file data to identify slow queries. For more information on accessing trace file data, please refer to Amazon RDS User Guide.

If you are using RDS for SQL Server, you can use the client side SQL Server traces to identify slow queries. For information on accessing server side trace file data, please refer to Amazon RDS User Guide.

Database Engine Versions

Which relational database engine versions does Amazon RDS support?

For the list of supported database engine versions, please refer to the documentation for each engine:

- Amazon RDS for MySQL
- Amazon RDS for MariaDB
- Amazon RDS for PostgreSQL
- Amazon RDS for Oracle
- Amazon RDS for SQL Server
- Amazon Aurora

How does Amazon RDS distinguish between “major” and “minor” DB engine versions?

Refer to the FAQs page for each Amazon RDS database engine for specifics on version numbering:

- Amazon RDS for MySQL
- Amazon RDS for MariaDB
- Amazon RDS for PostgreSQL
- Amazon RDS for Oracle
- Amazon RDS for SQL Server
- Amazon Aurora

Does Amazon RDS provide guidelines for support of new DB engine versions?

Over time, Amazon RDS adds support for new major and minor database engine versions. The number of new versions supported will vary based on the frequency and content of releases and patches from the engine’s vendor or development organization and the outcome of a thorough vetting of these releases and patches by our database engineering team. However, as a general guidance, we aim to support new engine versions within 5 months of their general availability.

How do I specify which supported DB engine version I would like my DB instance to run?

You can specify any currently supported version (major and minor) when creating a new DB instance via the Launch DB Instance operation in the AWS Management Console or the CreateDBInstance API. Please note that not every database engine version is available in every AWS region.

How do I control if and when the engine version of my DB instance is upgraded to new supported versions?

Amazon RDS strives to keep your database instance up to date by providing you with newer versions of the supported database engines. After a new version of a database engine is released by the vendor or development organization, it is thoroughly tested by our database engineering team before it is made available in Amazon RDS.

We recommend that you keep your database instance upgraded to the most current minor version as it will contain the latest security and functionality fixes. Unlike major version upgrades, minor version upgrades only include database changes that are backward-compatible with previous minor versions (of the same major version) of the database engine.

If a new minor version does not contain fixes that would benefit Amazon RDS customers, we may choose not to make it available in Amazon RDS. Soon after a new minor version is available in Amazon RDS, we will set it to be the preferred minor version for new DB instances.

To manually upgrade a database instance to a supported engine version, use the Modify DB Instance command on the AWS Management Console or the ModifyDBInstance API and set the DB Engine Version parameter to the desired version. By default, the upgrade will be applied during your next maintenance window. You can also choose to upgrade immediately by selecting the Apply Immediately option in the console API.

If we determine that a new engine minor version contains significant bug fixes compared to a previously released minor version, we will schedule automatic upgrades for DB instances that have the Auto Minor Version Upgrade setting to “Yes”. These upgrades will be scheduled to occur during customer-specified maintenance windows.

We schedule them so you can plan around them because downtime is required to upgrade a DB engine version, even for Multi-AZ instances. If you wish to turn off automatic minor version upgrades you can do so by setting the Auto Minor Version Upgrade setting to “No”.

In the case of RDS for Oracle and RDS for SQL Server, if the upgrade to the next minor version requires a change to a different edition, then we may not schedule automatic upgrades even if you have enabled the Auto Minor Version Upgrade setting. The determination on whether to schedule automatic upgrades in such situations will be made on a case-by-case basis.

Since major version upgrades involve some compatibility risk, they will not occur automatically and must be initiated by you (except in the case of major version deprecation, see below). For more information about upgrading a DB instance to a new DB engine version, refer to the Amazon RDS User Guide.

Can I test my DB instance with a new version before upgrading?

Yes. You can do so by creating a DB snapshot of your existing DB instance, restoring from the DB snapshot to create a new DB instance, and then initiating a version upgrade for the new DB instance. You can then experiment safely on the upgraded copy of your DB instance before deciding whether or not to upgrade your original DB instance.

For more information about restoring a DB snapshot, refer to the Amazon RDS User Guide.

Does Amazon RDS provide guidelines for deprecating database engine versions that are currently supported?

- We intend to support major version releases (e.g., MySQL 5.6, PostgreSQL 9.6) for at least 3 years after they are initially supported by Amazon RDS.
- We intend to support minor versions (e.g., MySQL 5.6.37, PostgreSQL 9.6.1) for at least 1 year after they are initially supported by Amazon RDS.

Periodically, we will deprecate major or minor engine versions. Major versions are made available at least until the community end of life for the corresponding community version or the version is no longer receiving software fixes or security updates. For minor versions, this is

when a minor version has significant bugs or security issues that have been resolved in a later minor version.

While we strive to meet these guidelines, in some cases we may deprecate specific major or minor versions sooner, such as when there are security issues. In the unlikely event that such cases occur, Amazon RDS will automatically upgrade your database engine to address the issue. Specific circumstances may dictate different timelines depending on the issue being addressed.

What happens when an Amazon RDS DB engine version is deprecated?

When a minor version of a database engine is deprecated in Amazon RDS, we will provide a three (3) month period after the announcement before beginning automatic upgrades. At the end of this period, all instances still running the deprecated minor version will be scheduled for automatic upgrade to the latest supported minor version during their scheduled maintenance windows.

When a major version of the database engine is deprecated in Amazon RDS, we will provide a minimum six (6) month period after the announcement of a deprecation for you to initiate an upgrade to a supported major version. At the end of this period, an automatic upgrade to the next major version will be applied to any instances still running the deprecated version during their scheduled maintenance windows.

Once a major or minor database engine version is deprecated in Amazon RDS, any DB instance restored from a DB snapshot created with the unsupported version will automatically and immediately be upgraded to a currently supported version.

Why can I not create a particular version?

In some cases, we may deprecate specific major or minor versions without prior notice, such as when we discover a version does not meet our high quality, performance, or security bar. In the unlikely event that such cases occur, Amazon RDS will discontinue the creation of new database instances and clusters with these versions. Existing customers may continue to be able to run their databases. Specific circumstances may dictate different timelines depending on the issue being addressed.

Free Tier

What does the AWS Free Tier for Amazon RDS offer?

The AWS Free Tier for Amazon RDS offer provides free use of Single-AZ Micro DB instances running MySQL, MariaDB, PostgreSQL, Oracle ("Bring-Your-Own-License (BYOL)" licensing model), and SQL Server Express Edition. The free usage tier is capped at 750 instance hours per month. Customers also receive 20 GB of General Purpose (SSD) database storage and 20 GB of backup storage for free per month.

For what time period will the AWS Free Tier for Amazon RDS be available to me?

New AWS accounts receive 12 months of AWS Free Tier access. Please see the AWS Free Tier FAQs for more information.

Can I run more than one DB instance under the AWS Free Usage Tier for Amazon RDS?

Yes. You can run more than one Single-AZ Micro DB instance simultaneously and be eligible for usage counted under the AWS Free Tier for Amazon RDS. However, any use exceeding 750 instance hours, across all Amazon RDS Single-AZ Micro DB instances and across all eligible database engines and regions, will be billed at standard Amazon RDS prices.

For example, if you run two Single-AZ Micro DB instances for 400 hours each in a single month, you will accumulate 800 instance hours of usage, of which 750 hours will be free. You will be billed for the remaining 50 hours at the standard Amazon RDS price.

Do I have access to 750 instance hours each of the MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server Micro DB instances under the AWS Free Tier?

No. A customer with access to the AWS Free Tier can use up to 750 instance hours of Micro instances running either MySQL, PostgreSQL, Oracle, or SQL Server Express Edition. Any use exceeding 750 instance hours, across all Amazon RDS Single-AZ Micro DB instances and across all eligible database engines and regions, will be billed at standard Amazon RDS prices.

How am I billed when my instance-hour usage exceeds the Free Tier benefit?

You are billed at standard Amazon RDS prices for instance hours beyond what the Free Tier provides.

Reserved Instances

What is a reserved instance (RI)?

Amazon RDS reserved instances give you the option to reserve a DB instance for a one or three year term and in turn receive a significant discount compared to the on-demand instance pricing for the DB instance. There are three RI payment options -- No Upfront, Partial Upfront, All Upfront -- which enable you to balance the amount you pay upfront with your effective hourly price.

How are reserved instances different from on-demand DB instances?

Functionally, reserved instances and on-demand DB instances are exactly the same. The only difference is how your DB instance(s) are billed. With Reserved Instances, you purchase a one- or three-year reservation and in return receive a lower effective hourly usage rate (compared with on-demand DB instances) for the duration of the term. Unless you purchase reserved instances in a Region, all DB instances will be billed at on-demand hourly rates.

How do I purchase and create reserved instances?

You can purchase a reserved instance in the "Reserved Instance" section of the AWS Management Console for Amazon RDS. Alternatively, you can use the Amazon RDS API or AWS Command Line Interface to list the reservations available for purchase and then purchase a DB instance reservation.

Once you have made a reserved purchase, using a reserved DB instance is no different than an On-Demand DB instance. Launch a DB instance using the same instance class, engine, and region for which you made the reservation. As long as your reservation purchase is active, Amazon RDS will apply the reduced hourly rate for which you are eligible to the new DB instance.

Do reserved instances include a capacity reservation?

Amazon RDS reserved instances are purchased for a Region rather than for a specific Availability Zone. As RIs are not specific to an Availability Zone, they are not capacity reservations. This means that even if capacity is limited in one Availability Zone, reservations can still be purchased in the Region and the discount will apply to matching usage in any Availability Zone within that Region.

How many reserved instances can I purchase?

You can purchase up to 40 reserved DB instances. If you wish to run more than 40 DB instances, please complete the Amazon RDS DB Instance request form.

What if I have an existing DB instance that I'd like to cover with a reserved instance?

Simply purchase a DB instance reservation with the same DB instance class, DB engine, Multi-AZ option, and License Model within the same Region as the DB instance you are currently running and would like to reserve. If the reservation purchase is successful, Amazon RDS will automatically apply your new hourly usage charge to your existing DB instance.

If I sign up for a reserved instance, when does the term begin? What happens to my DB instance when the term ends?

Pricing changes associated with a reserved instance are activated once your request is received while the payment authorization is processed. You can follow the status of your reservation on the AWS Account Activity page or by using the DescribeReservedDBInstances API or describe-reserved-db-instances command.If the one-time payment cannot be successfully authorized by the next billing period, the discounted price will not take effect.

When your reservation term expires, your reserved instance will revert to the appropriate On-Demand hourly usage rate for your DB instance class and Region.

How do I control which DB instances are billed at the reserved instance rate?

The Amazon RDS operations for creating, modifying, and deleting DB instances do not distinguish between on-demand and reserved instances. When computing your bill, our system will automatically apply your Reservation(s) such that all eligible DB instances are charged at the lower hourly reserved DB instance rate.

If I scale my DB instance class up or down, what happens to my reservation?

Each reservation is associated with the following set of attributes: DB engine, DB instance class, Multi-AZ deployment option, license model, and Region.

A reservation for a DB engine and license model that is eligible for size-flexibility (MySQL, MariaDB, PostgreSQL, Amazon Aurora, or Oracle "Bring Your Own License") will automatically apply to a running DB instance of any size within the same instance family (e.g. M4, T2, or R3) for the same database engine and Region. In addition, the reservation will also apply to DB instances running in either Single-AZ or Multi-AZ deployment options.

For example, let's say you purchased a db.m4.2xlarge MySQL reservation. If you decide to scale up the running DB instance to a db.m4.4xlarge, the discounted rate of this RI will cover 1/2 of the usage of the larger DB instance.

If you are running a DB engine or license model that is not eligible for size-flexibility (Microsoft SQL Server or Oracle "License Included"), each reservation can only be applied to a DB

instance with the same attributes for the duration of the term. If you decide to modify any of these attributes of your running DB instance before the end of the reservation term, your hourly usage rates for that DB instance will revert to on demand hourly rates.

For more details on about size flexibility, see the Amazon RDS User Guide.

Can I move a reserved instance from one Region or Availability Zone to another?

Each reserved instance is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each reservation can, however, be used in any of the available AZs within the associated Region.

Are reserved instances available for Multi-AZ deployments?

Yes. When you purchase a reserved instance, you can select the Multi-AZ option in the DB instance configuration available for purchase. In addition, if you are using a DB engine and license model that supports reserved instance size-flexibility, a Multi-AZ reserved instance will cover usage for two Single-AZ DB instances.

Are reserved instances available for read replicas?

A DB instance reservation can be applied to a read replica, provided the DB instance class and Region are the same. When computing your bill, our system will automatically apply your Reservation(s), such that all eligible DB instances are charged at the lower hourly reserved instance rate.

Can I cancel a reservation?

No, you cannot cancel your reserved DB instance and the one-time payment (if applicable) is not refundable. You will continue to pay for every hour during your Reserved DB instance term regardless of your usage.

How do the payment options impact my bill?

When you purchase an RI under the All Upfront payment option, you pay for the entire term of the RI in one upfront payment. You can choose to pay nothing upfront by choosing the No Upfront option. The entire value of the No Upfront RI is spread across every hour in the term and you will be billed for every hour in the term, regardless of usage. The Partial Upfront payment option is a hybrid of the All Upfront and No Upfront options. You make a small upfront payment, and you are billed a low hourly rate for every hour in the term regardless of usage.

Hardware and Scaling

How do I determine which initial DB instance class and storage capacity are appropriate for my needs?

In order to select your initial DB instance class and storage capacity, you will want to assess your application's compute, memory, and storage needs. For information about the DB instance classes available, please refer to the Amazon RDS User Guide.

How do I scale the compute resources and/or storage capacity associated with my Amazon RDS Database Instance?

You can scale the compute resources and storage capacity allocated to your DB instance with the AWS Management Console (selecting the desired DB instance and clicking the Modify button), the Amazon RDS API, or the AWS Command Line Interface. Memory and CPU resources are modified by changing your DB Instance class, and storage available is changed when you modify your storage allocation.

Please note that when you modify your DB Instance class or allocated storage, your requested changes will be applied during your specified maintenance window. Alternately, you can use the "apply-immediately" flag to apply your scaling requests immediately. Bear in mind that any other pending system changes will be applied as well.

Some older RDS for SQL Server instances may not be eligible for scaled storage. See the RDS for SQL Server FAQ for more information.

What is the hardware configuration for Amazon RDS storage?

Amazon RDS uses EBS volumes for database and log storage. Depending on the size of storage requested, Amazon RDS automatically stripes across multiple EBS volumes to enhance IOPS performance. For MySQL and Oracle, for an existing DB instance, you may observe some I/O capacity improvement if you scale up your storage. You can scale the storage capacity allocated to your DB Instance using the AWS Management Console, the ModifyDBInstance API, or the modify-db-instance command.

For more information, see Storage for Amazon RDS.

Will my DB instance remain available during scaling?

The storage capacity allocated to your DB Instance can be increased while maintaining DB Instance availability. However, when you decide to scale the compute resources available to your DB instance up or down, your database will be temporarily unavailable while the DB instance class is modified. This period of unavailability typically lasts only a few minutes, and will occur during the maintenance window for your DB Instance, unless you specify that the modification should be applied immediately.

How can I scale my DB instance beyond the largest DB instance class and maximum storage capacity?

Amazon RDS supports a variety of DB instance classes and storage allocations to meet different application needs. If your application requires more compute resources than the largest DB instance class or more storage than the maximum allocation, you can implement partitioning, thereby spreading your data across multiple DB instances.

What is Amazon RDS General Purpose (SSD) storage?

Amazon RDS General Purpose (SSD) Storage is suitable for a broad range of database workloads that have moderate I/O requirements. With the baseline of 3 IOPS/GB and ability to burst up to 3,000 IOPS, this storage option provides predictable performance to meet the needs of most applications.

What is Amazon RDS Provisioned IOPS (SSD) storage?

Amazon RDS Provisioned IOPS (SSD) Storage is an SSD-backed storage option designed to deliver fast, predictable, and consistent I/O performance. With Amazon RDS Provisioned IOPS (SSD) Storage, you specify an IOPS rate when creating a DB instance, and Amazon RDS provisions that IOPS rate for the lifetime of the DB instance. Amazon RDS Provisioned IOPS

(SSD) Storage is optimized for I/O-intensive, transactional (OLTP) database workloads. For more details, please see the Amazon RDS User Guide.

What is Amazon RDS magnetic storage?

Amazon RDS magnetic storage is useful for small database workloads where data is accessed less frequently. Magnetic storage is not recommended for production database instances.

How do I choose among the Amazon RDS storage types?

Choose the storage type most suited for your workload.

- High-performance OLTP workloads: Amazon RDS Provisioned IOPS (SSD) Storage
- Database workloads with moderate I/O requirements: Amazon RDS General Purpose (SSD) Storage

What are the minimum and maximum IOPS supported by Amazon RDS?

The IOPS supported by Amazon RDS varies by database engine. For more details, please see the Amazon RDS User Guide.

Automatic Backups and Database Snapshots

What is the difference between automated backups and DB Snapshots?

Amazon RDS provides two different methods for backing up and restoring your DB instance(s) automated backups and database snapshots (DB Snapshots).

The automated backup feature of Amazon RDS enables point-in-time recovery of your DB instance. When automated backups are turned on for your DB Instance, Amazon RDS automatically performs a full daily snapshot of your data (during your preferred backup window) and captures transaction logs (as updates to your DB Instance are made). When you initiate a point-in-time recovery, transaction logs are applied to the most appropriate daily backup in order to restore your DB instance to the specific time you requested.

Amazon RDS retains backups of a DB Instance for a limited, user-specified period of time called the retention period, which by default is 7 days but can be set to up to 35 days. You can initiate a point-in-time restore and specify any second during your retention period, up to the Latest Restorable Time. You can use the DescribeDBInstances API to return the latest restorable time for you DB instance, which is typically within the last five minutes.

Alternatively, you can find the Latest Restorable Time for a DB instance by selecting it in the AWS Management Console and looking in the “Description” tab in the lower panel of the Console.

DB Snapshots are user-initiated and enable you to back up your DB instance in a known state as frequently as you wish, and then restore to that specific state at any time. DB Snapshots can be created with the AWS Management Console, CreateDBSnapshot API, or create-db-snapshot command and are kept until you explicitly delete them.

The snapshots which Amazon RDS performs for enabling automated backups are available to you for copying (using the AWS console or the copy-db-snapshot command) or for the snapshot restore functionality. You can identify them using the "automated" Snapshot Type. In addition, you can identify the time at which the snapshot has been taken by viewing the "Snapshot Created Time" field.

Alternatively, the identifier of the "automated" snapshots also contains the time (in UTC) at which the snapshot has been taken.

Please note: When you perform a restore operation to a point in time or from a DB Snapshot, a new DB Instance is created with a new endpoint (the old DB Instance can be deleted if so desired). This is done to enable you to create multiple DB Instances from a specific DB Snapshot or point in time.

Do I need to enable backups for my DB Instance or is it done automatically?

By default, Amazon RDS enables automated backups of your DB Instance with a 7 day retention period. Free backup storage is limited to the size of your provisioned database and only applies to active DB Instances. For example, if you have 100 GB of provisioned database storage over the month, we will provide 100 GB-months of backup storage at no additional charge.

If you would like to modify your backup retention period, you can do so using the console or the CreateDBInstance API (when creating a new DB Instance) or the ModifyDBInstance API (for existing instances). You can use these APIs to change the RetentionPeriod parameter to any number from 0 (which will disable automated backups) to the desired number of days. The value can't be set to 0 if the DB instance is a source to Read Replicas. For more information on automated backups, please refer to the Amazon RDS User Guide.

What is a backup window and why do I need it? Is my database available during the backup window?

The preferred backup window is the user-defined period of time during which your DB Instance is backed up. Amazon RDS uses these periodic data backups in conjunction with your transaction logs to enable you to restore your DB Instance to any second during your retention period, up to the LatestRestorableTime (typically up to the last few minutes). During the backup window, storage I/O may be briefly suspended while the backup process initializes (typically under a few seconds) and you may experience a brief period of elevated latency. There is no I/O suspension for Multi-AZ DB deployments, since the backup is taken from the standby.

Where are my automated backups and DB snapshots stored and how do I manage their retention?

Amazon RDS DB snapshots and automated backups are stored in S3.

You can use the AWS Management Console, the ModifyDBInstance API, or the modify-db-instance command to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended). You can manage your user-created DB Snapshots via the "Snapshots" section of the Amazon RDS Console. Alternatively, you can see a list of the user-created DB Snapshots for a given DB Instance using the DescribeDBSnapshots API or describe-db-snapshots command Dand delete snapshots with the DeleteDBSnapshot API or delete-db-snapshot command.

Why do I have more automated DB snapshots than the number of days in the retention period for my DB instance?

It is normal to have 1 or 2 more automated DB snapshots than the number of days in your retention period. One extra automated snapshot is retained to ensure the ability to perform a point in time restore to any time during the retention period.

For example, if your backup window is set to 1 day, you will require 2 automated snapshots to support restores to any within the previous 24 hours. You may also see an additional automated snapshot as a new automated snapshot is always created before the oldest automated snapshot is deleted.

What happens to my backups and DB snapshots if I delete my DB instance?

When you delete a DB instance, you can create a final DB snapshot upon deletion; if you do, you can use this DB snapshot to restore the deleted DB instance at a later date. Amazon RDS retains this final user-created DB snapshot along with all other manually created DB snapshots after the DB instance is deleted.

Automated backups are deleted when the DB instance is deleted. Only manually created DB Snapshots are retained after the DB Instance is deleted.

Security

What is Amazon Virtual Private Cloud (VPC) and how does it work with Amazon RDS?

Amazon VPC lets you create a virtual networking environment in a private, isolated section of the AWS cloud where you can exercise complete control over aspects, such as private IP address ranges, subnets, routing tables, and network gateways. With Amazon VPC, you can define a virtual network topology and customize the network configuration to closely resemble a traditional IP network that you might operate in your own data center.

One way that you can take advantage of VPC is when you want to run a public-facing web application while still maintaining non-publicly accessible backend servers in a private subnet. You can create a public-facing subnet for your webservers that has access to the Internet, and place your backend Amazon RDS DB Instances in a private-facing subnet with no Internet access. For more information about Amazon VPC, refer to the Amazon Virtual Private Cloud User Guide.

How is using Amazon RDS inside a VPC different from using it on the EC2-Classic platform (non-VPC)?

If your AWS account was created before 2013-12-04, you may be able to run Amazon RDS in an Amazon Elastic Compute Cloud (EC2)-Classic environment. The basic functionality of Amazon RDS is the same regardless of whether EC2-Classic or EC2-VPC is used. Amazon RDS manages backups, software patching, automatic failure detection, read replicas, and recovery whether your DB Instances are deployed inside or outside a VPC. For more information about the differences between EC2-Classic and EC2-VPC, see the EC2 documentation.

What is a DB Subnet Group and why do I need one?

A DB Subnet Group is a collection of subnets that you may want to designate for your Amazon RDS DB Instances in a VPC. Each DB Subnet Group should have at least one subnet for every Availability Zone in a given Region. When creating a DB Instance in VPC, you will need to select a DB Subnet Group. Amazon RDS then uses that DB Subnet Group and your preferred Availability Zone to select a subnet and an IP address within that subnet. Amazon RDS creates and associates an Elastic Network Interface to your DB Instance with that IP address.

Please note that we strongly recommend you use the DNS Name to connect to your DB Instance as the underlying IP address can change (e.g., during failover).

For Multi-AZ deployments, defining a subnet for all Availability Zones in a Region will allow Amazon RDS to create a new standby in another Availability Zone should the need arise. You need to do this even for Single-AZ deployments, just in case you want to convert them to Multi-AZ deployments at some point.

How do I create an Amazon RDS DB Instance in VPC?

For a procedure that walks you through this process, refer to Creating a DB Instance in a VPC in the Amazon RDS User Guide.

How do I control network access to my DB Instance(s)?

Visit the Security Groups section of the Amazon RDS User Guide to learn about the different ways to control access to your DB Instances.

How do I connect to an Amazon RDS DB Instance in VPC?

DB Instances deployed within a VPC can be accessed by EC2 Instances deployed in the same VPC. If these EC2 Instances are deployed in a public subnet with associated Elastic IPs, you can access the EC2 Instances via the internet. DB Instances deployed within a VPC can be accessed from the Internet or from EC2 Instances outside the VPC via VPN or bastion hosts that you can launch in your public subnet or using Amazon RDS's Publicly Accessible option:

- To use a bastion host, you will need to set up a public subnet with an EC2 instance that acts as a SSH Bastion. This public subnet must have an internet gateway and routing rules that allow traffic to be directed via the SSH host, which must then forward requests to the private IP address of your Amazon RDS DB instance.
- To use public connectivity, simply create your DB Instances with the Publicly Accessible option set to yes. With Publicly Accessible active, your DB Instances within a VPC will be fully accessible outside your VPC by default. This means you do not need to configure a VPN or bastion host to allow access to your instances.

You can also set up a VPN Gateway that extends your corporate network into your VPC and allows access to the Amazon RDS DB instance in that VPC. Refer to the Amazon VPC User Guide for more details.

We strongly recommend you use the DNS Name to connect to your DB Instance as the underlying IP address can change (e.g., during failover).

Can I move my existing DB instances outside VPC into my VPC?

If your DB instance is not in a VPC, you can use the AWS Management Console to easily move your DB instance into a VPC. See the Amazon RDS User Guide for more details. You can also take a snapshot of your DB Instance outside VPC and restore it to VPC by specifying the DB Subnet Group you want to use. Alternatively, you can perform a “Restore to Point in Time” operation as well.

Can I move my existing DB instances from inside VPC to outside VPC?

Migration of DB Instances from inside to outside VPC is not supported. For security reasons, a DB Snapshot of a DB Instance inside VPC cannot be restored to outside VPC. The same is true with “Restore to Point in Time” functionality.

What precautions should I take to ensure that my DB Instances in VPC are accessible by my application?

You are responsible for modifying routing tables and networking ACLs in your VPC to ensure that your DB instance is reachable from your client instances in the VPC. For Multi-AZ deployments, after failover, your client EC2 instance and Amazon RDS DB Instance may be in different Availability Zones. You should configure your networking ACLs to ensure that cross-AZ communication is possible.

Can I change the DB Subnet Group of my DB Instance?

An existing DB Subnet Group can be updated to add more subnets, either for existing Availability Zones or for new Availability Zones added since the creation of the DB Instance. Removing subnets from an existing DB Subnet Group can cause unavailability for instances if they are running in a particular AZ that gets removed from the subnet group. View the Amazon RDS User Guide for more information.

What is an Amazon RDS primary user account and how is it different from an AWS account?

To begin using Amazon RDS you will need an AWS developer account. If you do not have one prior to signing up for Amazon RDS, you will be prompted to create one when you begin the sign-up process. A primary user account is different from an AWS developer account and used only within the context of Amazon RDS to control access to your DB Instance(s). The primary user account is a native database user account that you can use to connect to your DB Instance.

You can specify the primary user name and password you want associated with each DB Instance when you create the DB Instance. Once you have created your DB Instance, you can connect to the database using the primary user credentials. Subsequently, you may also want to create additional user accounts so that you can restrict who can access your DB Instance.

What privileges are granted to the primary user for my DB Instance?

For MySQL, the default privileges for the primary user include: create, drop, references, event, alter, delete, index, insert, select, update, create temporary tables, lock tables, trigger, create view, show view, alter routine, create routine, execute, trigger, create user, process, show databases, grant option.

For Oracle, the primary user is granted the "dba" role. The primary user inherits most of the privileges associated with the role. Please refer to the Amazon RDS User Guide for the list of restricted privileges and the corresponding alternatives to perform administrative tasks that may require these privileges.

For SQL Server, a user that creates a database is granted the "db_owner" role. Please refer to the Amazon RDS User Guide for the list of restricted privileges and the corresponding alternatives to perform administrative tasks that may require these privileges.

Is there anything different about user management with Amazon RDS?

No, everything works the way you are familiar with when using a relational database you manage yourself.

Can programs running on servers in my own data center access Amazon RDS databases?

Yes. You have to intentionally turn on the ability to access your database over the internet by configuring Security Groups. You can authorize access for only the specific IPs, IP ranges, or subnets corresponding to servers in your own data center.

Can I encrypt connections between my application and my DB Instance using SSL/TLS?

Yes, this option is supported for all Amazon RDS engines. Amazon RDS generates an SSL/TLS certificate for each DB Instance. Once an encrypted connection is established, data transferred between the DB Instance and your application will be encrypted during transfer.

While SSL offers security benefits, be aware that SSL/TLS encryption is a compute-intensive operation and will increase the latency of your database connection. SSL/TLS support within Amazon RDS is for encrypting the connection between your application and your DB Instance; it should not be relied on for authenticating the DB Instance itself.

For details on establishing an encrypted connection with Amazon RDS, please visit Amazon RDS's MySQL User Guide, MariaDB User Guide, PostgreSQL User Guide, or Oracle User Guide. To learn more about how SSL/TLS works with these engines, you can refer directly to the MySQL documentation, the MariaDB documentation, the MSDN SQL Server documentation, the PostgreSQL documentation, or the Oracle Documentation.

Can I encrypt data at rest on my Amazon RDS databases?

Amazon RDS supports encryption at rest for all database engines, using keys you manage using AWS Key Management Service (KMS). On a database instance running with Amazon RDS encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, read replicas, and snapshots. Encryption and decryption are handled transparently. For more information about the use of KMS with Amazon RDS, see the Amazon RDS User's Guide.

You can also add encryption to a previously unencrypted DB instance or DB cluster by creating a DB snapshot and then creating a copy of that snapshot and specifying a KMS encryption key. You can then restore an encrypted DB instance or DB cluster from the encrypted snapshot.

Amazon RDS for Oracle and SQL Server support those engines' Transparent Data Encryption (TDE) technologies. For more information, see the Amazon RDS User's Guide for Oracle and SQL Server.

How do I control the actions that my systems and users can take on specific Amazon RDS resources?

You can control the actions that your AWS IAM users and groups can take on Amazon RDS resources. You do this by referencing the Amazon RDS resources in the AWS IAM policies that you apply to your users and groups. Amazon RDS resources that can be referenced in an AWS IAM policy include DB instances, DB snapshots, read replicas, DB security groups, DB option groups, DB parameter groups, event subscriptions, and DB subnet groups.

In addition, you can tag these resources to add additional metadata to your resources. By using tagging, you can categorize your resources (e.g. "Development" DB instances, "Production" DB instances, and "Test" DB instances), and write AWS IAM policies that list the permissions (i.e. actions) that can be taken on resources with the same tags. For more information, refer to Tagging Amazon RDS Resources.

I wish to perform security analysis or operational troubleshooting on my Amazon RDS deployment. Can I get a history of all Amazon RDS API calls made on my account?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more at CloudTrail's AWS Management Console home page.

Can I use Amazon RDS with applications that require HIPAA compliance?

Yes, all Amazon RDS database engines are HIPAA-eligible, so you can use them to build HIPAA-compliant applications and store healthcare-related information, including protected health information (PHI) under an executed Business Associate Agreement (BAA) with AWS.

If you already have an executed BAA, no action is necessary to begin using these services in the account(s) covered by your BAA. If you do not have an executed BAA with AWS, or have any other questions about HIPAA-compliant applications on AWS, please contact your account manager.

Database Configuration

How do I choose the right configuration parameters for my DB Instance(s)?

By default, Amazon RDS chooses the optimal configuration parameters for your DB Instance taking into account the instance class and storage capacity. However, if you want to change them, you can do so using the AWS Management Console, the Amazon RDS APIs, or the AWS Command Line Interface. Please note that changing configuration parameters from recommended values can have unintended effects, ranging from degraded performance to system crashes, and should only be attempted by advanced users who wish to assume these risks.

What are DB Parameter groups? How are they helpful?

A database parameter group (DB Parameter Group) acts as a “container” for engine configuration values that can be applied to one or more DB Instances. If you create a DB Instance without specifying a DB Parameter Group, a default DB Parameter Group is used. This default group contains engine defaults and Amazon RDS system defaults optimized for the DB Instance you are running.

However, if you want your DB Instance to run with your custom-specified engine configuration values, you can simply create a new DB Parameter Group, modify the desired parameters, and modify the DB Instance to use the new DB Parameter Group. Once associated, all DB Instances that use a particular DB Parameter Group get all the parameter updates to that DB Parameter Group.

For more information on configuring DB Parameter Groups, please read the Amazon RDS User Guide.

How can I monitor the configuration of my Amazon RDS resources?

You can use AWS Config to continuously record configuration changes to Amazon RDS DB Instances, DB Subnet Groups, DB Snapshots, DB Security Groups, and Event Subscriptions and receive notification of changes through Amazon Simple Notification Service (SNS). You can

also create AWS Config Rules to evaluate whether these Amazon RDS resources have the desired configurations.

Multi-AZ Deployments

What does it mean to run a DB instance as a Multi-AZ deployment?

When you create or modify your DB instance to run as a Multi-AZ deployment, Amazon RDS automatically provisions and maintains a synchronous "standby" replica in a different Availability Zone. Updates to your DB Instance are synchronously replicated across Availability Zones to the standby in order to keep both in sync and protect your latest database updates against DB instance failure.

During certain types of planned maintenance, or in the unlikely event of DB instance failure or Availability Zone failure, Amazon RDS will automatically failover to the standby so that you can resume database writes and reads as soon as the standby is promoted. Since the name record for your DB instance remains the same, your application can resume database operation without the need for manual administrative intervention. With Multi-AZ deployments, replication is transparent. You do not interact directly with the standby, and it cannot be used to serve read traffic. More information about Multi-AZ deployments is in the Amazon RDS User Guide.

What is an Availability Zone?

Availability Zones are distinct locations within a Region that are engineered to be isolated from failures in other Availability Zones. Each Availability Zone runs on its own physically distinct, independent infrastructure, and is engineered to be highly reliable. Common points of failures like generators and cooling equipment are not shared across Availability Zones. Additionally, they are physically separate, such that even extremely uncommon disasters such as fires, tornados, or flooding would only affect a single Availability Zone. Availability Zones within the same Region benefit from low-latency network connectivity.

What do "primary" and "standby" mean in the context of a Multi-AZ deployment?

When you run a DB instance as a Multi-AZ deployment, the "primary" serves database writes and reads. In addition, Amazon RDS provisions and maintains a "standby" behind the scenes, which is an up-to-date replica of the primary. The standby is "promoted" in failover scenarios. After failover, the standby becomes the primary and accepts your database operations. You do not interact directly with the standby (e.g. for read operations) at any point prior to promotion. If you are interested in scaling read traffic beyond the capacity constraints of a single DB instance, please see the FAQs on Read Replicas.

What are the benefits of a Multi-AZ deployment?

The chief benefits of running your DB instance as a Multi-AZ deployment are enhanced database durability and availability. The increased availability and fault tolerance offered by Multi-AZ deployments make them a natural fit for production environments.

Running your DB instance as a Multi-AZ deployment safeguards your data in the unlikely event of a DB instance component failure or loss of availability in one Availability Zone. For example, if a storage volume on your primary fails, Amazon RDS automatically initiates a failover to the standby, where all of your database updates are intact. This provides additional data durability relative to standard deployments in a single AZ, where a user-initiated restore operation would be required and updates that occurred after the latest restorable time (typically within the last five minutes) would not be available.

You also benefit from enhanced database availability when running your DB instance as a Multi-AZ deployment. If an Availability Zone failure or DB instance failure occurs, your availability impact is limited to the time automatic failover takes to complete. The availability benefits of Multi-AZ also extend to planned maintenance.
For example, with automated backups, I/O activity is no longer suspended on your primary during your preferred backup window, since backups are taken from the standby. In the case of patching or DB instance class scaling, these operations occur first on the standby, prior to automatic failover. As a result, your availability impact is limited to the time required for automatic failover to complete.

Another implied benefit of running your DB instance as a Multi-AZ deployment is that DB instance failover is automatic and requires no administration. In an Amazon RDS context, this means you are not required to monitor DB instance events and initiate manual DB instance recovery (via the RestoreDBInstanceToPointInTime or RestoreDBInstanceFromSnapshot APIs) in the event of an Availability Zone failure or DB instance failure.

Are there any performance implications of running my DB instance as a Multi-AZ deployment?

You may observe elevated latencies relative to a standard DB instance deployment in a single Availability Zone as a result of the synchronous data replication performed on your behalf.

When running my DB instance as a Multi-AZ deployment, can I use the standby for read or write operations?

No, a Multi-AZ standby cannot serve read requests. Multi-AZ deployments are designed to provide enhanced database availability and durability, rather than read scaling benefits. As such, the feature uses synchronous replication between primary and standby. Our implementation makes sure the primary and the standby are constantly in sync, but precludes using the standby for read or write operations. If you are interested in a read scaling solution, please see the FAQs on Read Replicas.

How do I set up a Multi-AZ DB instance deployment?

In order to create a Multi-AZ DB instance deployment, simply click the “Yes” option for “Multi-AZ Deployment” when launching a DB Instance with the AWS Management Console.

Alternatively, if you are using the Amazon RDS APIs, you would call the CreateDBInstance API and set the “Multi-AZ” parameter to the value “true.” To convert an existing standard (single-AZ) DB instance to Multi-AZ, modify the DB instance in the AWS Management Console or use the ModifyDBInstance API and set the Multi-AZ parameter to true.

What happens when I convert my Amazon RDS instance from Single-AZ to Multi-AZ?

For the RDS for MySQL, MariaDB, PostgreSQL, and Oracle database engines, when you elect to convert your Amazon RDS instance from Single-AZ to Multi-AZ, the following happens:

- A snapshot of your primary instance is taken.
- A new standby instance is created in a different Availability Zone, from the snapshot.
- Synchronous replication is configured between primary and standby instances.

As such, there should be no downtime incurred when an instance is converted from Single-AZ to Multi-AZ. However, you may see increased latency while the data on the standby is caught up to match to the primary.

What events would cause Amazon RDS to initiate a failover to the standby replica?

Amazon RDS detects and automatically recovers from the most common failure scenarios for Multi-AZ deployments so that you can resume database operations as quickly as possible without administrative intervention. Amazon RDS automatically performs a failover in the event of any of the following:

- Loss of availability in primary Availability Zone
- Loss of network connectivity to primary
- Compute unit failure on primary
- Storage failure on primary

Note: When operations such as DB instance scaling or system upgrades, like OS patching, are initiated for Multi-AZ deployments, for enhanced availability they are applied first on the standby prior to automatic failover. As a result, your availability impact is limited only to the time required for automatic failover to complete. Note that Amazon RDS Multi-AZ deployments do not failover automatically in response to database operations, such as long running queries, deadlocks, or database corruption errors.

Will I be alerted when automatic failover occurs?

Yes, Amazon RDS will emit a DB instance event to inform you that automatic failover occurred. You can click the "Events" section of the Amazon RDS Console or use the DescribeEvents API to return information about events related to your DB instance. You can also use Amazon RDS Event Notifications to be notified when specific DB events occur.

What happens during Multi-AZ failover and how long does it take?

Failover is automatically handled by Amazon RDS so that you can resume database operations as quickly as possible without administrative intervention. When failing over, Amazon RDS simply flips the canonical name record (CNAME) for your DB instance to point at the standby, which is in turn promoted to become the new primary. We encourage you to follow best practices and implement database connection retry at the application layer.

Failovers, as defined by the interval between the detection of the failure on the primary and the resumption of transactions on the standby, typically complete within one to two minutes. Failover time can also be affected by whether large uncommitted transactions must be recovered; the use of adequately large instance types is recommended with Multi-AZ for best results. AWS also recommends the use of Provisioned IOPS with Multi-AZ instances, for fast, predictable, and consistent throughput performance.

Can I initiate a "forced failover" for my Multi-AZ DB instance deployment?

Amazon RDS will automatically failover without user intervention under a variety of failure conditions. In addition, Amazon RDS provides an option to initiate a failover when rebooting your instance. You can access this feature via the AWS Management Console or when using the RebootDBInstance API call.

How do I control/configure Multi-AZ synchronous replication?

With Multi-AZ deployments, you simply set the “Multi-AZ” parameter to true. The creation of the standby, synchronous replication, and failover are all handled automatically. This means you cannot select the Availability Zone your standby is deployed in or alter the number of standbys available (Amazon RDS provisions one dedicated standby per DB instance primary). The standby also cannot be configured to accept database read activity. Learn more about Multi-AZ configurations.

Will my standby be in the same Region as my primary?

Yes. Your standby is automatically provisioned in a different Availability Zone of the *same Region* as your DB instance primary.

Can I see which Availability Zone my primary is currently located in?

Yes, you can gain visibility into the location of the current primary by using the AWS Management Console or DescribeDBInstances API.

After failover, my primary is now located in a different Availability Zone than my other AWS resources (e.g. EC2 instances). Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same Region. In addition, you may want to consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone failure. Multi-AZ deployments address this need for the database tier without administration on your part.

How do DB Snapshots and automated backups work with my Multi-AZ deployment?

You interact with automated backup and DB Snapshot functionality in the same way whether you are running a standard deployment in a Single-AZ or Multi-AZ deployment. If you are running a Multi-AZ deployment, automated backups and DB Snapshots are simply taken from the standby to avoid I/O suspension on the primary. Please note that you may experience increased I/O latency (typically lasting a few minutes) during backups for both Single-AZ and Multi-AZ deployments.

Initiating a restore operation (point-in-time restore or restore from DB Snapshot) also works the same with Multi-AZ deployments as standard, Single-AZ deployments. New DB instance deployments can be created with either the RestoreDBInstanceFromSnapshot or RestoreDBInstanceToPointInTime APIs. These new DB instance deployments can be either standard or Multi-AZ, regardless of whether the source backup was initiated on a standard or Multi-AZ deployment.

Read Replicas

What does it mean to run a DB Instance as a read replica?

Read replicas make it easy to take advantage of supported engines' built-in replication functionality to elastically scale out beyond the capacity constraints of a single DB instance for read-heavy database workloads.

You can create a read replica with a few clicks in the AWS Management Console or using the CreateDBInstanceReadReplica API. Once the read replica is created, database updates on the source DB instance will be replicated using a supported engine's native, asynchronous replication. You can create multiple read replicas for a given source DB Instance and distribute your application's read traffic amongst them.

Since read replicas use supported engines' built-in replication, they are subject to its strengths and limitations. In particular, updates are applied to your read replica(s) after they occur on the source DB instance, and replication lag can vary significantly. Read replicas can be associated with Multi-AZ deployments to gain read scaling benefits in addition to the enhanced database write availability and data durability provided by Multi-AZ deployments.

When would I want to consider using an Amazon RDS read replica?

There are a variety of scenarios where deploying one or more read replicas for a given source DB instance may make sense. Common reasons for deploying a read replica include:

- Scaling beyond the compute or I/O capacity of a single DB instance for read-heavy database workloads. This excess read traffic can be directed to one or more read replicas.
- Serving read traffic while the source DB instance is unavailable. If your source DB Instance cannot take I/O requests (e.g. due to I/O suspension for backups or scheduled maintenance), you can direct read traffic to your read replica(s). For this use case, keep in mind that the data on the read replica may be "stale" since the source DB Instance is unavailable.
- Business reporting or data warehousing scenarios. You may want business reporting queries to run against a read replica rather than your primary, production DB Instance.
- You may use a read replica for disaster recovery of the source DB instance either in the same AWS Region or in another Region.

Do I need to enable automatic backups on my DB instance before I can create read replicas?

Yes. Enable automatic backups on your source DB Instance before adding read replicas by setting the backup retention period to a value other than 0. Backups must remain enabled for read replicas to work.

Which versions of database engines support Amazon RDS read replicas?

Amazon Aurora: All DB clusters.

Amazon RDS for MySQL: All DB instances support creation of read replicas. Automatic backups must be and remain enabled on the source DB instance for read replica operations. Automatic backups on the replica are supported only for Amazon RDS read replicas running MySQL 5.6 and later, not 5.5.

Amazon RDS for PostgreSQL: DB instances with PostgreSQL version 9.3.5 or newer support creation of read replicas. Existing PostgreSQL instances prior to version 9.3.5 need to be upgraded to PostgreSQL version 9.3.5 to take advantage of Amazon RDS read replicas.

Amazon RDS for MariaDB: All DB instances support creation of read replicas. Automatic backups must be and remain enabled on the source DB Instance for read replica operations.

Amazon RDS for Oracle: Supported for Oracle version 12.1.0.2.v12 and higher and for all 12.2 versions using the Bring Your Own License model with Oracle Database Enterprise Edition and licensed for the Active Data Guard Option.

Amazon RDS for SQL Server: Read replicas are supported on Enterprise Edition in the Multi-AZ configuration when the underlying replication technology is using Always On availability groups for SQL Server versions 2016 and 2017.

How do I deploy a read replica for a given DB instance?

You can create a read replica in minutes using the standard CreateDBInstanceReadReplica API or a few clicks on the AWS Management Console. When creating a read replica, you can identify it as a read replica by specifying a SourceDBInstanceIdentifier. The SourceDBInstanceIdentifier is the DB Instance Identifier of the “source” DB Instance from which you wish to replicate. As with a standard DB Instance, you can also specify the Availability Zone, DB instance class, and preferred maintenance window. The engine version (e.g., PostgreSQL 9.3.5) and storage allocation of a read replica is inherited from the source DB instance.

When you initiate the creation of a read replica, Amazon RDS takes a snapshot of your source DB instance and begins replication. As a result, you will experience a brief I/O suspension on your source DB instance as the snapshot occurs. The I/O suspension typically lasts on the order of one minute and is avoided if the source DB instance is a Multi-AZ deployment (in the case of Multi-AZ deployments, snapshots are taken from the standby).

Amazon RDS is also currently working on an optimization (to be released shortly) such that if you create multiple Read Replicas within a 30 minute window, all of them will use the same source snapshot to minimize I/O impact (“catch-up” replication for each Read Replica will begin after creation).

How do I connect to my read replica(s)?

You can connect to a read replica just as you would connect to a standard DB instance, using the DescribeDBInstance API or AWS Management Console to retrieve the endpoint(s) for your read replica(s). If you have multiple read replicas, it is up to your application to determine how read traffic will be distributed amongst them.

How many read replicas can I create for a given source DB instance?

Amazon RDS for MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server allow you to create up to 5 read replicas for a given source DB instance.

Can I create a read replica in an AWS Region different from that of the source DB instance?

Yes, Amazon RDS (except RDS for SQL Server) supports cross-region read replicas. The amount of time between when data is written to the source DB instance and when it is available in the read replica will depend on the network latency between the two regions.

Do Amazon RDS read replicas support synchronous replication?

No. Read replicas in Amazon RDS for MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server are implemented using those engines' native asynchronous replication. Amazon Aurora uses a different, but still asynchronous, replication mechanism.

Can I use a read replica to enhance database write availability or protect the data on my source DB instance against failure scenarios?

If you are looking to use replication to increase database write availability and protect recent database updates against various failure conditions, we recommend you run your DB instance as a Multi-AZ deployment. With Amazon RDS Read Replicas, which employ supported engines' native, asynchronous replication, database writes occur on a read replica after they have already occurred on the source DB instance, and this replication “lag” can vary significantly.

In contrast, the replication used by Multi-AZ deployments is synchronous, meaning that all database writes are concurrent on the primary and standby. This protects your latest database updates, since they should be available on the standby in the event failover is required.

In addition, with Multi-AZ deployments replication is fully managed. Amazon RDS automatically monitors for DB instance failure conditions or Availability Zone failure and initiates automatic failover to the standby (or to a read replica, in the case of Amazon Aurora) if an outage occurs.

Can I create a read replica with a Multi-AZ DB instance deployment as its source?

Yes. Since Multi-AZ DB instances address a different need than read replicas, it makes sense to use the two in conjunction for production deployments and to associate a read replica with a Multi-AZ DB Instance deployment. The “source” Multi AZ-DB instance provides you with enhanced write availability and data durability, and the associated read replica would improve read traffic scalability.

Can I configure my Amazon RDS read replicas themselves Multi-AZ?

Yes. Amazon RDS for MySQL, MariaDB, PostgreSQL, and Oracle allow you to enable Multi-AZ configuration on read replicas to support disaster recovery and minimize downtime from engine upgrades.

If my read replica(s) use a Multi-AZ DB instance deployment as a source, what happens if Multi-AZ failover occurs?

In the event of Multi-AZ failover, any associated and available read replicas will automatically resume replication once failover has completed (acquiring updates from the newly promoted primary).

Can I create a read replica of another read replica?

Amazon Aurora, Amazon RDS for MySQL, and MariaDB: You can create a second-tier read replica from an existing first-tier read replica. By creating a second-tier read replica, you may be able to move some of the replication load from the primary database instance to a first-tier Read Replica.

Please note that a second-tier Read Replica may lag further behind the primary because of additional replication latency introduced as transactions are replicated from the primary to the first tier replica and then to the second-tier replica.

Amazon RDS for PostgreSQL, Oracle, and SQL Server: Read Replicas of Read Replicas are not currently supported.

Can my read replicas only accept database read operations?

Read replicas are designed to serve read traffic. However, there may be use cases where advanced users wish to complete Data Definition Language (DDL) SQL statements against a read replica. Examples might include adding a database index to a read replica that is used for business reporting without adding the same index to the corresponding source DB instance.

Amazon RDS for MySQL can be configured to permit DDL SQL statements against a read replica. If you wish to enable operations other than reads for a given read replica, modify the active DB parameter group for the read replica setting the “read_only” parameter to “0.”

Amazon RDS for PostgreSQL does not currently support the execution of DDL SQL statements against a read replica.

Can I promote my read replica into a “standalone” DB Instance?

Yes. Refer to the Amazon RDS User Guide for more details.

Will my read replica be kept up-to-date with its source DB instance?

Updates to a source DB instance will automatically be replicated to any associated read replicas. However, with supported engines' asynchronous replication technology, a read replica can fall behind its source DB instance for a variety of reasons. Typical reasons include:

- Write I/O volume to the source DB instance exceeds the rate at which changes can be applied to the read replica (this problem is particularly likely to arise if the compute capacity of a read replica is less than the source DB Instance)
- Complex or long-running transactions to the source DB Instance hold up replication to the read replica
- Network partitions or latency between the source DB instance and a read replica

Read Replicas are subject to the strengths and weaknesses of supported engines' native replication. If you are using Read Replicas, you should be aware of the potential for a lag between a Read Replica and its source DB Instance or “inconsistency”.

How do I see the status of my active read replica(s)?

You can use the standard DescribeDBInstances API to return a list of all the DB Instances you have deployed (including Read Replicas) or simply click on the "Instances" tab of the Amazon RDS Console.

Amazon RDS allows you to gain visibility into how far a read replica has fallen behind its source DB instance. The number of seconds that the read replica is behind the master is published as an Amazon CloudWatch metric ("Replica Lag") available via the AWS Management Console or Amazon CloudWatch APIs.

For Amazon RDS for MySQL, the source of this information is the same as that displayed by issuing a standard "Show Replica Status" MySQL command against the read replica. For Amazon RDS for PostgreSQL, you can use the pg_stat_replication view on the source DB instance to explore replication metrics.

Amazon RDS monitors the replication status of your Read Replicas and updates the Replication State field in the AWS Management console to "Error" if replication stops for any reason (e.g. attempting DML queries on your replica that conflict with the updates made on the primary database instance could result in a replication error). You can review the details of the associated error thrown by the MySQL engine by viewing the Replication Error field and take appropriate action to recover from it. You can learn more about troubleshooting replication

issues in the Troubleshooting a Read Replica Problem section of the User Guide for Amazon RDS for MySQL or PostgreSQL.

If a replication error is fixed, the Replication State changes to Replicating.

I scaled the compute and/or storage capacity of my source DB instance. Should I scale the resources for associated read replica(s) as well?

For replication to work effectively, we recommend that read replicas have as much or more compute and storage resources as their respective source DB instances. Otherwise replication lag is likely to increase or your read replica may run out of space to store replicated updates.

How do I delete a read replica? Will it be deleted automatically if its source DB Instance is deleted?

You can easily delete a read replica with a few clicks of the AWS Management Console or by passing its DB Instance identifier to the DeleteDBInstance API.

An Amazon Aurora replica will stay active and continue accepting read traffic even after its corresponding source DB Instance has been deleted. One of the replicas in the cluster will automatically be promoted as the new primary and will start accepting write traffic.

An Amazon RDS for MySQL or MariaDB read replica will stay active and continue accepting read traffic even after its corresponding source DB instance has been deleted. If you desire to delete the Read Replica in addition to the source DB instance, you must explicitly do so using the DeleteDBInstance API or AWS Management Console.

If you delete an Amazon RDS for PostgreSQL DB Instance that has read replicas, all Read Replicas will be promoted to standalone DB Instances and will be able to accept both read and write traffic. The newly promoted DB Instances will operate independently of one another. If you desire to delete these DB Instances in addition to the original source DB Instance, you must explicitly do so using the DeleteDBInstance API or AWS Management Console.

How much do read replicas cost? When does billing begin and end?

A read replica is billed as a standard DB Instance and at the same rates. Just like a standard DB instance, the rate per “DB Instance hour” for a read replica is determined by the DB instance class of the read replica. You are not charged for the data transfer incurred in replicating data between your source DB instance and read replica within the same AWS Region.

Billing for a read replica begins as soon as the replica has been successfully created (i.e. when the status is listed as “active”). The read replica will continue being billed at standard Amazon RDS DB instance hour rates until you issue a command to delete it.

Enhanced Monitoring

What is Enhanced Monitoring for Amazon RDS?

Enhanced Monitoring for Amazon RDS gives you deeper visibility into the health of your Amazon RDS instances. Just turn on the “Enhanced Monitoring” option for your Amazon RDS DB Instance and set a granularity and Enhanced Monitoring will collect vital operating system metrics and process information, at the defined granularity.

For an even deeper level of diagnostics and visualization of your database load, and a longer data retention period, you can try Performance Insights.

Which metrics and processes can I monitor in Enhanced Monitoring?

Enhanced Monitoring captures your Amazon RDS instance system level metrics, such as the CPU, memory, file system, and disk I/O among others. The complete list of metrics can be found in the documentation.

Which engines are supported by Enhanced Monitoring?

Enhanced Monitoring supports all Amazon RDS database engines.

Which instance types are supported by Enhanced Monitoring?

Enhanced Monitoring supports every instance type except t1.micro and m1.small. The software uses a small amount of CPU, memory, and I/O, and for general purpose monitoring, we recommend switching on higher granularities for instances that are medium or larger. For non-production DB Instances, the default setting for Enhanced Monitoring is “off” and you have the choice of leaving it disabled or modifying the granularity when it is on.

What information can I view on the Amazon RDS dashboard?

You can view all the system metrics and process information for your Amazon RDS DB Instances in a graphical format on the console. You can manage which metrics you want to monitor for each instance and customize the dashboard according to your requirements.

Will all the instances in my Amazon RDS account sample metrics at the same granularity?

No. You can set different granularities for each DB Instance in your Amazon RDS account. You can also choose the instances on which you want to enable Enhanced Monitoring as well as modify the granularity of any instance whenever you want.

How far back can I see the historical metrics on the Amazon RDS console?

You can see the performance values for all the metrics up to 1 hour back at a granularity of up to 1 second based on your settings.

How can I visualize the metrics generated by Amazon RDS Enhanced Monitoring in CloudWatch?

The metrics from Amazon RDS Enhanced Monitoring are delivered into your CloudWatch Logs account. You can create metrics filters in CloudWatch from CloudWatch Logs and display the graphs on the CloudWatch dashboard. For more details, please visit the Amazon CloudWatch page.

When should I use CloudWatch instead of the Amazon RDS console dashboard?

You should use CloudWatch if you want to view historical data beyond what is available on the Amazon RDS console dashboard. You can monitor your Amazon RDS instances in CloudWatch to diagnose the health of your entire AWS stack in a single location. Currently, CloudWatch supports granularities of up to 1 minute and the values will be averaged out for granularities less than that.

Can I set up alarms and notifications based on specific metrics?

Yes. You can create an alarm in CloudWatch that sends a notification when the alarm changes state. The alarm watches a single metric over a time period that you specify and performs one or more actions based on the value of the metric relative to the specified threshold over a number of time periods. For more details on CloudWatch alarms, please visit the Amazon CloudWatch Developer Guide.

How do I integrate Enhanced Monitoring with my tool that I currently use?

Amazon RDS Enhanced Monitoring provides a set of metrics formed as JSON payloads that are delivered into your CloudWatch Logs account. The JSON payloads are delivered at the granularity last configured for the Amazon RDS instance.

There are two ways you can consume the metrics via a third-party dashboard or application. Monitoring tools can use CloudWatch Logs Subscriptions to set up a near real time feed for the metrics. Alternatively, you can use filters in CloudWatch Logs to bridge metrics across to CloudWatch and integrate your application with CloudWatch. Please visit Amazon CloudWatch Documentation for more details.

How can I delete historical data?

Since Enhanced Monitoring delivers JSON payloads into a log in your CloudWatch Logs account, you can control its retention period just like any other CloudWatch Logs stream. The default retention period configured for Enhanced Monitoring in CloudWatch Logs is 30 days. For details on how to change retention settings, please visit Amazon CloudWatch Developer Guide.

What impact does Enhanced Monitoring have on my monthly bills?

Since the metrics are ingested into CloudWatch Logs, your charges will be based on CloudWatch Logs data transfer and storage rates once you exceed CloudWatch Logs free tier. The amount of information transferred for an Amazon RDS instance is directly proportional to the defined granularity for the Enhanced Monitoring feature. Administrators can set different granularities for different instances in their accounts to manage costs.

The approximate volume of data ingested into CloudWatch Logs by Enhanced Monitoring for an instance is as shown below:

Granularity	60 seconds	30 seconds	15 seconds	10 seconds	5 seconds	1 second
Data ingested in CloudWatch Logs* (GB per month)	0.27	0.53	1.07	1.61	3.21	16.07

Amazon RDS Proxy

What is Amazon RDS Proxy?

Amazon RDS Proxy is a fully managed, highly available database proxy feature for Amazon RDS. RDS Proxy makes applications more scalable, more resilient to database failures, and more secure.

Why would I use Amazon RDS Proxy?

Amazon RDS Proxy is a fully managed, highly available, and easy-to-use database proxy feature of Amazon RDS that enables your applications to: 1) improve scalability by pooling and sharing database connections, 2) improve availability by reducing database failover times by up to 66% and preserving application connections during failovers, and 3) improve security by optionally enforcing AWS IAM authentication to databases and securely storing credentials in AWS Secrets Manager.

What use cases does Amazon RDS Proxy address?

Amazon RDS Proxy addresses a number of use cases related to scalability, availability, and security of your applications, including:

Applications with unpredictable workloads: Applications that support highly variable workloads may attempt to open a burst of new database connections. Amazon RDS Proxy's connection governance allows you to gracefully scale applications dealing with unpredictable workloads by efficiently reusing database connections. First, RDS Proxy enables multiple application connections to share a database connection for efficient use of database resources. Second, RDS Proxy allows you to maintain predictable database performance by regulating the number of database connections that are opened. Third, RDS Proxy removes requests that cannot be served to preserve the overall performance and availability of the application.

Applications that frequently open and close database connections: Applications built on technologies such as Serverless, PHP, or Ruby on Rails may open and close database connections frequently to serve application requests. Amazon RDS Proxy maintains a pool of database connections to avoid unnecessary stress on database compute and memory for establishing new connections.

Applications that keep connections open but idle: Applications in industries such as SaaS or eCommerce may keep database connections idling to minimize the response time when a customer reengages. Instead of overprovisioning databases to support mostly idling connections, you can use Amazon RDS Proxy to hold idling connections while only establishing database connections as required to optimally serve active requests.

Applications requiring availability through transient failures: With Amazon RDS Proxy, you can build applications that can transparently tolerate database failures without needing to write complex failure handling code. RDS Proxy automatically routes traffic to a new database instance while preserving application connections. RDS Proxy also bypasses Domain Name System (DNS) caches to reduce failover times by up to 66% for Amazon RDS and Aurora Multi-AZ databases. During database failovers, the application may experience increased latencies and ongoing transactions may have to be retried.

Improved security and centralized credentials management: Amazon RDS Proxy aids you in building more secure applications by giving you a choice to enforce IAM based authentication with relational databases. RDS Proxy also enables you to centrally manage database credentials through AWS Secrets Manager.

When should I connect to the database directly versus using Amazon RDS Proxy?

Depending on your workload, Amazon RDS Proxy can add an average of 5 milliseconds of network latency to query or transaction response time. If your application cannot tolerate 5 milliseconds of latency or does not need connection management and other features enabled by RDS Proxy, you may want your application to connect directly to the database endpoint.

How will serverless applications benefit from Amazon RDS Proxy?

Amazon RDS Proxy transforms your approach to building modern serverless applications that leverage the power and simplicity of relational databases. First, RDS Proxy enables serverless applications to scale efficiently by pooling and reusing database connections. Second, with RDS Proxy, you no longer need to handle database credentials in your Lambda code. You can use the IAM execution role associated with your Lambda function to authenticate with RDS Proxy and your database. Third, you don't need to manage any new infrastructure or code to utilize the full potential of serverless applications backed by relational databases. RDS Proxy is fully managed and scales its capacity automatically based on your application demands.

Which database engines does Amazon RDS Proxy support?

The Amazon RDS Proxy is available for Amazon Aurora with MySQL compatibility and Amazon RDS for MySQL. Support for additional database engines will come soon.

How can I enable Amazon RDS Proxy?

You enable Amazon RDS Proxy for your Amazon RDS database with just a few clicks in the Amazon RDS console. While enabling RDS Proxy, you specify the VPC and subnets you want to access RDS Proxy from. As a Lambda user, you can enable Amazon RDS Proxy for your Amazon RDS database and set up a Lambda function to access it with just a few clicks in the Lambda console.

Can I access Amazon RDS Proxy using APIs?

Yes. You can use Amazon RDS Proxy APIs to create a proxy and then define target groups to associate the proxy with specific database instances or clusters. For example:

```
aws rds create-db-proxy
        --db-proxy-name '…'
        --engine-family <mysql|postgresql>
        --auth [{}, {}]
        --role-arn '…'
        --subnet-ids {}
        --require-tls <true|false>
        --tags {}
aws rds register-db-proxy-targets
        --target-group-name '…'
        --db-cluster-identifier  '…'
        --db-instance-identifier '…'
```

Amazon Aurora

Designed for unparalleled high performance and availability at global scale with full MySQL and PostgreSQL compatibility.

How it works

Amazon Aurora provides built-in security, continuous backups, serverless compute, up to 15 read replicas, automated multi-Region replication, and integrations with other AWS services.

Introduction to Amazon Aurora

Amazon Aurora is a relational database management system (RDBMS) built for the cloud with full MySQL and PostgreSQL compatibility. Aurora gives you the performance and availability of commercial-grade databases at one-tenth the cost.

Use cases

Modernize enterprise applications
Operate enterprise applications, such as customer relationship management (CRM), enterprise resource planning (ERP), supply chain, and billing applications, with high availability and performance.

Build SaaS applications
Support reliable, high- performance, and multi-tenant Software-as a-Service (SaaS) applications with flexible instance and storage scaling.

Deploy globally distributed applications
Develop internet-scale applications, such as mobile games, social media apps, and online services, that require multi-Region scalability and resilience.

Go serverless
Hands-off capacity management, and pay only for capacity consumed with instantaneous and fine-grained scaling to save up to 90% of cost.

What is Amazon Aurora?

Amazon Aurora (Aurora) is a fully managed relational database engine that's compatible with MySQL and PostgreSQL. You already know how MySQL and PostgreSQL combine the speed and reliability of high-end commercial databases with the simplicity and cost-effectiveness of open-source databases. The code, tools, and applications you use today with your existing MySQL and PostgreSQL databases can be used with Aurora. With some workloads, Aurora can deliver up to five times the throughput of MySQL and up to three times the throughput of PostgreSQL without requiring changes to most of your existing applications.

Aurora includes a high-performance storage subsystem. Its MySQL- and PostgreSQL-compatible database engines are customized to take advantage of that fast distributed storage. The underlying storage grows automatically as needed. An Aurora cluster volume can grow to a maximum size of 128 tebibytes (TiB). Aurora also automates and standardizes database clustering and replication, which are typically among the most challenging aspects of database configuration and administration.

Aurora is part of the managed database service Amazon Relational Database Service (Amazon RDS). Amazon RDS is a web service that makes it easier to set up, operate, and scale a relational database in the cloud. If you are not already familiar with Amazon RDS, see the *Amazon Relational Database Service User Guide*.

The following points illustrate how Aurora relates to the standard MySQL and PostgreSQL engines available in Amazon RDS:

- You choose Aurora as the DB engine option when setting up new database servers through Amazon RDS.
- Aurora takes advantage of the familiar Amazon Relational Database Service (Amazon RDS) features for management and administration. Aurora uses the Amazon RDS AWS Management Console interface, AWS CLI commands, and API operations to handle routine database tasks such as provisioning, patching, backup, recovery, failure detection, and repair.
- Aurora management operations typically involve entire clusters of database servers that are synchronized through replication, instead of individual database instances. The automatic clustering, replication, and storage allocation make it simple and cost-effective to set up, operate, and scale your largest MySQL and PostgreSQL deployments.
- You can bring data from Amazon RDS for MySQL and Amazon RDS for PostgreSQL into Aurora by creating and restoring snapshots, or by setting up one-way replication. You can use push-button migration tools to convert your existing Amazon RDS for MySQL and Amazon RDS for PostgreSQL applications to Aurora.

Amazon Aurora FAQs

General

What is Amazon Aurora?

Amazon Aurora is a modern relational database service offering performance and high availability at scale, fully open source MySQL and PostgreSQL-compatible editions, and a range of developer tools for building serverless and machine learning (ML)-driven applications.

Aurora features a distributed, fault-tolerant, and self-healing storage system that is decoupled from compute resources and auto-scales up to 128 TiB per database instance. It delivers high performance and availability with up to 15 low-latency read replicas, point-in-time recovery, continuous backup to Amazon Simple Storage Service (Amazon S3), and replication across three Availability Zones (AZs).

Amazon Aurora is also a fully managed service that automates time-consuming administration tasks like hardware provisioning, database setup, patching, and backups while providing the security, availability, and reliability of commercial databases at 1/10th the cost.

Is Amazon Aurora MySQL compatible?

Amazon Aurora is drop-in compatible with existing MySQL open-source databases and adds support for new releases regularly. This means you can easily migrate MySQL databases to *and from Aurora* using standard import/export tools or snapshots. It also means that most of the code, applications, drivers, and tools you already use with MySQL databases today can be used with Aurora with little or no change. This makes it easy to move applications between the two engines.

You can see the current Amazon Aurora MySQL release compatibility information in the documentation.

Is Amazon Aurora PostgreSQL compatible?

Amazon Aurora is drop-in compatible with existing PostgreSQL open-source databases and adds support for new releases regularly. This means you can easily migrate PostgreSQL databases to and from Aurora using standard import/export tools or snapshots. It also means that most of the code, applications, drivers, and tools you already use with PostgreSQL databases today can be used with Aurora with little or no change.

You can see the current Amazon Aurora PostgreSQL release compatibility information in the documentation.

How is Aurora PostgreSQL supported for issues related to PostgreSQL extensions?

Amazon fully supports Aurora PostgreSQL and all extensions available with Aurora. If you need support for Aurora PostgreSQL, please reach out to AWS Support. If you have an active AWS Premium Support account, you can contact AWS Premium Support for Amazon Aurora specific issues.

How do I get started with Amazon Aurora?

To try Amazon Aurora, sign in to the AWS Management Console, select RDS under the Database category, and choose Amazon Aurora as your database engine. For detailed guidance and resources, check out our Getting started with Aurora page.

How much does Amazon Aurora cost?

For provisioned Aurora, you can choose On-Demand Instances and pay for your database by the hour with no long-term commitments or upfront fees, or choose Reserved Instances for additional savings. Alternatively, Aurora Serverless automatically starts up, shuts down, and scales capacity up or down based on your application's needs and you pay only for capacity consumed.

Amazon Aurora replicates each chunk of my database volume six ways across three Availability Zones. Does that mean that my effective storage price will be three or six times what is shown on the pricing page?

No. Amazon Aurora's replication is bundled into the price. You are charged based on the storage your database consumes at the database layer, not the storage consumed in Amazon Aurora's virtualized storage layer.

In which AWS regions is Amazon Aurora available?

You can see region availability for Amazon Aurora here.

How can I migrate from MySQL to Amazon Aurora and vice versa?

If you want to migrate from MySQL to Amazon Aurora (and vice versa), you have several options:

- You can use the standard mysqldump utility to export data from MySQL and mysqlimport utility to import data to Amazon Aurora, and vice-versa.
- You can also use Amazon RDS's DB Snapshot migration feature to migrate an Amazon RDS for MySQL DB Snapshot to Amazon Aurora using the AWS Management Console.

Migration to Aurora completes for most customers in under an hour, though the duration depends on format and data set size. For more information see Best Practices for Migrating MySQL Databases to Amazon Aurora.

How can I migrate from PostgreSQL to Amazon Aurora and vice versa?

If you want to migrate from PostgreSQL to Amazon Aurora (and vice versa), you have several options:

- You can use the standard pg_dump utility to export data from PostgreSQL and pg_restore utility to import data to Amazon Aurora, and vice-versa.
- You can also use Amazon RDS's DB Snapshot migration feature to migrate an Amazon RDS for PostgreSQL DB Snapshot to Amazon Aurora using the AWS Management Console.

Migration to Aurora completes for most customers in under an hour, though the duration depends on format and data set size.

To migrate SQL Server databases to Aurora PostgreSQL-Compatible Edition, you can use Babelfish for Aurora PostgreSQL. Your applications will work without any changes. See the Babelfish documentation for more information.

Does Amazon Aurora participate in the AWS Free Tier?

Not at this time. The AWS Free Tier for Amazon RDS offers benefits for Micro DB Instances; Amazon Aurora does not currently offer Micro DB Instance support.

To try Amazon Aurora, sign in to the AWS Management Console, select RDS under the Database category, and choose Amazon Aurora as your database engine.

What are I/Os in Amazon Aurora and how are they calculated?

I/Os are input/output operations performed by the Aurora database engine against its solid state drive (SSD)-based virtualized storage layer. Every database page read operation counts as one I/O.

The Aurora database engine issues reads against the storage layer in order to fetch database pages not present in memory in the cache:

- If your query traffic can be totally served from memory or the cache, you will not be charged for retrieving any data pages from memory.
- If your query traffic cannot be served entirely from memory, you will be charged for any data pages that need to be retrieved from storage.

Each database page is 16 KB in Aurora MySQL-Compatible Edition and 8 KB in Aurora PostgreSQL-Compatible Edition.

Aurora was designed to eliminate unnecessary I/O operations in order to reduce costs and to ensure resources are available for serving read/write traffic. Write I/Os are only consumed when persisting redo log records in Aurora MySQL-Compatible Edition or write ahead log records in Aurora PostgreSQL-Compatible Edition to the storage layer for the purpose of making writes durable.

Write I/Os are counted in 4 KB units. For example, a log record that is 1024 bytes will count as one write I/O operation. However, if the log record is larger than 4 KB, more than one write I/O operation will be needed to persist it.

Concurrent write operations whose log records are less than 4 KB may be batched together by the Aurora database engine in order to optimize I/O consumption, if they are persisted on the same storage protection groups. Unlike traditional database engines, Aurora never flushes dirty data pages to storage.

You can see how many I/O requests your Aurora instance is consuming by checking the AWS Management Console. To find your I/O consumption, go to the Amazon RDS section of the console, look at your list of instances, select your Aurora instances, then look for the “Billed read operations” and “Billed write operations” metrics in the monitoring section.

Do I need to change client drivers to use Amazon Aurora PostgreSQL-Compatible Edition?

No, Amazon Aurora will work with standard PostgreSQL database drivers.

Performance

What does "five times the performance of MySQL" mean?

Amazon Aurora delivers significant increases over MySQL performance by tightly integrating the database engine with an SSD-based virtualized storage layer purpose-built for database workloads, reducing writes to the storage system, minimizing lock contention, and eliminating delays created by database process threads.

Our tests with SysBench on r3.8xlarge instances show that Amazon Aurora delivers over 500,000 SELECTs/sec and 100,000 UPDATEs/sec, five times higher than MySQL running the same benchmark on the same hardware. Detailed instructions on this benchmark and how to replicate it yourself are provided in the Amazon Aurora MySQL-Compatible Edition Performance Benchmarking Guide.

What does "three times the performance of PostgreSQL" mean?

Amazon Aurora delivers significant increases over PostgreSQL performance by tightly integrating the database engine with an SSD-based virtualized storage layer purpose-built for database workloads, reducing writes to the storage system, minimizing lock contention, and eliminating delays created by database process threads.

Our tests with SysBench on r4.16xlarge instances show that Amazon Aurora delivers SELECTs/sec and UPDATEs/sec over three times higher than PostgreSQL running the same benchmark on the same hardware. Detailed instructions on this benchmark and how to replicate it yourself are provided in the Amazon Aurora PostgreSQL-Compatible Edition Performance Benchmarking Guide.

How do I optimize my database workload for Amazon Aurora MySQL-Compatible Edition?

Amazon Aurora is designed to be compatible with MySQL so that existing MySQL applications and tools can run without requiring modification. However, one area where Amazon Aurora improves upon MySQL is with highly concurrent workloads. In order to maximize your workload's throughput on Amazon Aurora, we recommend building your applications to drive a large number of concurrent queries and transactions.

How do I optimize my database workload for Amazon Aurora PostgreSQL-Compatible Edition?

Amazon Aurora is designed to be compatible with PostgreSQL so that existing PostgreSQL applications and tools can run without requiring modification. However, one area where Amazon Aurora improves upon PostgreSQL is with highly concurrent workloads. In order to maximize your workload's throughput on Amazon Aurora, we recommend building your applications to drive a large number of concurrent queries and transactions.

Hardware and Scaling

What are the minimum and maximum storage limits of an Amazon Aurora database?

The minimum storage is 10 GB. Based on your database usage, your Amazon Aurora storage will automatically grow, up to 128 TiB, in 10 GB increments with no impact to database performance. There is no need to provision storage in advance.

How do I scale the compute resources associated with my Amazon Aurora DB Instance?

There are two ways to scale the compute resources associated with my Amazon Aurora DB Instance – via Aurora Serverless and via manual adjustment.

You can use Aurora Serverless, an on-demand, autoscaling configuration for Amazon Aurora to scale database compute resources based on application demand. It enables you to run your database in the cloud without worrying about database capacity management. You can specify the desired database capacity range and your database will scale based on your application's needs. Read more in the Aurora Serverless User Guide.

You can also manually scale your compute resources associated with your database by selecting the desired DB instance type in the AWS Management Console. Your requested change will be applied during your specified maintenance window or you can use the "Apply Immediately" flag to change the DB instance type immediately.

Both of these options will have an availability impact for a few minutes as the scaling operation is performed. Note that any other pending system changes will also be applied.

Backup and Restore

How do I enable backups for my DB Instance?

Automated continuous backups are always enabled on Amazon Aurora DB Instances. Backups do not impact database performance.

Can I take DB Snapshots and keep them around as long as I want?

Yes, and there is no performance impact when taking snapshots. Note that restoring data from DB Snapshots requires the creation of a new DB Instance.

If my database fails, what is my recovery path?

Amazon Aurora automatically maintains six copies of your data across three Availability Zones (AZs) and will automatically attempt to recover your database in a healthy AZ with no data loss. In the unlikely event your data is unavailable within Amazon Aurora storage, you can restore from a DB Snapshot or perform a point-in-time restore operation to a new instance. Note that the latest restorable time for a point-in-time restore operation can be up to five minutes in the past.

What happens to my automated backups and DB Snapshots if I delete my DB Instance?

You can choose to create a final DB Snapshot when deleting your DB Instance. If you do, you can use this DB Snapshot to restore the deleted DB Instance at a later date. Amazon Aurora retains this final user-created DB Snapshot along with all other manually created DB Snapshots after the DB Instance is deleted. Only DB Snapshots are retained after the DB Instance is deleted (i.e., automated backups created for point-in-time restore are not kept).

Can I share my snapshots with another AWS account?

Yes. Aurora gives you the ability to create snapshots of your databases, which you can use later to restore a database. You can share a snapshot with a different AWS account, and the owner of the recipient account can use your snapshot to restore a DB that contains your data. You can even choose to make your snapshots public – that is, anybody can restore a DB containing your (public) data.

You can use this feature to share data between your various environments (production, dev/test, staging, etc.) that have different AWS accounts, as well as keep backups of all your data secure in a separate account in case your main AWS account is ever compromised.

Will I be billed for shared snapshots?

There is no charge for sharing snapshots between accounts. However, you may be charged for the snapshots themselves, as well as any databases you restore from shared snapshots.

Can I automatically share snapshots?

We do not support automatic sharing of DB snapshots. To share a snapshot, you must manually create a copy of the snapshot, and then share the copy.

How many accounts can I share snapshots with?

You may share manual snapshots with up to 20 AWS account IDs. If you want to share the snapshot with more than 20 accounts, you can either share the snapshot as public, or contact support for increasing your quota.

In which regions can I share my Aurora snapshots?

You can share your Aurora snapshots within each AWS region where Aurora is available.

Can I share my Aurora snapshots across different regions?

No. Your shared Aurora snapshots will only be accessible by accounts in the same region as the account that shares them.

Can I share an encrypted Aurora snapshot?

Yes, you can share encrypted Aurora snapshots.

High Availability and Replication

How does Amazon Aurora improve my database's fault tolerance to disk failures?

Amazon Aurora automatically divides your database volume into 10 GB segments spread across many disks. Each 10 GB chunk of your database volume is replicated six ways, across three AZs. Amazon Aurora is designed to transparently handle the loss of up to two copies of data without affecting database write availability and up to three copies without affecting read availability.

Amazon Aurora storage is also self-healing. Data blocks and disks are continuously scanned for errors and repaired automatically.

How does Aurora improve recovery time after a database crash?

Unlike other databases, after a database crash Amazon Aurora does not need to replay the redo log from the last database checkpoint (typically five minutes) and confirm that all changes have been applied before making the database available for operations. This reduces database restart times to less than 60 seconds in most cases.

Amazon Aurora moves the buffer cache out of the database process and makes it available immediately at restart time. This prevents you from having to throttle access until the cache is repopulated to avoid brownouts.

What kind of replicas does Aurora support?

Amazon Aurora MySQL-Compatible Edition and Amazon Aurora PostgreSQL-Compatible Edition support Amazon Aurora replicas, which share the same underlying volume as the primary instance in the same AWS region. Updates made by the primary are visible to all Amazon Aurora Replicas.

With Amazon Aurora MySQL-Compatible Edition, you can also create cross-region MySQL Read Replicas based on MySQL's binlog-based replication engine. In MySQL Read Replicas, data from your primary instance is replayed on your replica as transactions. For most use cases, including read scaling and high availability, we recommend using Amazon Aurora Replicas.

You have the flexibility to mix and match these two replica types based on your application needs:

Feature	Amazon Aurora Replicas	MySQL Replicas
Number of replicas	Up to 15	Up to 5
Replication type	Asynchronous (milliseconds)	Asynchronous (seconds)
Performance impact on primary	Low	High
Replica location	In-region	Cross-region
Act as failover target	Yes (no data loss)	Yes (potentially minutes of data loss)
Automated failover	Yes	No
Support for user-defined replication delay	No	Yes
Support for different data or schema vs. primary	No	Yes

You have two additional replication options in addition to the ones listed above. You can use Amazon Global Database for much faster physical replication between Aurora clusters in different regions. And for replication between Aurora and non-Aurora MySQL-Compatible Edition databases (even outside of AWS), you can set up your own, self-managed binlog replication.

Can I have cross-region replicas with Amazon Aurora?

Yes, you can set up cross-region Aurora replicas using either physical or logical replication. Physical replication, called Amazon Aurora Global Database, uses dedicated infrastructure that leaves your databases entirely available to serve your application, and can replicate up to five secondary regions with typical latency of under a second. It's available for both Aurora MySQL-Compatible Edition and Aurora PostgreSQL-Compatible Edition.

For low-latency global reads and disaster recovery, we recommend using Amazon Aurora Global Database.
Aurora supports native logical replication in each database engine (binlog for MySQL and PostgreSQL replication slots for PostgreSQL), so you can replicate to Aurora and non-Aurora databases, even across Regions.

Aurora MySQL-Compatible Edition also offers an easy-to-use logical cross-region read replica feature that supports up to five secondary AWS regions. It is based on single threaded MySQL binlog replication, so the replication lag will be influenced by the change/apply rate and delays in network communication between the specific regions selected.

Can I create Aurora Replicas on the cross-region replica cluster?

Yes, you can add up to 15 Aurora Replicas on each cross-region cluster, and they will share the same underlying storage as the cross-region replica. A cross-region replica acts as the primary

on the cluster and the Aurora Replicas on the cluster will typically lag behind the primary by tens of milliseconds.

Can I fail over my application from my current primary to the cross-region replica?

Yes, you can promote your cross-region replica to be the new primary from the Amazon RDS console. For logical (binlog) replication, the promotion process typically takes a few minutes depending on your workload. The cross-region replication will stop once you initiate the promotion process.

With Amazon Aurora Global Database, you can promote a secondary region to take full read/write workloads in under a minute.

Can I prioritize certain replicas as failover targets over others?

Yes. You can assign a promotion priority tier to each instance on your cluster. When the primary instance fails, Amazon RDS will promote the replica with the highest priority to primary. If two or more Aurora Replicas share the same priority, then Amazon RDS promotes the replica that is largest in size. If two or more Aurora Replicas share the same priority and size, then Amazon RDS promotes an arbitrary replica in the same promotion tier.

For more information on failover logic, read the Amazon Aurora User Guide.

Can I modify priority tiers for instances after they have been created?

Yes, you can modify the priority tier for an instance at any time. Simply modifying priority tiers will not trigger a failover.

Can I prevent certain replicas from being promoted to the primary instance?

You can assign lower priority tiers to replicas that you don't want promoted to the primary instance. However, if the higher priority replicas on the cluster are unhealthy or unavailable for some reason, then Amazon RDS will promote the lower priority replica.

How can I improve upon the availability of a single Amazon Aurora database?

You can add Amazon Aurora Replicas. Aurora Replicas in the same AWS Region share the same underlying storage as the primary instance. Any Aurora Replica can be promoted to primary without any data loss, and therefore can be used to enhance fault tolerance in the event of a primary DB Instance failure.

To increase database availability, simply create one to 15 replicas, in any of three AZs, and Amazon RDS will automatically include them in failover primary selection in the event of a database outage. You can use Amazon Aurora Global Database if you want your database to span multiple AWS Regions. This will replicate your data with no impact on database performance and provide disaster recovery from region-wide outages.

What happens during failover and how long does it take?

Failover is handled automatically by Amazon Aurora so your applications can resume database operations as quickly as possible without manual administrative intervention.

- If you have an Amazon Aurora Replica, in the same or a different Availability Zone, when failing over, Aurora flips the canonical name record (CNAME) for your DB Instance to

point at the healthy replica, which is in turn promoted to become the new primary. Start-to-finish, failover typically completes within 30 seconds.

- If you are running Aurora Serverless and the DB instance or AZ become unavailable, Aurora will automatically recreate the DB instance in a different AZ.
- If you do not have an Amazon Aurora Replica (i.e., single instance) and are not running Aurora Serverless, Aurora will attempt to create a new DB Instance in the same Availability Zone as the original instance. This replacement of the original instance is done on a best-effort basis and may not succeed, for example, if there is an issue that is broadly affecting the Availability Zone.

Your application should retry database connections in the event of connection loss. Disaster recovery across regions is a manual process, where you promote a secondary region to take read/write workloads.

If I have a primary database and an Amazon Aurora Replica actively taking read traffic and a failover occurs, what happens?

Amazon Aurora will automatically detect a problem with your primary instance and trigger a failover. If you are using the Cluster Endpoint, your read/write connections will be automatically redirected to an Amazon Aurora Replica that will be promoted to primary.

In addition, the read traffic that your Aurora Replicas were serving will be briefly interrupted. If you are using the Cluster Reader Endpoint to direct your read traffic to the Aurora Replica, the read only connections will be directed to the newly promoted Aurora Replica until the old primary node is recovered as a replica.

How far behind the primary will my replicas be?

Since Amazon Aurora Replicas share the same data volume as the primary instance in the same AWS Region, there is virtually no replication lag. We typically observe lag times in the tens of milliseconds.

For cross-region replication, binlog-based logical replication lag can grow indefinitely based on change/apply rate as well as delays in network communication. However, under typical conditions, under a minute of replication lag is common. Cross-region replicas using Amazon Aurora Global Database's physical replication will have a typical lag of under a second.

Can I set up replication between my Aurora MySQL-Compatible Edition database and an external MySQL database?

Yes, you can set up binlog replication between an Aurora MySQL-Compatible Edition instance and an external MySQL database. The other database can run on Amazon RDS, or as a self-managed database on AWS, or completely outside of AWS.

If you're running Aurora MySQL-Compatible Edition 5.7, consider setting up GTID-based binlog replication. This will provide complete consistency so your replication won't miss transactions or generate conflicts, even after failover or downtime.

What is Amazon Aurora Global Database?

Amazon Aurora Global Database is a feature that allows a single Amazon Aurora database to span multiple AWS regions. It replicates your data with no impact on database performance, enables fast local reads in each Region with typical latency of less than a second, and provides disaster recovery from region-wide outages. In the unlikely event of a regional degradation or

outage, a secondary region can be promoted to full read/write capabilities in less than one minute. This feature is available for both Aurora MySQL-Compatible Edition and Aurora PostgreSQL-Compatible Edition.

How do I create an Amazon Aurora Global Database?

You can create an Aurora Global Database with just a few clicks in the Amazon RDS console. Alternatively, you can use the AWS Software Development Kit (SDK) or AWS Command-Line Interface (CLI). You need to provision at least one instance per region in your Amazon Aurora Global Database.

How many secondary regions can an Amazon Aurora Global Database have?

You can create up to five secondary regions for an Amazon Aurora Global Database.

If I use Amazon Aurora Global Database, can I also use logical replication (binlog) on the primary database?

Yes. If your goal is to analyze database activity, consider using Aurora advanced auditing, general logs, and slow query logs instead, to avoid impacting the performance of your database.

Will Aurora automatically fail over to a secondary region of an Amazon Aurora Global Database?

No. If your primary region becomes unavailable, you can manually remove a secondary region from an Amazon Aurora Global Database and promote it to take full reads and writes. You will also need to point your application to the newly promoted region.

What is Amazon Aurora Multi-Master?

Amazon Aurora Multi-Master is a new Aurora MySQL-Compatible Edition feature that adds the ability to scale out write performance across multiple Availability Zones, allowing applications to direct read/write workloads to multiple instances in a database cluster and operate with higher availability.

How can I get started with Amazon Aurora Multi-Master?

Amazon Aurora Multi-Master is now generally available. You can read the Amazon Aurora documentation to learn more. You can create an Aurora Multi-Master cluster with just a few clicks in the Amazon RDS console or download the latest AWS SDK or CLI.

Security

Can I use Amazon Aurora in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, all Amazon Aurora DB Instances must be created in a VPC. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network you might operate in your own datacenter. This gives you complete control over who can access your Amazon Aurora databases.

Does Amazon Aurora encrypt my data in transit and at rest?

Yes. Amazon Aurora uses SSL (AES-256) to secure the connection between the database instance and the application. Amazon Aurora allows you to encrypt your databases using keys you manage through AWS Key Management Service (AWS KMS).

On a database instance running with Amazon Aurora encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, snapshots, and replicas in the same cluster. Encryption and decryption are handled seamlessly. For more information about the use of AWS KMS with Amazon Aurora, see the Amazon RDS User's Guide.

Can I encrypt an existing unencrypted database?

Currently, encrypting an existing unencrypted Aurora instance is not supported. To use Amazon Aurora encryption for an existing unencrypted database, create a new DB Instance with encryption enabled and migrate your data into it.

How do I access my Amazon Aurora database?

Amazon Aurora databases must be accessed through the database port entered on database creation. This provides an additional layer of security for your data. Step-by-step instructions on how to connect to your Amazon Aurora database are provided in the Amazon Aurora Connectivity Guide.

Can I use Amazon Aurora with applications that require HIPAA compliance?

Yes, the MySQL- and PostgreSQL-compatible editions of Aurora are Health Insurance Portability and Accountability Act (HIPAA)-eligible, so you can use them to build HIPAA-compliant applications and store healthcare related information, including protected health information (PHI) under an executed Business Associate Agreement (BAA) with AWS. If you already have an executed BAA, no action is necessary to begin using these services in the account(s) covered by your BAA. For more information about building compliant applications on AWS, see Healthcare Providers & Insurers in the Cloud.

Where can I access a list of Common Vulnerabilities and Exposures (CVE) entries for publicly known cybersecurity vulnerabilities for Amazon Aurora releases?

You can currently find a list of CVEs at Amazon Aurora Security Updates.

Serverless

What is Amazon Aurora Serverless?

Aurora Serverless is an on-demand, auto-scaling configuration for Amazon Aurora. It enables you to run your database in the cloud without managing database capacity. Manually managing database capacity can take up valuable time and can lead to inefficient use of database resources. With Aurora Serverless, you simply create a database, specify the desired database capacity range, and connect your application. Aurora automatically adjusts the capacity within the range your specified based on your application’s needs.

You pay on a per-second basis for the database capacity you use when the database is active. Learn more about Aurora Serverless and get started with a few clicks in the Amazon RDS Management Console.

What is the difference between Aurora Serverless v2 and v1?

Aurora Serverless v2 supports every type of database workload, from development and test environments, websites, and applications that have infrequent, intermittent, or unpredictable workloads to the most demanding, business critical applications that require high scale and high availability. It scales in place by adding more CPU and memory without having to failover the database to a larger or smaller database instance. As a result, it can scale even when there are long running transactions, table locks etc.

In addition, it scales database capacity in increments as small as 0.5 Aurora Capacity Unit (ACUs) so your database capacity closely matches your application's needs.

Aurora Serverless v1 is a simple, cost-effective option for infrequent, intermittent, or unpredictable workloads. It automatically starts up, scales compute capacity to match your application's usage, and shuts down when it's not in use. Visit the Aurora User Guide to learn more.

Which Aurora features does Aurora Serverless v2 support?

Aurora Serverless v2 supports all features of provisioned Aurora, including read replica, multi-AZ configuration, Global Database, RDS proxy, and Performance Insights.

Can I start using Aurora Serverless v2 with my existing Aurora DB cluster?

Yes, you can start using Aurora Serverless v2 to manage database compute capacity in your existing Aurora DB cluster. A cluster containing both provisioned instances as well as Aurora Serverless v2 is referred to as a mixed-configuration cluster. You can choose to have any combination of provisioned instances and Aurora Serverless v2 in your cluster.

To test Aurora Serverless v2, you add a reader to your Aurora DB cluster and select Serverless v2 as the instance type. Once the reader is created and available, you can start using it for read-only workloads. Once you confirm that the reader is working as expected, you can initiate a failover to start using Aurora Serverless v2 for both reads and writes. This option provides a minimal downtime experience to get started with Aurora Serverless v2.

Can I migrate from Aurora Serverless v1 to Aurora Serverless v2?

Yes, you can migrate from Aurora Serverless v1 to Aurora Serverless v2. Refer to the Aurora User Guide to learn more.

Which versions of Amazon Aurora are supported for Aurora Serverless?

Aurora Serverless v1 compatibility information can be seen here. Aurora Serverless v2 compatibility information can be seen here.

Can I migrate an existing Aurora DB cluster to Aurora Serverless?

Yes, you can restore a snapshot taken from an existing Aurora provisioned cluster into an Aurora Serverless DB Cluster (and vice versa).

How do I connect to an Aurora Serverless DB cluster?

You access an Aurora Serverless DB cluster from within a client application running in the same VPC. You can't give a public IP address to an Aurora Serverless DB.

Can I explicitly set the capacity of an Aurora Serverless cluster?

While Aurora Serverless automatically scales based on the active database workload, in some cases, capacity might not scale fast enough to meet a sudden workload change, such as a large number of new transactions. In these cases, you can set the capacity explicitly to a specific value with the AWS Management Console, the AWS CLI, or the Amazon RDS API.

Why isn't my Aurora Serverless DB Cluster automatically scaling?

Once a scaling operation is initiated, Aurora Serverless attempts to find a scaling point, which is a point in time at which the database can safely complete scaling. Aurora Serverless might not be able to find a scaling point if you have long-running queries or transactions in progress, or temporary tables or table locks in use.

How am I billed for Aurora Serverless?

In Aurora Serverless, database capacity is measured in Aurora Capacity Units (ACUs). You pay a flat rate per second of ACU usage. Storage and I/O prices are the same for provisioned and serverless configurations.

Parallel Query

What is Amazon Aurora Parallel Query?

Amazon Aurora Parallel Query refers to the ability to push down and distribute the computational load of a single query across thousands of CPUs in Aurora's storage layer. Without Parallel Query, a query issued against an Amazon Aurora database would be executed wholly within one instance of the database cluster; this would be similar to how most databases operate.

What's the target use case?

Parallel Query is a good fit for analytical workloads requiring fresh data and good query performance, even on large tables. Workloads of this type are often operational in nature.

What benefits does Parallel Query provide?

Parallel Query results in faster performance, speeding up analytical queries by up to two orders of magnitude. It also delivers operational simplicity and data freshness as you can issue a query directly over the current transactional data in your Aurora cluster. And, Parallel Query enables transactional and analytical workloads on the same database by allowing Aurora to maintain high transaction throughput alongside concurrent analytical queries.

What specific queries improve under Parallel Query?

Most queries over large data sets that are not already in the buffer pool can expect to benefit. The initial version of Parallel Query can push down and scale out of the processing of more than 200 SQL functions, equijoins, and projections.

What performance improvement can I expect?

The improvement to a specific query's performance depends on how much of the query plan can be pushed down to the Aurora storage layer. Customers have reported more than an order of magnitude improvement to query latency.

Is there any chance that performance will be slower?

Yes, but we expect such cases to be rare.

What changes do I need to make to my query to take advantage of Parallel Query?

Changes in query syntax are not required. The query optimizer will automatically decide whether to use Parallel Query for your specific query. To check if a query is using Parallel Query, you can view the query execution plan by running the EXPLAIN command. If you wish to bypass the heuristics and force Parallel Query for test purposes, use the aurora_pq_force session variable.

How do I turn Parallel Query feature on or off?

Parallel Query can be enabled and disabled dynamically at both the global and session level using the aurora_pq parameter.

Are there any additional charges associated with using Parallel Query?

No. You aren't charged for anything other than what you already pay for instances, I/O, and storage.

Since Parallel Query reduces I/O, will turning it on reduce my Aurora IO charges?

No, Parallel Query I/O costs for your query are metered at the storage layer, and will be the same or larger with Parallel Query turned on. Your benefit is the improvement in query performance.

There are two reasons for potentially higher I/O costs with Parallel Query. First, even if some of the data in a table is in the buffer pool, Parallel Query requires all data to be scanned at the storage layer, incurring I/O. Second, a side effect of avoiding contention in the buffer pool is that running a Parallel Query does not warm up the buffer pool. As a result, consecutive runs of the same Parallel Query query will incur the full I/O cost.

Learn more about Parallel Query in the Documentation.

Is Parallel Query available with all instance types?

No. At this time, you can use Parallel Query with instances in the R* instance family.

Is Parallel Query compatible with all other Aurora features?

Not initially. At this time, you can only turn it on for database clusters that aren't running the Serverless or Backtrack features. Further, it doesn't support functionality specific to Aurora with MySQL 5.7 compatibility.

If Parallel Query speeds up queries with only rare performance losses, should I simply turn it on all the time?

No. While we expect Parallel Query to improve query latency in most cases, you may incur higher I/O costs. We recommend that you thoroughly test your workload with the feature enabled and disabled. Once you're convinced that Parallel Query is the right choice, you can rely on the query optimizer to automatically decide which queries will use Parallel Query. In the rare case when the optimizer doesn't make the optimal decision, you can override the setting.

Can Aurora Parallel Query replace my data warehouse?

Aurora Parallel Query is not a data warehouse and doesn't provide the functionality typically found in such products. It's designed to speed up query performance on your relational database and is suitable for use cases such as operational analytics, when you need to perform fast analytical queries on fresh data in your database.

For an exabyte scale cloud data warehouse, please consider Amazon Redshift.

Amazon DevOps Guru for RDS

What is Amazon DevOps Guru for RDS?

Amazon DevOps Guru for RDS is a new ML-powered capability for Amazon RDS (which includes Amazon Aurora) that is designed to automatically detect and diagnose database performance and operational issues, enabling you to resolve issues in minutes rather than days.

Amazon DevOps Guru for RDS is a feature of Amazon DevOps Guru, which is designed to detect operational and performance issues for all Amazon RDS engines and dozens of other resource types. DevOps Guru for RDS expands the capabilities of DevOps Guru to detect, diagnose, and remediate a wide variety of database-related issues in Amazon RDS (e.g. resource over-utilization, and misbehavior of certain SQL queries).

When an issue occurs, Amazon DevOps Guru for RDS is designed to immediately notify developers and DevOps engineers and provides diagnostic information, details on the extent of the problem, and intelligent remediation recommendations to help customers quickly resolve database-related performance bottlenecks and operational issues.

Why should I use DevOps Guru for RDS?

Amazon DevOps Guru for RDS is designed to remove manual effort and shorten time (from hours and days to minutes) to detect and resolve hard to find performance bottlenecks in your relational database workload.

You can enable DevOps Guru for RDS for every Amazon Aurora database, and it will automatically detect performance issues for your workloads, send alerts to you on each issue, explain findings, and recommend actions to resolve.
DevOps Guru for RDS helps make database administration more accessible to non-experts and assists database experts so that they can manage even more databases.

How does Amazon DevOps Guru for RDS work?

Amazon DevOps Guru for RDS uses ML to analyze telemetry data collected by Amazon RDS Performance Insights (PI). DevOps Guru for RDS does not use any of your data stored in the database in its analysis. PI measures database load, a metric that characterizes how an application spends time in the database and selected metrics generated by the database, such as server status variables in MySQL and pg_stat tables in PostgreSQL.

How can I get started with Amazon DevOps Guru for RDS?

To get started with DevOps Guru for RDS, ensure Performance Insights is enabled through the RDS console, and then simply enable DevOps Guru for your Amazon Aurora databases. With DevOps Guru, you can choose your analysis coverage boundary to be your entire AWS account, prescribe the specific AWS CloudFormation stacks that you want DevOps Guru to analyze, or use AWS tags to create the resource grouping you want DevOps Guru to analyze.

What types of issues can Amazon DevOps Guru for RDS detect?

Amazon DevOps Guru for RDS helps identify a wide range of performance issues that may affect application service quality, such as lock pile-ups, connection storms, SQL regressions, CPU and I/O contention, and memory issues.

How is DevOps Guru for RDS different from Amazon RDS Performance insights?

Amazon RDS Performance Insights is a database performance tuning and monitoring feature that collects and visualizes Amazon RDS database performance metrics, helping you quickly assess the load on your database, and determine when and where to take action. Amazon DevOps Guru for RDS is designed to monitor those metrics, detect when your database is experiencing performance issues, analyze the metrics, and then tell you what's wrong and what you can do about it.

Amazon Neptune

Serverless graph database designed for superior scalability and availability.

Amazon Neptune is a fast, reliable, fully managed graph database service that makes it easy to build and run applications that work with highly connected datasets. The core of Neptune is a purpose-built, high-performance graph database engine that is optimized for storing billions of relationships and querying the graph with milliseconds latency. Neptune supports the popular graph query languages Apache TinkerPop Gremlin and W3C's SPARQL, allowing you to build queries that efficiently navigate highly connected datasets. Neptune powers graph use cases such as recommendation engines, fraud detection, knowledge graphs, drug discovery, and network security.

How it works

Amazon Neptune is a fully managed database service built for the cloud that makes it easier to build and run graph applications. Neptune provides built-in security, continuous backups, serverless compute, and integrations with other AWS services.

Use cases

Transform personalization with customer 360
Easily build identity graphs for identity resolution solutions such as social graphs and 360-degree views of your customers. Accelerate updates for ad targeting, personalization, and analytics.

Detect fraud patterns
Build graph queries for near real-time fraud pattern detection by modeling relationships between people, places, and transactions to discover relationships that might not be obvious.

Unleash machine learning predictions
Amazon Neptune machine learning (ML) uses graph Neural Networks (GNNs) to improve the accuracy of most predictions for graphs by over 50% when compared to predictions using non-graph methods.

Improve IT security
Proactively detect and investigate IT infrastructure using a layered security approach. Model assets to relationships to see how different dimensions of your IT environment interact.

What Is Amazon Neptune?

Amazon Neptune is a fast, reliable, fully managed graph database service that makes it easy to build and run applications that work with highly connected datasets. The core of Neptune is a purpose-built, high-performance graph database engine. This engine is optimized for storing billions of relationships and querying the graph with milliseconds latency. Neptune supports the popular graph query languages Apache TinkerPop Gremlin, the W3C's SPARQL, and Neo4j's openCypher, enabling you to build queries that efficiently navigate highly connected datasets. Neptune powers graph use cases such as recommendation engines, fraud detection, knowledge graphs, drug discovery, and network security.

Neptune is highly available, with read replicas, point-in-time recovery, continuous backup to Amazon S3, and replication across Availability Zones. Neptune provides data security features, with support for encryption at rest and in transit. Neptune is fully managed, so you no longer need to worry about database management tasks like hardware provisioning, software patching, setup, configuration, or backups.

Key Service Components

- *Primary DB instance* – Supports read and write operations, and performs all of the data modifications to the cluster volume. Each Neptune DB cluster has one primary DB instance that is responsible for writing (that is, loading or modifying) graph database contents.

- *Neptune replica* – Connects to the same storage volume as the primary DB instance and supports only read operations. Each Neptune DB cluster can have up to 15 Neptune Replicas in addition to the primary DB instance. This provides high availability by locating Neptune Replicas in separate Availability Zones and distribution load from reading clients.
- *Cluster volume* – Neptune data is stored in the cluster volume, which is designed for reliability and high availability. A cluster volume consists of copies of the data across multiple Availability Zones in a single AWS Region. Because your data is automatically replicated across Availability Zones, it is highly durable, and there is little possibility of data loss.

Supports Open Graph APIs

Amazon Neptune supports open graph APIs for both Gremlin and SPARQL. It provides high performance for both of these graph models and their query languages. You can choose the Property Graph (PG) model and its open source query language, or the Apache TinkerPop Gremlin graph traversal language. Or, you can use the W3C standard Resource Description Framework (RDF) model and its standard SPARQL Query Language.

Highly Secure

Neptune provides multiple levels of security for your database. Security features include network isolation using Amazon VPC, and encryption at rest using keys that you create and control through AWS Key Management Service (AWS KMS). On an encrypted Neptune instance, data in the underlying storage is encrypted, as are the automated backups, snapshots, and replicas in the same cluster.

Fully Managed

With Amazon Neptune, you don't have to worry about database management tasks like hardware provisioning, software patching, setup, configuration, or backups.

You can use Neptune to create sophisticated, interactive graph applications that can query billions of relationships in milliseconds. SQL queries for highly connected data are complex and hard to tune for performance. With Neptune, you can use the popular graph query languages TinkerPop Gremlin and SPARQL to execute powerful queries that are easy to write and perform well on connected data. This capability significantly reduces code complexity so that you can quickly create applications that process relationships.

Neptune is designed to offer greater than 99.99 percent availability. It increases database performance and availability by tightly integrating the database engine with an SSD-backed virtualized storage layer that is built for database workloads. Neptune storage is fault-tolerant and self-healing. Disk failures are repaired in the background without loss of database availability. Neptune automatically detects database crashes and restarts without the need for crash recovery or rebuilding the database cache. If the entire instance fails, Neptune automatically fails over to one of up to 15 read replicas.

Amazon Neptune FAQs

General

What is Amazon Neptune?

Amazon Neptune is a fast, reliable, fully-managed graph database service that makes it easy to build and run applications that work with highly connected datasets. SQL queries for highly connected data are complex and hard to tune for performance. Instead, with Amazon Neptune you can use open and popular graph query languages to execute powerful queries that are easy to write and perform well on connected data. The core of Neptune is a purpose-built, high-performance graph database engine optimized for storing billions of relationships and querying the graph with milliseconds latency. You can use Neptune for graph use cases such as recommendation engines, fraud detection, knowledge graphs, drug discovery, and network security. Amazon Neptune is fully managed and handles the time-consuming tasks such as provisioning, patching, backup, recovery, failure detection and repair. You pay a simple monthly charge for each Amazon Neptune database instance you use. There are no upfront costs or long-term commitments required.

What popular graph query languages does Amazon Neptune support?

Amazon Neptune supports both the open source Apache TinkerPop Gremlin graph traversal language and the W3C standard Resource Description Framework's (RDF) SPARQL query language.

Can I use Apache TinkerPop Gremlin and RDF/SPARQL on the same Neptune instance?

Yes, each Neptune instance provides both a Gremlin Websocket Server and a SPARQL 1.1 Protocol REST endpoint. The data is separated between the stacks, i.e. you cannot execute a Gremlin traversal over RDF data or vice-versa. This is intended to allow you to experiment and see which one is the best fit for your application. In production, as the resources are shared on a single instance, we recommend clients accessing an instance via only one of the languages, i.e. Gremlin or SPARQL.

How can I migrate from an existing Apache TinkerPop Gremlin application to Amazon Neptune?

Amazon Neptune provides an Apache TinkerPop Gremlin Server that supports both Websocket and REST connections. Once you provision an instance of Amazon Neptune, you can configure

your existing TinkerPop application to use the endpoint provided by the service. See also Accessing the Graph via Gremlin.

Do I need to change client drivers to use Amazon Neptune's Gremlin Server?

No, Amazon Neptune's Gremlin server supports clients that are compatible with Apache TinkerPop version 3.3 using both Websocket and REST connections.

How can I migrate from a triple store with a SPARQL endpoint to Amazon Neptune?

Amazon Neptune provides an HTTP REST endpoint that implements the SPARQL 1.1 Protocol. Once you provision a service instance, you can configure your application to point to the SPARQL endpoint. See also Accessing the Graph via SPARQL.

Do I need to change client drivers to use Amazon Neptune's SPARQL Endpoint?

No, Amazon Neptune's SPARQL endpoint will work with any client that supports the SPARQL 1.1 Protocol.

Is Neptune ACID (Atomicity, Consistency, Isolation, Durability) compliant?

Yes, Neptune is ACID compliant with immediate consistency.

Why are Amazon RDS permissions and resources required to use Amazon Neptune?

Amazon Neptune is a purpose-built, high-performance graph database engine. For certain management features such as instance lifecycle management, encryption-at-rest with Amazon Key Management Service (KMS) keys and security groups management, Neptune leverages operational technology that is shared with Amazon RDS.

Does Amazon Neptune have a service level agreement (SLA)?

Yes. Please see the Amazon Neptune SLA.

Performance

What types of graph query workloads are optimized to work with Amazon Neptune?

Amazon Neptune is designed to support graph applications that require high throughput and low latency graph queries. With support for up to 15 read replicas, Amazon Neptune can support 100,000s of queries per second.

Does Amazon Neptune perform query optimization?

Yes, Amazon Neptune uses query optimization for both SPARQL queries and Gremlin traversals.

Is Amazon Neptune built on a relational database?

No, Amazon Neptune is a purpose-built, high-performance graph database engine. Neptune efficiently stores and navigates graph data, and uses a scale-up, in-memory optimized architecture to allow for fast query evaluation over large graphs.

Amazon Neptune replicates each chunk of my database volume six ways across three Availability Zones. Does that mean that my effective storage price will be three or six times what is shown on the pricing page?

No. Amazon Neptune's replication is bundled into the price. You are charged based on the storage your database consumes at the database layer, not the storage consumed in Amazon Neptune's virtualized storage layer .

What are IOs in Amazon Neptune and how are they calculated?

Amazon Neptune was designed to eliminate unnecessary IO operations in order to reduce costs and to ensure resources are available for serving read/write traffic. Write IOs are only consumed when pushing transaction log records to the storage layer for the purpose of making writes durable. Write IOs are counted in 4KB units. For example, a transaction log record that is 1024 bytes will count as one IO operation. However, concurrent write operations whose transaction log is less than 4KB can be batched together by the Amazon Neptune database engine in order to optimize I/O consumption. Unlike traditional database engines Amazon Neptune never pushes modified database pages to the storage layer, resulting in further IO consumption savings.

Hardware and Scaling

What are the minimum and maximum storage limits of an Amazon Neptune database?

The minimum storage is 10GB. Based on your database usage, your Amazon Neptune storage will automatically grow, up to 64 TB, in 10GB increments with no impact to database performance. There is no need to provision storage in advance.

How do I scale the compute resources associated with my Amazon Neptune DB Instance?

You can scale the compute resources allocated to your DB Instance in the AWS Management Console by selecting the desired DB Instance and clicking the Modify button. Memory and CPU resources are modified by changing your DB Instance class.

When you modify your DB Instance class, your requested changes will be applied during your specified maintenance window. Alternatively, you can use the "Apply Immediately" flag to apply your scaling requests immediately. Both of these options will have an availability impact for a few minutes as the scaling operation is performed. Bear in mind that any other pending system changes will also be applied.

Backup and Restore

How do I enable backups for my DB Instance?

Automated backups are always enabled on Amazon Neptune DB Instances. Backups do not impact database performance.

Can I take DB Snapshots and keep them around as long as I want?

Yes, and there is no performance impact when taking snapshots. Note that restoring data from DB Snapshots requires creating a new DB Instance.

If my database fails, what is my recovery path?

Amazon Neptune automatically maintains 6 copies of your data across 3 Availability Zones and will automatically attempt to recover your database in a healthy AZ with no data loss. In the unlikely event your data is unavailable within Amazon Neptune storage, you can restore from a DB Snapshot or perform a point-in-time restore operation to a new instance. Note that the latest restorable time for a point-in-time restore operation can be up to 5 minutes in the past.

What happens to my automated backups and DB Snapshots if I delete my DB Instance?

You can choose to create a final DB Snapshot when deleting your DB Instance. If you do, you can use this DB Snapshot to restore the deleted DB Instance at a later date. Amazon Neptune retains this final user-created DB Snapshot along with all other manually created DB Snapshots after the DB Instance is deleted. Only DB Snapshots are retained after the DB Instance is deleted (i.e., automated backups created for point-in-time restore are not kept).

Can I share my snapshots with another AWS account?

Yes. Neptune gives you the ability to create snapshots of your databases, which you can use later to restore a database. You can share a snapshot with a different AWS account, and the owner of the recipient account can use your snapshot to restore a DB that contains your data. You can even choose to make your snapshots public – that is, anybody can restore a DB containing your (public) data. You can use this feature to share data between your various environments (production, dev/test, staging, etc.) that have different AWS accounts, as well as keep backups of all your data secure in a separate account in case your main AWS account is ever compromised.

Will I be billed for shared snapshots?

There is no charge for sharing snapshots between accounts. However, you may be charged for the snapshots themselves, as well as any databases you restore from shared snapshots. Learn more about Amazon Neptune pricing.

Can I automatically share snapshots?

We do not support sharing automatic DB snapshots. To share an automatic snapshot, you must manually create a copy of the snapshot, and then share the copy.

How many accounts can I share snapshots with?

You may share manual snapshots with up to 20 AWS account IDs. If you want to share the snapshot with more than 20 accounts, you can either share the snapshot as public, or contact support to increase your quota.

In which regions can I share my Amazon Neptune snapshots?

You can share your Neptune snapshots in all AWS regions where Amazon Neptune is available.

Can I share my Amazon Neptune snapshots across different regions?

No. Your shared Amazon Neptune snapshots will only be accessible by accounts in the same region as the account that shares them.

Can I share an encrypted Amazon Neptune snapshot?

Yes, you can share encrypted Amazon Neptune snapshots.

Can I use Amazon Neptune snapshots outside of the service?

No, Amazon Neptune snapshots can only be used inside of the service.

High Availability and Replication

How does Amazon Neptune improve my database's fault tolerance to disk failures?

Amazon Neptune automatically divides your database volume into 10GB segments spread across many disks. Each 10GB chunk of your database volume is replicated six ways, across three Availability Zones. Amazon Neptune is designed to transparently handle the loss of up to two copies of data without affecting database write availability and up to three copies without affecting read availability. Amazon Neptune storage is also self-healing. Data blocks and disks are continuously scanned for errors and repaired automatically.

How does Amazon Neptune improve recovery time after a database crash?

Unlike other databases, after a database crash Amazon Neptune does not need to replay the redo log from the last database checkpoint (typically 5 minutes) and confirm that all changes have been applied, before making the database available for operations. This reduces database restart times to less than 60 seconds in most cases. Amazon Neptune moves the buffer cache out of the database process and makes it available immediately at restart time. This prevents you from having to throttle access until the cache is repopulated to avoid brownouts.

What kind of replicas does Neptune support?

Amazon Neptune supports Read Replicas, which share the same underlying volume as the primary instance. Updates made by the primary are visible to all Amazon Neptune Replicas.

Feature	Amazon Neptune Replicas
Number of replicas	Up to 15
Replication Type	Asynchronous (milliseconds)
Performance impact on primary	Low
Act as failover target	Yes (no data loss)
Automated failover	Yes

Can I have cross-region replicas with Amazon Neptune?

No, Amazon Neptune does not support cross-region replicas.

Can I prioritize certain replicas as failover targets over others?

Yes. You can assign a promotion priority tier to each instance on your cluster. When the primary instance fails, Amazon Neptune will promote the replica with the highest priority to primary. If there is contention between 2 or more replicas in the same priority tier, then Amazon Neptune will promote the replica that is the same size as the primary instance.

Can I modify priority tiers for instances after they have been created?

You can modify the priority tier for an instance at any time. Simply modifying priority tiers will not trigger a failover.

Can I prevent certain replicas from being promoted to the primary instance?

You can assign lower priority tiers to replicas that you don’t want promoted to the primary instance. However, if the higher priority replicas on the cluster are unhealthy or unavailable for some reason, then Amazon Neptune will promote the lower priority replica.

How can I improve upon the availability of a single Amazon Neptune database?

You can add Amazon Neptune Replicas. Amazon Neptune Replicas share the same underlying storage as the primary instance. Any Amazon Neptune Replica can be promoted to become primary without any data loss and therefore can be used for enhancing fault tolerance in the event of a primary DB Instance failure. To increase database availability, simply create 1 to 15 replicas, and Amazon Neptune will automatically include them in failover primary selection in the event of a database outage.

What happens during failover and how long does it take?

Failover is automatically handled by Amazon Neptune so that your applications can resume database operations as quickly as possible without manual administrative intervention.

If you have an Amazon Neptune Replica, in the same or a different Availability Zone, when failing over, Amazon Neptune flips the canonical name record (CNAME) for your DB primary endpoint to a healthy replica, which is in turn is promoted to become the new primary. Start-to-finish, failover typically completes within 30 seconds. Additionally, the read replicas endpoint doesn't require any CNAME updates during failover.
If you do not have an Amazon Neptune Replica (i.e. single instance), Neptune will first attempt to create a new DB Instance in the same Availability Zone as the original instance. If unable to do so, Neptune will attempt to create a new DB Instance in a different Availability Zone. From start to finish, failover typically completes in under 15 minutes.

Your application should retry database requests in the event of connection loss.

If I have a primary database and an Amazon Neptune Replica actively taking read traffic and a failover occurs, what happens?

Amazon Neptune will automatically detect a problem with your primary instance and begin routing your read/write traffic to an Amazon Neptune Replica. On average, this failover will complete within 30 seconds. In addition, the read traffic that your Amazon Neptune Replicas were serving will be briefly interrupted.

How far behind the primary will my replicas be?

Since Amazon Neptune Replicas share the same data volume as the primary instance, there is virtually no replication lag. We typically observe lag times in the 10s of milliseconds.

Security

Can I use Amazon Neptune in Amazon Virtual Private Cloud (Amazon VPC)?

Yes, all Amazon Neptune DB Instances must be created in a VPC. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. This gives you complete control over who can access your Amazon Neptune databases.

Does Amazon Neptune support encrypting my data in transit and at rest?

Amazon Neptune supports HTTPS encrypted client connections and also allows you to encrypt your databases using keys you manage through AWS Key Management Service (KMS). On a database instance running with Amazon Neptune encryption, data stored at rest in the underlying storage is encrypted, as are its automated backups, snapshots, and replicas in the same cluster. Encryption and decryption are handled seamlessly. For more information about the use of KMS with Amazon Neptune, see the Amazon Neptune User Guide.

Can I encrypt an existing unencrypted database?

Currently, encrypting an existing unencrypted Neptune instance is not supported. To use Amazon Neptune encryption for an existing unencrypted database, create a new DB Instance with encryption enabled and migrate your data into it.

How do I access my Amazon Neptune database?

Access to Amazon Neptune databases must be done through the HTTP port entered on database creation within your VPC. This is done to provide an additional layer of security for your data. Step by step instructions on how to connect to your Amazon Neptune database is provided in the Amazon Neptune User Guide.

Amazon ElastiCache

Unlock microsecond latency and scale with in-memory caching.

How it works

Amazon ElastiCache is a fully managed, in-memory caching service supporting flexible, real-time use cases. You can use ElastiCache for caching, which accelerates application and database performance, or as a primary data store for use cases that don't require durability like session stores, gaming leaderboards, streaming, and analytics. ElastiCache is compatible with Redis and Memcached.

Use cases

Accelerate application performance
Access data with microsecond latency and high throughput for lightning-fast application performance.

Ease backend database load

Cache your data to reduce pressure on your backend database, enabling higher application scalability and reducing operational burden.

Build low-latency data stores
Use ElastiCache to store non-durable datasets in memory and support real-time applications with microsecond latency.

Amazon ElastiCache

Amazon ElastiCache makes it easy to set up, manage, and scale distributed in-memory cache environments in the AWS Cloud. It provides a high performance, resizable, and cost-effective in-memory cache, while removing complexity associated with deploying and managing a distributed cache environment. ElastiCache works with both the Redis and Memcached engines; to see which works best for you, see the Comparing Memcached and Redis topic in either user guide.

Amazon ElastiCache for Redis

Redis compatible in-memory data store built for the cloud. Power real-time applications with sub-millisecond latency.

Amazon ElastiCache for Redis is a blazing fast in-memory data store that provides sub-millisecond latency to power internet-scale real-time applications. Built on open-source Redis and compatible with the Redis APIs, ElastiCache for Redis works with your Redis clients and uses the open Redis data format to store your data. Your self-managed Redis applications can work seamlessly with ElastiCache for Redis without any code changes. ElastiCache for Redis combines the speed, simplicity, and versatility of open-source Redis with manageability, security, and scalability from Amazon to power the most demanding real-time applications in Gaming, Ad-Tech, E-Commerce, Healthcare, Financial Services, and IoT.

Benefits

EXTREME PERFORMANCE
Amazon ElastiCache for Redis works as an in-memory data store to support the most demanding applications requiring sub-millisecond response times. In comparison to disk-based databases where most operations need a roundtrip to disk, in-memory data stores manage data in-memory which is an order of magnitude faster than disks. The result is blazing fast performance with average read or write operations taking less than a millisecond and support for hundreds of millions of operations per second within a cluster. ElastiCache gives you an optimized end-to-end hardware and software stack for blazing fast performance.

SECURE
Starting with Amazon ElastiCache for Redis 6, ElastiCache provides you with the ability to create and manage users and user groups that can be used to set up Role-Based Access Control (RBAC) for Redis commands. You can simplify your architecture while maintaining security boundaries, and also take advantage of granular access control to manage groups. You can use AWS Identity and Access Management (IAM) to connect to ElastICache for Redis using IAM identities. ElastiCache for Redis supports Amazon VPC, so you can isolate your cluster to the IP ranges you choose for your nodes. The ElastiCache team continuously monitors for known security vulnerabilities in open-source Redis, operating system, and firmware to keep your Redis environment secure. It is PCI compliant, HIPAA eligible, FedRAMP authorized, and offers encryption in transit, at rest (including customer managed CMK stored in AWS KMS), and Redis AUTH for secure internode communications to help keep sensitive data such as personally identifiable information (PII) safe.

FULLY MANAGED AND HARDENED
Amazon ElastiCache for Redis is a fully managed service. You no longer need to perform management tasks such as hardware provisioning, software patching, setup, configuration, monitoring, failure recovery, and backups. ElastiCache continuously monitors your clusters to keep your Redis up and running so that you can focus on higher value application development. It provides detailed monitoring metrics of your Redis usage, enabling you to track application trends and adjust cluster configuration, as needed. ElastiCache adds automatic write throttling, intelligent swap memory management, and failover enhancements to improve upon the availability and manageability of open source Redis.

REDIS-COMPATIBLE
Redis is a widely adopted in-memory data store for use as a database, cache, message broker, queue, session store, and leaderboard. Amazon ElastiCache for Redis maintains compatibility with the open source Redis data formats, Redis APIs, and works with Redis clients. You can migrate your self-managed Redis workloads to ElastiCache for Redis without any code change.

HIGHLY AVAILABLE AND RELIABLE
Amazon ElastiCache for Redis supports both Redis cluster and non-cluster modes and provides high availability via support for automatic failover by detecting primary node failures and promoting a replica to be primary with minimal impact. It allows for read availability for your application by supporting read replicas (across availability zones), to enable the reads to be served when the primary is busy with the increased workload. ElastiCache for Redis supports enhanced failover logic to allow for automatic failover in cases when majority of the primary nodes for Redis cluster mode are unavailable. On Redis 5.0.5 onwards, auto-failover enabled clusters provide online configuration changes for all planned operations.

EASILY SCALABLE
With Amazon ElastiCache for Redis, you can start small and easily scale your Redis data as your application grows - all the way up to a cluster with 310 TiB of in-memory data, or 982 TiB when using clusters with data tiering. It allows you to scale your Redis Cluster environment up to 500 nodes and 500 shards. It is designed to support online cluster resizing to scale-out and scale-in your Redis clusters without downtime and adapts to changing demand. To scale read capacity, ElastiCache allows you to add up to five read replicas across multiple availability zones. To scale write capacity, ElastiCache supports Redis cluster mode which enables you to partition your write traffic across multiple primaries.

How it works

Use cases

Amazon ElastiCache for Redis is a great choice for real-time transactional and analytical processing use cases such as caching, chat/messaging, gaming leaderboards, geospatial, machine learning, media streaming, queues, real-time analytics, and session store.

Caching

Amazon ElastiCache for Redis is a great choice for implementing a highly available, distributed, and secure in-memory cache to decrease access latency, increase throughput, and ease the load off your relational or NoSQL databases and applications. ElastiCache can serve frequently requested items at sub-millisecond response times, and enables you to easily scale for higher loads without growing the costlier backend databases. Database query results caching, persistent session caching, and full-page caching are all popular examples of caching with ElastiCache for Redis. Learn how to build a caching application with ElastiCache for Redis.

Chat and Messaging

Amazon ElastiCache for Redis supports the PUB/SUB standard with pattern matching. This allows ElastiCache for Redis to support high performance chat rooms, real-time comment streams, and server intercommunication. You can also use PUB/SUB to trigger actions based on published events. Learn how to build a chat application with ElastiCache for Redis.

Gaming Leaderboards

Real-time gaming leaderboards are easy to create with Amazon ElastiCache for Redis. Just use the Redis Sorted Set data structure, which provides uniqueness of elements while maintaining the list sorted by their scores. Creating a real-time ranked list is as simple as updating a user's score each time it changes. You can also use Sorted Sets to handle time series data by using timestamps as the score.

Geospatial

Amazon ElastiCache for Redis offers purpose-built in-memory data structures and operators to manage real-time geospatial data at scale and speed. You can use ElastiCache for Redis to add location-based features such as drive time, drive distance, and points of interests to your applications. Learn how to build a geospatial application with ElastiCache for Redis.

Machine Learning

Amazon ElastiCache for Redis gives you a fast in-memory data store to build and deploy machine learning models quickly. Use ElastiCache for Redis for use cases such as fraud

detection in gaming and financial services, real-time bidding in ad-tech, and matchmaking in dating and ride sharing to process live data and make decisions within tens of milliseconds. Learn how Coffee Meets Bagel uses ElastiCache for real-time machine learning-based dating recommendations.

Media Streaming

Amazon ElastiCache for Redis offers a fast, in-memory data store to power live streaming use cases. ElastiCache for Redis can be used to store metadata for user profile and viewing history, authentication information/tokens for millions of users, and manifest files to enable CDNs to stream videos to millions of mobile and desktop users at a time.

Queues

Amazon ElastiCache for Redis offers List data structure making it easy to implement a lightweight, persistent queue. Lists offer atomic operations as well as blocking capabilities, making them suitable for a variety of applications that require a reliable message broker or a circular list.

Real-time Analytics

Use Amazon ElastiCache for Redis with streaming solutions such as Apache Kafka and Amazon Kinesis as an in-memory data store to ingest, process, and analyze real-time data with sub-millisecond latency. ElastiCache is an ideal choice for real-time analytics use cases such as social media, ad targeting, personalization, and IoT and time-series data analytics.

Session Store

Amazon ElastiCache for Redis is highly suited as a session store to manage session information such as user authentication tokens, session state, and more. Simply use ElastiCache for Redis as a fast key-value store with appropriate TTL on session keys to manage your session information. Session management is commonly required for online applications, including games, e-commerce websites, and social media platforms. Learn how to use ElastiCache for Redis as a session store.

What is Amazon ElastiCache for Redis?

Amazon ElastiCache is a web service that makes it easy to set up, manage, and scale a distributed in-memory data store or cache environment in the cloud. It provides a high-performance, scalable, and cost-effective caching solution. At the same time, it helps remove the complexity associated with deploying and managing a distributed cache environment.

Note

Amazon ElastiCache works with both the Redis and Memcached engines. Use the guide for the engine that you're interested in. If you're unsure which engine you want to use, see Comparing Memcached and Redis in this guide.

Overview of ElastiCache for Redis

Existing applications that use Redis can use ElastiCache with almost no modification. Your applications simply need information about the host names and port numbers of the ElastiCache nodes that you have deployed.

ElastiCache for Redis has multiple features that help make the service more reliable for critical production deployments:

- Automatic detection of and recovery from cache node failures.
- Multi-AZ for a failed primary cluster to a read replica, in Redis clusters that support replication.
- Redis (cluster mode enabled) supports partitioning your data across up to 500 shards.
- For Redis version 3.2 and later, all versions support encryption in transit and encryption at rest encryption with authentication. This support helps you build HIPAA-compliant applications.
- Flexible Availability Zone placement of nodes and clusters for increased fault tolerance.
- Integration with other AWS services such as Amazon EC2, Amazon CloudWatch, AWS CloudTrail, and Amazon SNS. This integration helps provide a managed in-memory caching solution that is high-performance and highly secure.
- ElastiCache for Redis manages backups, software patching, automatic failure detection, and recovery.
- You can have automated backups performed when you need them, or manually create your own backup snapshot. You can use these backups to restore a cluster. The ElastiCache for Redis restore process works reliably and efficiently.
- You can get high availability with a primary instance and a synchronous secondary instance that you can fail over to when problems occur. You can also use read replicas to increase read scaling.
- You can control access to your ElastiCache for Redis clusters by using AWS Identity and Access Management to define users and permissions. You can also help protect your clusters by putting them in a virtual private cloud (VPC).
- By using the Global Datastore for Redis feature, you can work with fully managed, fast, reliable, and secure replication across AWS Regions. Using this feature, you can create cross-Region read replica clusters for

ElastiCache for Redis to enable low-latency reads and disaster recovery across AWS Regions.

- Data tiering provides a price-performance option for Redis workloads by utilizing lower-cost solid state drives (SSDs) in each cluster node in addition to storing data in memory. It is ideal for workloads that access up to 20 percent of their overall dataset regularly, and for applications that can tolerate additional latency when accessing data on SSD. For more information, see Data tiering.

Clusters

The basic building block of ElastiCache for Redis is the cluster. A cluster is a collection of one or more cache nodes, all of which run an instance of the Redis cache engine software. When you create a cluster, you specify the engine and version for all of the nodes to use. Your ElastiCache for Redis instances are designed to be accessed through an Amazon EC2 instance. You can create and modify a cluster by using the AWS CLI, the ElastiCache for Redis API, or the AWS Management Console.

Each ElastiCache for Redis cluster runs a Redis engine version. Each Redis engine version has its own supported features. Additionally, each Redis engine version has a set of parameters in a parameter group that control the behavior of the clusters that it manages.

The computation and memory capacity of a cluster is determined by its instance, or node, class. You can select the node type that best meets your needs. If your needs change over time, you can change node types. For information, see Supported node types.

You can also leverage data-tiering when considering your node type needs. Data tiering is a feature where some least frequently used data is stored on disk to mitigate against memory limitations on applications that can tolerate additional latency when data on SSD (solid state drives) is accessed.

Cluster node storage comes in two types: Standard and memory-optimized. They differ in performance characteristics and price, allowing you to tailor your storage performance and cost to your needs. Each instance has minimum and maximum storage requirements depending on the storage type. It's important to have sufficient storage so that your clusters have room to grow. Also, sufficient storage makes sure that features have room to write content or log entries.

You can run a cluster on a virtual private cloud (VPC) using the Amazon Virtual Private Cloud (Amazon VPC) service. When you use a VPC, you have control over your virtual networking environment. You can choose your own IP address range, create subnets, and configure routing and access control lists. ElastiCache manages backups, software patching, automatic failure detection, and recovery. There's no additional cost to run your cluster in a VPC. For more information on using Amazon VPC with ElastiCache for Redis, see Amazon VPCs and ElastiCache security.

Security

A security group controls the access to a cluster. It does so by allowing access to IP address ranges or Amazon EC2 instances that you specify. For more information about security groups, see Security in ElastiCache for Redis.

Monitoring an ElastiCache for Redis cluster

There are several ways that you can track the performance and health of a ElastiCache for Redis cluster. You can use the CloudWatch service to monitor the performance and health of a cluster. CloudWatch performance charts are shown in the ElastiCache for Redis console. You can also subscribe to ElastiCache for Redis events to be notified about changes to a cluster, snapshot, parameter group, or security group. For more information, see Monitoring Use with CloudWatch metrics.

What is Amazon ElastiCache for Memcached?

Amazon ElastiCache is a web service that makes it easy to set up, manage, and scale a distributed in-memory data store or cache environment in the cloud. It provides a high-performance, scalable, and cost-effective caching solution. At the same time, it helps remove the complexity associated with deploying and managing a distributed cache environment.

Note

Amazon ElastiCache works with both the Redis and Memcached engines. Use the guide for the engine that you're interested in. If you're unsure which engine you want to use, see Comparing Memcached and Redis in this guide.

Existing applications that use Memcached can use ElastiCache with almost no modification. Your applications simply need information about the host names and port numbers of the ElastiCache nodes that you have deployed. The ElastiCache

Auto Discovery feature for Memcached lets your applications identify all of the nodes in a cache cluster and connect to them. This means that you don't have to maintain a list of available host names and port numbers. In this way, your applications are effectively insulated from changes to node membership in a cluster.

ElastiCache for Memcached has multiple features to enhance reliability for critical production deployments:

- Automatic detection and recovery from cache node failures.
- Automatic discovery of nodes within a cluster enabled for automatic discovery, so that no changes need to be made to your application when you add or remove nodes.
- Flexible Availability Zone placement of nodes and clusters.
- Integration with other AWS services such as Amazon EC2, Amazon CloudWatch, AWS CloudTrail, and Amazon SNS to provide a secure, high-performance, managed in-memory caching solution.

Amazon ElastiCache for Memcached

Managed, Memcached-compatible, in-memory store. Sub-millisecond latency to power real-time applications.

Amazon ElastiCache for Memcached is a Memcached-compatible, in-memory, key-value store service that can be used as a cache or a data store. It delivers the performance, ease-of-use, and simplicity of Memcached. ElastiCache for Memcached is fully managed, scalable, and secure - making it an ideal candidate for use cases where frequently accessed data must be in-memory. It is a popular choice for use cases such as web, mobile apps, gaming, ad-tech, and e-commerce.

Benefits

EXTREME PERFORMANCE

Amazon ElastiCache for Memcached works as an in-memory data store and cache to support the most demanding applications requiring sub-millisecond response times. By utilizing an end-to-end optimized stack running on customer dedicated nodes, Amazon ElastiCache provides you secure, blazing fast performance.

SECURE AND HARDENED

Amazon ElastiCache for Memcached supports Amazon VPC, enabling you to isolate your cluster to the IP ranges you choose for your nodes, and use them to connect to your application. ElastiCache continuously monitors your nodes and applies the necessary patches

to keep your environment secure. ElastiCache for Memcached offers encryption in transit using Transport Layer Security (TLS) version 1.2. Learn more »

MEMCACHED-COMPATIBLE

Memcached is a widely adopted memory object caching system. ElastiCache is compliant with Memcached, so popular tools that you use today with existing Memcached environments will work seamlessly with the service. Learn more about open-source Memcached »

EASILY SCALABLE

Amazon ElastiCache with Memcached includes sharding to scale in-memory cache with up to 20 nodes and 12.7 TiB per cluster. Learn more »

FULLY MANAGED

You no longer need to perform management tasks such as hardware provisioning, software patching, setup, configuration, monitoring, and failure recovery. ElastiCache continuously monitors your clusters to keep your workloads up and running so that you can focus on higher value application development. Learn more »

AUTO DISCOVERY

ElastiCache for Memcached Cluster Client with Auto Discovery saves you time by simplifying the way an application connects to a Memcached cluster. The client connects to a single cluster endpoint and gets configuration information about other nodes. When nodes are added or removed, the client reconfigures itself to use the most current set of nodes. Learn more »

How it works

Use cases

Caching
Amazon ElastiCache for Memcached is a great choice for implementing an in-memory cache to decrease access latency, increase throughput, and ease the load off your relational or NoSQL database. Amazon ElastiCache can serve frequently requested items at sub-millisecond

response times, and enables you to easily scale for higher loads without growing the costlier backend database layer. Database query results caching, persistent session caching, and full-page caching are all popular examples of caching with ElastiCache for Memcached.

Session Store
Session stores are easy to create with Amazon ElastiCache for Memcached. Simply use the Memcached hash table, which can be distributed across multiple nodes. Scaling the session store is as easy as adding a node and updating the clients to take advantage of the new node.

Amazon ElastiCache FAQs

General

The Basics

Q: What is Amazon ElastiCache?

Amazon ElastiCache is a web service that makes it easy to deploy and run Memcached or Redis protocol-compliant server nodes in the cloud. Amazon ElastiCache improves the performance of web applications by allowing you to retrieve information from a fast, managed, in-memory system, instead of relying entirely on slower disk-based databases. The service simplifies and offloads the management, monitoring and operation of in-memory environments, enabling your engineering resources to focus on developing applications. Using Amazon ElastiCache, you can not only improve load and response times to user actions and queries, but also reduce the cost associated with scaling web applications.

Amazon ElastiCache automates common administrative tasks required to operate a distributed in-memory key-value environment. Using Amazon ElastiCache, you can add a caching or in-memory layer to your application architecture in a matter of minutes via a few clicks of the AWS Management Console. Once a cluster is provisioned, Amazon ElastiCache automatically detects and replaces failed nodes, providing a resilient system that mitigates the risk of overloaded databases, which slow website and application load times. Through integration with Amazon CloudWatch monitoring, Amazon ElastiCache provides enhanced visibility into key performance metrics associated with your nodes. Amazon ElastiCache is protocol-compliant with Memcached and Redis, so code, applications, and popular tools that you use today with your existing Memcached or Redis environments will work seamlessly with the service. With the support for clustered configuration in Amazon ElastiCache, you get the benefits of fast, scalable and easy to use managed service that can meet the needs of your most demanding applications. As with all Amazon Web Services, there are no up-front investments required, and you pay only for the resources you use.

Q: What is in-memory caching and how does it help my applications?

The in-memory caching provided by Amazon ElastiCache can be used to significantly improve latency and throughput for many read-heavy application workloads (such as social networking, gaming, media sharing and Q&A portals) or compute-intensive workloads (such as a recommendation engine). In-memory caching improves application performance by storing critical pieces of data in memory for low-latency access. Cached information may include the results of I/O-intensive database queries or the results of computationally-intensive calculations.

Q: Can I use Amazon ElastiCache for use cases other than caching?

A: Yes. ElastiCache for Redis can be used as a primary in-memory key-value data store, providing fast, sub millisecond data performance, high availability and scalability. You can choose to configure a 500-node cluster that ranges between 83 shards (one master and five replicas per shard) and 500 shards (single master and no replicas), giving you up to 340 TB of

memory. Support for 500-node cluster is available with Amazon ElastiCache for Redis starting with Redis version 5.0.6. See here for other use cases, such as leaderboards, rate limiting, queues, and chat.

Q: Can I use Amazon ElastiCache through AWS CloudFormation?

AWS CloudFormation simplifies provisioning and management by providing AWS CloudFormation templates for quick and reliable provisioning of the services or applications. AWS CloudFormation provides comprehensive support for Amazon ElastiCache by providing templates to create cluster (both MemCached and Redis) and Replication Groups. The templates are up to date with the latest ElastiCache Redis announcement for clustered Redis configuration and provide flexibility and ease of use to Amazon ElastiCache customers.

Q: What does Amazon ElastiCache manage on my behalf?

Amazon ElastiCache manages the work involved in setting up a distributed in-memory environment, from provisioning the server resources you request to installing the software. Once your environment is up and running, the service automates common administrative tasks such as failure detection and recovery, and software patching. Amazon ElastiCache provides detailed monitoring metrics associated with your nodes, enabling you to diagnose and react to issues very quickly. For example, you can set up thresholds and receive alarms if one of your nodes is overloaded with requests.

Q: What are Amazon ElastiCache nodes, shards and clusters?

A node is the smallest building block of an Amazon ElastiCache deployment. It is a fixed-size chunk of secure, network-attached RAM. Each node runs an instance of the Memcached or Redis protocol-compliant service and has its own DNS name and port. Multiple types of nodes are supported, each with varying amount of associated memory. A Redis shard is a subset of the cluster's keyspace, that can include a primary node and zero or more read-replicas. For more details on Redis deployments see the Redis section below. The shards add up to form a cluster.

Q: Which engines does Amazon ElastiCache support?

Amazon ElastiCache offers fully managed Redis, voted the most loved database by developers in the Stack Overflow Developer Survey for 5 years in a row, and Memcached for your most demanding applications that require sub-millisecond response times.

Q: How do I get started with Amazon ElastiCache?

If you are not already signed up for Amazon ElastiCache, you can click the "Get started" button on the Amazon ElastiCache page and complete the sign-up process. You must have an Amazon Web Services account; if you do not already have one, you will be prompted to create one when you begin the Amazon ElastiCache sign-up process. After you are signed up for ElastiCache, please refer to the Amazon ElastiCache documentation, which includes the Getting Started Guide for Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached.

Once you have familiarized yourself with Amazon ElastiCache, you can launch a cluster within minutes by using the AWS Management Console or Amazon ElastiCache APIs.

Q: How do I create a cluster?

Clusters are simple to create, using the AWS Management Console, Amazon ElastiCache APIs, or Command Line Tools. To launch a cluster using the AWS Management Console, click on the "Create" button in either the "Memcached" or "Redis" tab. From there, all you need to specify is your Cluster Identifier, Node Type, and Number of Nodes to create a cluster with the amount of memory you require. Alternatively, you can create your cluster using the CreateCacheCluster

API or elasticache-create-cache-cluster command. If you do not specify an Availability Zone when creating a cluster, AWS will place it automatically based upon your memory requirements and available capacity.

Q: What Node Types can I select?

Amazon ElastiCache supports Nodes of the following types:

Current Generation Nodes:

- cache.m4.large: 6.42 GiB
- cache.m4.xlarge: 14.28 GiB
- cache.m4.2xlarge: 29.7 GiB
- cache.m4.4xlarge: 60.78 GiB
- cache.m4.10xlarge: 154.64 GiB
- cache.m5.large: 6.38 GiB
- cache.m5.xlarge: 12.93 GiB
- cache.m5.2xlarge: 26.04 GiB
- cache.m5.4xlarge: 52.26 GiB
- cache.m5.12xlarge: 157.12 GiB
- cache.m5.24xlarge: 314.32 GiB
- cache.m6g.large: 6.38 GiB
- cache.m6g.xlarge: 12.94 GiB
- cache.m6g.2xlarge: 26.05 GiB
- cache.m6g.4xlarge: 52.26 GiB
- cache.m6g.8xlarge: 103.68 GiB
- cache.m6g.12xlarge: 157.13 GiB
- cache.m6g.16xlarge: 209.55 GiB
- cache.r4.large: 12.3 GiB
- cache.r4.xlarge: 25.05 GiB
- cache.r4.2xlarge: 50.47 GiB
- cache.r4.4xlarge: 101.38 GiB
- cache.r4.8xlarge: 203.26 GiB
- cache.r4.16xlarge: 407 GiB
- cache.r5.large: 13.07 GiB
- cache.r5.xlarge: 26.32 GiB
- cache.r5.2xlarge: 52.82 GiB
- cache.r5.4xlarge: 105.81 GiB
- cache.r5.12xlarge: 317.77 GiB
- cache.r5.24xlarge: 635.61 GiB
- cache.r6g.large: 13.07 GiB
- cache.r6g.xlarge: 26.32 GiB
- cache.r6g.2xlarge: 52.82 GiB
- cache.r6g.4xlarge: 105.81 GiB
- cache.r6g.8xlarge: 209.55 GiB
- cache.r6g.12xlarge: 317.77 GiB
- cache.r6g.16xlarge: 419.1 GiB
- cache.t2.micro: 555 MB
- cache.t2.small: 1.55 GiB
- cache.t2.medium: 3.22 GiB
- cache.t3.micro: 0.5 GiB
- cache.t3.small: 1.37 GiB
- cache.t3.medium: 3.09 GiB

- cache.t4g.micro: 0.5 GiB

- cache.t4g.small: 1.37 GiB
- cache.t4g.medium: 3.09 GiB

Current Generation Nodes with data tiering:

- cache.r6gd.xlarge: 26.32 GiB memory, 99.33 GiB SSD
- cache.r6gd.2xlarge: 52.82 GiB memory, 199.07 GiB SSD
- cache.r6gd.4xlarge: 105.81 GiB memory, 398.14 GiB SSD
- cache.r6gd.8xlarge: 209.55 GiB memory, 796.28 GiB SSD
- cache.r6gd.12xlarge: 317.77 GiB memory, 1194.42 GiB SSD
- cache.r6gd.16xlarge: 419.1 GiB memory, 1592.56 GiB SSD

Previous Generation Nodes:

- cache.m1.small: 1.3 GiB
- cache.m1.medium: 3.35 GiB
- cache.m1.large: 7.1 GiB
- cache.m1.xlarge: 14.6 GiB
- cache.m2.xlarge: 16.7 GiB
- cache.m2.2xlarge: 33.8 GiB
- cache.m2.4xlarge: 68 GiB
- cache.m3.medium: 2.78 GiB
- cache.m3.large: 6.05 GiB
- cache.m3.xlarge: 13.3 GiB
- cache.m3.2xlarge: 27.9 GiB
- cache.r3.large: 13.5 GiB
- cache.r3.xlarge: 28.4 GiB
- cache.r3.2xlarge: 58.2 GiB
- cache.r3.4xlarge: 118 GiB
- cache.r3.8xlarge: 237 GiB
- cache.t1.micro: 213 MB
- cache.c1.xlarge: 6.6 GiB

Each Node Type above lists the memory available to Memcached or Redis after taking Amazon ElastiCache System Software overhead into account. The total amount of memory in a cluster is

an integer multiple of the memory available in each shard. For example, a cluster consisting of ten shards of 6 GB each will provide 60 GB of total memory.

Q: How do I access my nodes?

Once your cluster is available, you can retrieve your node endpoints using the following steps on the AWS Management Console:

- Navigate to the "Amazon ElastiCache" tab.
- Click on the "(Number of) Nodes" link and navigate to the "Nodes" tab.
- Click on the "Copy Node Endpoint(s)" button.

Alternatively, you can use the DescribeCacheClusters API to retrieve the Endpoint list.

You can then configure your Memcached or Redis client with this endpoint list and use your favorite programming language to add or delete data from your ElastiCache Nodes. In order to allow network requests to your nodes, you will need to authorize access. For a detailed explanation to get started, please refer to our Getting Started Guide for Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached.

Q: What is a maintenance window? Will my nodes be available during software maintenance?

You can think of the Amazon ElastiCache maintenance window as an opportunity to control when software patching occurs, in the event either are requested or required. If a "maintenance" event is scheduled for a given week, it will be initiated and completed at some point during the 60 minute maintenance window you identify.

Your nodes could incur some downtime during your maintenance window if software patching is scheduled. Please refer to Engine Version Management for more details. Patching can be user requested - for example cache software upgrade, or determined as required (if we identify any security vulnerabilities in the system or caching software). Software patching occurs infrequently (typically once every few months) and should seldom require more than a fraction of your maintenance window. If you do not specify a preferred weekly maintenance window when creating your Cluster, a 60 minute default value is assigned. If you wish to modify when maintenance is performed on your behalf, you can do so by modifying your DB Instance in the AWS Management Console or by using the ModifyCacheCluster API. Each of your Clusters can have different preferred maintenance windows, if you so choose.

Reserved Nodes

Q: What are Amazon ElastiCache Reserved Nodes?

Reserved Nodes or Reserved Instance (RI) is an offering that provides you with a significant discount over on-demand usage when you commit to a one-year or three-year term. With Reserved Nodes, you can make a one-time, up-front payment to create a one or three year reservation to run your node in a specific Region and receive a significant discount off of the ongoing hourly usage charge. There are three Reserved Node types All Upfront, No Upfront and Partial Upfront that enable you to balance the amount you pay upfront with your effective hourly price.

Q: How are Reserved Nodes different from On-Demand Nodes?

Functionally, Reserved Nodes and On-Demand Nodes are exactly the same. The only difference is how your Node(s) are billed; with Reserved Nodes, you make a one-time up-front payment and receive a lower ongoing hourly usage rate (compared with On-Demand Nodes) for the duration of the term.

Q: How do I purchase and create Reserved Nodes?

You can use the "Purchase Reserved Nodes" option in the AWS Management Console. Alternatively, you can use the API tools to list the reservations available for purchase with the DescribeReservedCacheNodesOfferings API method and then purchase a cache node reservation by calling the PurchaseReservedCacheNodesOffering method.

Creating a Reserved Node is no different than launching an On-Demand Node. You simply specify the node class and Region for which you made the reservation. So long as your reservation purchase was successful, Amazon ElastiCache will apply the reduced hourly rate for which you are eligible to the new node.

Q: Will there always be reservations available for purchase?

Yes. Reserved Nodes are purchased for the Region rather than for the Availability Zone. This means that even if capacity is limited in one Availability Zone, reservations can still be purchased in that Region and used in a different Availability Zone within that Region.

Q: How many Reserved Cache can I purchase?

You can purchase up to 300 Reserved Nodes. If you wish to run more than 300 Nodes please complete the Amazon ElastiCache Node request form.

Q: What if I have an existing node that I'd like to convert to a Reserved Node?

Simply purchase a node reservation with the same node class, within the same region as the node you are currently running and would like to reserve. If the reservation purchase is successful, Amazon ElastiCache will automatically apply your new hourly usage charge to your existing node.

Q: If I sign up for a Reserved Node, when does the term begin? What happens to my node when the term ends?

Pricing changes associated with a Reserved Node are activated once your request is received while the payment authorization is processed. You can follow the status of your reservation on the AWS Account Activity page or by using the DescribeReservedCacheNodes API. If the one-time payment cannot be successfully authorized by the next billing period, the discounted price will not take effect.

When your reservation term expires, your Reserved Node will revert to the appropriate On-Demand hourly usage rate for your node class and region.

Q: How do I control which nodes are billed at the Reserved Node rate?

The Amazon ElastiCache APIs for creating, modifying, and deleting nodes do not distinguish between On-Demand and Reserved Nodes so that you can seamlessly use both. When computing your bill, our system will automatically apply your Reservation(s), such that all eligible nodes are charged at the lower hourly Reserved Cache Node rate.

Q: Can I move a Reserved Node from one Region or Availability Zone to another?

Each Reserved Node is associated with a specific Region, which is fixed for the lifetime of the reservation and cannot be changed. Each reservation can, however, be used in any of the available AZs within the associated Region.

Q: Can I cancel a reservation?

No, you cannot cancel your reserved DB instance and the one-time payment (if applicable) is not refundable. You will continue to pay for every hour during your Reserved DB instance term regardless of your usage.

Q: How do the payment options impact my bill?

When you purchase an RI under the All Upfront payment option, you pay for the entire term of the RI in one upfront payment. You can choose to pay nothing upfront by choosing the No Upfront option. The entire value of the No Upfront RI is spread across every hour in the term and you will be billed for every hour in the term, regardless of usage. The Partial Upfront payment option is a hybrid of the All Upfront and No Upfront options. You make a small upfront payment, and you are billed a low hourly rate for every hour in the term regardless of usage.

Security

Q: How do I control access to Amazon ElastiCache?

When not using VPC, Amazon ElastiCache allows you to control access to your clusters through Cache Security Groups. A Security Group acts like a firewall, controlling network access to your cluster. By default, network access is turned off to your clusters. If you want your applications to access your cluster, you must explicitly enable access from hosts in specific EC2 security groups. This process is called ingress.

To allow network access to your cluster, create a Security Group and link the desired EC2 security groups (which in turn specify the EC2 instances allowed) to it. The Security Group can be associated with your cluster at the time of creation, or using the "Modify" option on the AWS Management Console.

Please note that IP-range based access control is currently not enabled for clusters. All clients to a cluster must be within the EC2 network, and authorized via security groups as described above.

When using VPC, please see here for more information.

Q: Can programs running on servers in my own data center access Amazon ElastiCache?

Yes. You can access an Amazon ElastiCache cluster from an application running in your data center providing there is connectivity between your VPC and the data center either through VPN or Direct Connect. The details are described here.

Q: Can programs running on EC2 instances in a VPC access Amazon ElastiCache?

Yes, EC2 instances in a VPC can access Amazon ElastiCache if the ElastiCache cluster was created within the VPC. Details on how to create an Amazon ElastiCache cluster within a VPC are given here.

Q: What is Amazon Virtual Private Cloud (VPC) and why may I want to use with Amazon ElastiCache?

Amazon VPC lets you create a virtual networking environment in a private, isolated section of the Amazon Web Services (AWS) cloud, where you can exercise complete control over aspects such as private IP address ranges, subnets, routing tables and network gateways. With Amazon VPC, you can define a virtual network topology and customize the network configuration to closely resemble a traditional IP network that you might operate in your own datacenter.

One of the scenarios where you may want to use Amazon ElastiCache in a VPC is if you want to run a public-facing web application, while still maintaining non-publicly accessible backend servers in a private subnet. You can create a public-facing subnet for your webservers that has access to the Internet, and place your backend infrastructure in a private-facing subnet with no

Internet access. Your backend infrastructure could include RDS DB Instances and an Amazon ElastiCache Cluster providing the in-memory layer. For more information about Amazon VPC, refer to the Amazon Virtual Private Cloud User Guide.

Q: How do I create an Amazon ElastiCache Cluster in VPC?

For a walk through example of creating an Amazon ElastiCache Cluster in VPC, refer to the Amazon ElastiCache User Guide.

Following are the pre-requisites necessary to create a cluster within a VPC:

- You need to have a VPC set up with at least one subnet. For information on creating Amazon VPC and subnets refer to the Getting Started Guide for Amazon VPC.
- You need to have a Subnet Group (for Redis or Memcached) defined for your VPC.
- You need to have a VPC Security Group defined for your VPC (or you can use the default provided).
- In addition, you should allocate adequately large CIDR blocks to each of your subnets so that there are enough spare IP addresses for Amazon ElastiCache to use during maintenance activities such as cache node replacement.

Q: How do I create an Amazon ElastiCache Cluster in an existing VPC?

Creating an Amazon ElastiCache Cluster in an existing VPC is the same as that for a newly created VPC. Here are more details for Redis or Memcached.

Q: How do I connect to an ElastiCache Node in VPC?

Amazon ElastiCache Nodes, deployed within a VPC, can be accessed by EC2 Instances deployed in the same VPC. If these EC2 Instances are deployed in a public subnet with associated Elastic IPs, you can access the EC2 Instances via the internet.

If you want to access Amazon ElastiCache Nodes, deployed within a VPC, from the Internet or from EC2 Instances outside the VPC, please see guidelines for Redis or Memcached.

ElastiCache ensures that both the DNS name and the IP address of the cache node remain the same when cache nodes are recovered in case of failure.

Q: What is a Subnet Group and why do I need one?

A Subnet Group is a collection of subnets that you must designate for your Amazon ElastiCache Cluster in a VPC. A Subnet Group is created using the Amazon ElastiCache Console. Each Subnet Group should have at least one subnet. Amazon ElastiCache uses the Subnet Group to select a subnet. The IP Addresses from the selected subnet are then associated with the Node Endpoints. Furthermore, Amazon ElastiCache creates and associates Elastic Network Interfaces to nodes with the previously mentioned IP addresses.

Please note that, we strongly recommend you use the DNS Names to connect to your nodes as the underlying IP addresses can change (e.g., after cache node replacement).

Q: Can I change the Subnet Group of my ElastiCache Cluster?

An existing Subnet Group can be updated to add more subnets either for existing Availability Zones or for new Availability Zones added since the creation of the ElastiCache Cluster. However, changing the Subnet Group of a deployed cluster is not currently allowed.

Q: How is using Amazon ElastiCache inside a VPC different from using it outside?

The basic functionality of Amazon ElastiCache remains the same whether VPC is used or not. Amazon ElastiCache manages automatic failure detection, recovery, scaling, auto discovery, and software patching whether your ElastiCache Cluster is inside or outside a VPC.

Within a VPC, nodes of an ElastiCache cluster only have a private IP address (within a subnet that you define). Outside of a VPC, the access to the ElastiCache cluster can be controlled using Security Groups as described here.

Q: Can I move my existing ElastiCache Cluster from outside VPC into my VPC?

No, you cannot move an existing Amazon ElastiCache Cluster from outside VPC into a VPC. You will need to create a new Amazon ElastiCache Cluster inside the VPC.

Q: Can I move my existing ElastiCache Cluster from inside VPC to outside VPC?

Currently, direct migration of ElastiCache Cluster from inside to outside VPC is not supported. You will need to create a new Amazon ElastiCache Cluster outside VPC.

Q: How do I control network access to my cluster?

Amazon ElastiCache allows you to control access to your cluster and therefore the nodes using Security Groups in non-VPC deployments. A Security Group acts like a firewall controlling network access to your node. By default, network access is turned off to your nodes. If you want your applications to access your node, you can set your Security Group to allow access from EC2 Instances with specific EC2 Security Group membership or IP ranges. This process is called ingress. Once ingress is configured for a Security Group, the same rules apply to all nodes associated with that Security Group. Security Groups can be configured with the "Security Groups" section of the Amazon ElastiCache Console or using the Amazon ElastiCache APIs.

In VPC deployments, access to your nodes is controlled using the VPC Security Group and the Subnet Group. The VPC Security Group is the VPC equivalent of the Security Group.

Q: What precautions should I take to ensure that my ElastiCache Nodes in VPC are accessible by my application?

You are responsible for modifying routing tables and networking ACLs in your VPC to ensure that your ElastiCache Nodes are reachable from your client instances in the VPC. To learn more see the Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached Documentation.

Q: Can I use Security Groups to configure the clusters that are part of my VPC?

No, Security Groups are not used when operating in a VPC. Instead they are used in the non VPC settings. When creating a cluster in a VPC you will need to use VPC Security Groups.

Q: Can I associate a regular EC2 security group with a cluster that is launched within a VPC?

No, you can only associate VPC security groups that are part of the same VPC as your cluster.

Q: Can nodes of an ElastiCache cluster span multiple subnets?

Yes, nodes of an Amazon ElastiCache cluster can span multiple subnets as long as the subnets are part of the same Subnet Group that was associated with the ElastiCache Cluster at creation time.

Parameter Groups

Q: What are Parameter Groups? How are they helpful?

A Parameter Group acts as a "container" for engine configuration values that can be applied to one or more clusters. If you create a cluster without specifying a Parameter Group, a default Parameter Group is used. This default group contains engine defaults and Amazon ElastiCache system defaults optimized for the cluster you are running. However, if you want your cluster to run with your custom-specified engine configuration values, you can simply create a new Parameter Group, modify the desired parameters, and modify the cluster to use the new Parameter Group. Once associated, all clusters that use a particular Parameter Group get all the parameter updates to that Parameter Group. For more information on configuring Parameter Groups, please refer to the Amazon ElastiCache for Redis or Amazon ElastiCache for Memcached User Guide.

Q: How do I choose the right configuration parameters for my Cluster(s)?

Amazon ElastiCache by default chooses the optimal configuration parameters for your cluster taking into account the Node Type's memory/compute resource capacity. However, if you want to change them, you can do so using our configuration management APIs. Please note that changing configuration parameters from recommended values can have unintended effects, ranging from degraded performance to system crashes, and should only be attempted by advanced users who wish to assume these risks. For more information on changing parameters, please refer to the Amazon ElastiCache User Guide.

Q: How do I see the current setting for my parameters for a given Parameter Group?

You can use the AWS Management Console, Amazon ElastiCache APIs, or Command Line Tools to see information about your Parameter Groups and their corresponding parameter settings.

Memcached

Features

Q: What can I cache using Amazon ElastiCache for Memcached?

You can cache a variety of objects using the service, from the content in persistent data stores (such as Amazon RDS, DynamoDB, or self-managed databases hosted on EC2) to dynamically generated web pages (with Nginx for example), or transient session data that may not require a persistent backing store. You can also use it to implement high-frequency counters to deploy admission control in high volume web applications.

Q: Can I use Amazon ElastiCache for Memcached with an AWS persistent data store such as Amazon RDS or Amazon DynamoDB?

Yes, Amazon ElastiCache is an ideal front-end for data stores like Amazon RDS or Amazon DynamoDB, providing a high-performance middle tier for applications with extremely high request rates and/or low latency requirements.

Q: I use Memcached today. How do I migrate to Amazon ElastiCache?

Amazon ElastiCache is protocol-compliant with Memcached. Therefore, you can use standard Memcached operations like get, set, incr and decr in exactly the same way as you would in your existing Memcached deployments. Amazon ElastiCache supports both the text and binary protocols. It also supports most of the standard stats results, which can also be viewed as graphs via CloudWatch. As a result, you can switch to using Amazon ElastiCache without recompiling or re-linking your applications - the libraries you use will continue to work. To configure the cache servers your application accesses, all you will need to do is to update your

application's Memcached config file to include the endpoints of the servers (nodes) we provision for you. You can simply use the "Copy Node Endpoints" option on the AWS Management Console or the "DescribeCacheClusters" API to get a list of the endpoints. As with any migration process, we recommend thorough testing of your new Amazon ElastiCache deployment before completing the cut over from your current solution.

You can access Amazon ElastiCache cluster in an Amazon VPC from either Amazon EC2 network, or from your own Data Center, please refer to Amazon VPC access patterns for more details.

Amazon ElastiCache uses DNS entries to allow client applications to locate servers (nodes). The DNS name for a node remains constant, but the IP address of a node can change over time, for example, when nodes are auto replaced after a failure on a non-VPC installation. See this FAQ for recommendations to deal with node failures.

Configuration and Scaling

Q: How do I select an appropriate Node Type for my application?

Though there is no precise answer for this question, with Amazon ElastiCache, you don't need to worry about getting the number of nodes exactly right, as you can very easily add or remove nodes later. The following two inter-related aspects could be considered for the choice of your initial configuration:

- The total memory required for your data to achieve your target cache-hit rate, and
- The number of nodes required to maintaining acceptable application performance without overloading the database backend in the event of node failure(s).

The amount of memory required is dependent upon the size of your data set and the access patterns of your application. To improve fault tolerance, once you have a rough idea of the total memory required, divide that memory into enough nodes such that your application can survive the loss of one or two nodes. For example, if your memory requirement is 13GB, you may want to use two cache.m4.large nodes instead of using one cache.m4.xlarge node. It is important that other systems such as databases will not be overloaded if the cache-hit rate is temporarily reduced during failure recovery of one or more of nodes. Please refer to the Amazon ElastiCache User Guide for more details.

Q: Can a cluster span multiple Availability Zones?

Yes. When creating a cluster or adding nodes to an existing cluster, you can chose the availability zones for the new nodes. You can either specify the requested amount of nodes in each availability zones or select “spread nodes across zones”. If the cluster is in VPC, nodes can only be placed in availability zones that are part of the selected cache subnet group. For additional details please see ElastiCache VPC documentation.

Q: How many nodes can I run per region in Amazon ElastiCache Memcached?

You can run a maximum of 300 nodes per region. If you need more nodes, please fill in the ElastiCache Limit Increase Request form.

Q: How does Amazon ElastiCache respond to node failure?

The service will detect the node failure and react with the following automatic steps:

- Amazon ElastiCache will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. For VPC installations, ElastiCache will ensure that both the DNS name and the IP address of the node remain the same when nodes are recovered in case of failure.

For non-VPC installations, ElastiCache will ensure that the DNS name of a node is unchanged; however, the underlying IP address of the node can change.
- If you associated an SNS topic with your cluster, when the new node is configured and ready to be used, Amazon ElastiCache will send an SNS notification to let you know that node recovery occurred. This allows you to optionally arrange for your applications to force the Memcached client library to attempt to reconnect to the repaired nodes. This may be important, as some Memcached libraries will stop using a server (node) indefinitely if they encounter communication errors or timeouts with that server.

Q: If I determine that I need more memory to support my application, how do I increase the total memory with Amazon ElastiCache?

You could add more nodes to your existing Memcached Cluster by using the "Add Node" option on "Nodes" tab for your Cache Cluster on the AWS Management Console or calling the ModifyCacheCluster API.

Compatibility

Q: How does Amazon ElastiCache interact with other Amazon Web Services?

Amazon ElastiCache is ideally suited as a front-end for Amazon Web Services like Amazon RDS and Amazon DynamoDB, providing extremely low latency for high performance applications and offloading some of the request volume while these services provide long lasting data durability. The service can also be used to improve application performance in conjunction with Amazon EC2 and EMR.

Q: Is Amazon ElastiCache better suited to any specific programming language?

Memcached client libraries are available for many, if not all of the popular programming languages. If you encounter any issues with specific Memcached clients when using Amazon ElastiCache, please engage us via the Amazon ElastiCache community forum.

Q: What popular Memcached libraries are compatible with Amazon ElastiCache?

Amazon ElastiCache does not require specific client libraries and works with existing Memcached client libraries without recompilation or application re-linking (Memcached 1.4.5 and later); examples include libMemcached (C) and libraries based on it (e.g. PHP, Perl, Python), spyMemcached (Java) and fauna (Ruby).

Auto Discovery

Q: What is Auto Discovery and what can I do with it?

Auto Discovery is a feature that saves developers time and effort, while reducing complexity of their applications. Auto Discovery enables automatic discovery of cache nodes by clients when they are added to or removed from an Amazon ElastiCache cluster. Until now to handle cluster membership changes, developers must update the list of cache node endpoints manually. Depending on how the client application is architected, typically a client initialization, by shutting down the application and restarting it, is needed resulting in downtime. Through Auto Discovery we are eliminating this complexity. With Auto Discovery, in addition to being backwards protocol-compliant with the Memcached protocol, Amazon ElastiCache provides clients with information on cache cluster membership. A client capable of processing the additional information reconfigures itself, without any initialization, to use the most current nodes of an Amazon ElastiCache cluster.

Q: How does Auto Discovery work?

An Amazon ElastiCache cluster can be created with nodes that are addressable via named endpoints. With Auto Discovery the Amazon ElastiCache cluster is also given a unique Configuration Endpoint which is a DNS Record that is valid for the lifetime of the cluster. This DNS Record contains the DNS Names of the nodes that belong to the cluster. Amazon ElastiCache will ensure that the Configuration Endpoint always points to at least one such "target" node. A query to the target node then returns endpoints for all the nodes of the cluster in question. After this, you can connect to the cluster nodes just as before and use the Memcached protocol commands such as get, set, incr and decr. For more details, see here. To use Auto Discovery, you will need an Auto Discovery capable client. Auto Discovery clients for .Net , Java and PHP are available for download from the Amazon ElastiCache console. Upon initialization, the client will automatically determine the current members of the Amazon ElastiCache cluster using the Configuration Endpoint. When you make changes to your cache cluster by adding or removing nodes or if a node is replaced upon failure, the Auto Discovery client automatically determines the changes and you do not need to initialize your clients manually.

Q: How can I get started using Auto Discovery?

To get started, download the Amazon ElastiCache Cluster Client by clicking the "Download ElastiCache Cluster Client" link on the Amazon ElastiCache console. Before you can download, you must have an Amazon ElastiCache account; if you do not already have one, you can sign up from the Amazon ElastiCache detail page. After you download the client, you can begin setting up and activating your Amazon ElastiCache cluster by visiting the Amazon ElastiCache console. More details can be found here.

Q: If I continue to use my own Memcached clients with my ElastiCache cluster – will I be able to get this feature?

No, you will not get the Auto Discovery feature with the existing Memcached clients. To use the Auto Discovery feature a client must be able to use a Configuration Endpoint and determine the cluster node endpoints. You may either use the Amazon ElastiCache Cluster Client or extend your existing Memcached client to include the Auto Discovery command set.

Q: What are the minimum hardware / software requirements for Auto Discovery?

To take advantage of Auto Discovery, an Auto Discovery capable client must be used to connect to an Amazon ElastiCache Cluster. Amazon ElastiCache currently supports Auto Discovery capable clients for .Net , Java and PHP. These can be downloaded from the Amazon ElastiCache console. Our customers can create clients for any other language by building upon the popular Memcached clients available.

Q: How do I modify or write my own Memcached client to support auto-discovery?

You can take any Memcached Client Library and add support for Auto Discovery. If you would like to add or modify your own client to enable Auto Discovery, please refer to the Auto Discovery command set documentation.

Q: Can I continue to work with my existing Memcached client if I don't need Auto-discovery?

Yes, Amazon ElastiCache is still Memcached protocol compliant and does not require you to change your clients. However, for taking advantage of auto-discovery feature, we had to enhance the Memcached client capabilities. If you choose to not use the Amazon ElastiCache Cluster Client, you can continue to use your own clients or modify your own client library to understand the auto-discovery command set.

Q: Can I have heterogeneous clients when using Auto Discovery?

Yes, the same Amazon ElastiCache cluster can be connected through an Auto Discovery capable Client and the traditional Memcached client at the same time. Amazon ElastiCache remains 100% Memcached compliant.

Q: Can I stop using Auto Discovery?

Yes, you can stop using Auto Discovery anytime. You can disable Auto Discovery by specifying the mode of operation during the Amazon ElastiCache Cluster client initialization. Also, since Amazon ElastiCache continues to support Memcached 100% you may use any Memcached protocol-compliant client as before.

Engine Version Management

Q: Can I control if and when the engine version powering Amazon ElastiCache Cluster is upgraded to new supported versions?

Amazon ElastiCache allows you to control if and when the Memcached protocol-compliant software powering your cluster is upgraded to new versions supported by Amazon ElastiCache. This provides you with the flexibility to maintain compatibility with specific Memcached versions, test new versions with your application before deploying in production, and perform version upgrades on your own terms and timelines. Version upgrades involve some compatibility risk, thus they will not occur automatically and must be initiated by you. This approach to software patching puts you in the driver's seat of version upgrades, but still offloads the work of patch application to Amazon ElastiCache. You can learn more about version management by reading the FAQs that follow. Alternatively, you can refer to the Amazon ElastiCache User Guide. While Engine Version Management functionality is intended to give you as much control as possible over how patching occurs, we may patch your cluster on your behalf if we determine there is any security vulnerability in the system or cache software.

Q: How do I specify which supported Memcached Version my Cluster should run?

You can specify any currently supported version (minor and/or major) when creating a new cluster. If you wish to initiate an upgrade to a supported engine version release, you can do so using the "Modify" option for your cluster. Simply specify the version you wish to upgrade to via the "Cache Engine Version" field. The upgrade will then be applied on your behalf either immediately (if the "Applied Immediately" option is checked) or during the next scheduled maintenance window for your cluster.

Q: Can I test my cluster against a new version before upgrading?

Yes. You can do so by creating a new cluster with the new engine version. You can point your development/staging application to this cluster, test it and decide whether or not to upgrade your original cluster.

Q: Does Amazon ElastiCache provide guidelines for supporting new Memcached version releases and/or deprecating versions that are currently supported?

Over time, we plan to support additional Memcached versions for Amazon ElastiCache, both major and minor. The number of new version releases supported in a given year will vary based on the frequency and content of the Memcached version releases and the outcome of a thorough vetting of the release by our engineering team.

Q: Which version of the Memcached wire protocol does Amazon ElastiCache support?

Amazon ElastiCache supports the Memcached text and binary protocol of versions 1.6.6, 1.5.16, 1.5.10, 1.4.34, 1.4.33, 1.4.24, 1.4.14, and 1.4.5 of Memcached.

Q: What should I do to upgrade to the latest Memcached version?

You can upgrade your existing Memcached cluster by using the Modify process. When upgrading from an older version of Memcached to Memcached version 1.4.33 or newer, please ensure that your existing parameter max_chunk_size values satisfies conditions needed for slab_chunk_max parameter. Please review upgrade prerequisites here.

Redis

Features

Q: What is Amazon ElastiCache for Redis?

Amazon ElastiCache for Redis is a web service that makes it easy to deploy and run Redis protocol-compliant server nodes in the cloud. The service enables the management, monitoring, and operation of Redis nodes; creation, deletion, and modification of the nodes can be carried out through the Amazon ElastiCache console, the command line interface (CLI), or the web service APIs. Amazon ElastiCache for Redis supports high-availability configurations, including Redis cluster-mode enabled and cluster-mode disabled with auto-failover from primary to replica.

Q: Is Amazon ElastiCache for Redis protocol-compliant with open source Redis?

Yes, Amazon ElastiCache for Redis is designed to be protocol-compliant with open source Redis. Code, applications, drivers and tools a customer uses today with their existing standalone Redis data store will continue to work with Amazon ElastiCache for Redis and no code changes will be required for existing Redis deployments migrating to Amazon ElastiCache for Redis unless noted. We currently support Redis 6.2.5, 6.0.5, 5.0.6, 5.0.5, 5.0.4, 5.0.3, 5.0.0, 4.0.10, 3.2.10, 3.2.6, 3.2.4, 2.8.24, 2.8.23, 2.8.22, 2.8.21, 2.8.19, 2.8.6, and 2.6.13.

Q: How much does Amazon ElastiCache for Redis cost?

Please see our pricing page for current pricing information.

Q: What are Amazon ElastiCache for Redis nodes, clusters, and replication groups?

An Amazon ElastiCache for Redis node is the smallest building block of an Amazon ElastiCache for Redis deployment. Each Amazon ElastiCache for Redis node supports the Redis protocol and has its own DNS name and port. Multiple types of Amazon ElastiCache for Redis nodes are supported, each with varying amount of CPU capability, and associated memory. An Amazon ElastiCache for Redis node may take on a primary or a read replica role. A primary node can be replicated to multiple read replica nodes. An Amazon ElastiCache for Redis cluster is a collection of one or more Amazon ElastiCache for Redis nodes; the primary node will be in the primary cluster and the read replica node will be in a read replica cluster. A cluster manages a logical key space, where each node is responsible for a part of the key space. Most of your management operations will be performed at the cluster level. An Amazon ElastiCache for Redis replication group encapsulates the primary and read replica clusters for a Redis installation. A replication group will have only one primary cluster and zero or many read replica clusters. All nodes within a replication group (and consequently cluster) will be of the same node type, and have the same parameter and security group settings.

Q: Does Amazon ElastiCache for Redis support Redis persistence?

Amazon ElastiCache for Redis doesn’t support the AOF (Append Only File) feature but you can achieve persistence by snapshotting your Redis data using the Backup and Restore feature. Please see here for details.

Q: How can I migrate from Amazon ElastiCache for Memcached to Amazon ElastiCache for Redis and vice versa?

We currently do not support automatically migrating from Memcached to Redis or vice versa. You may, however, use a Memcached client to read from a Memcached cluster and use a Redis client to write to a Redis cluster. Similarly, you may read from a Redis cluster using a Redis client and use a Memcached client to write to a Memcached cluster. Make sure to consider the differences in data format, and cluster configuration between the two engines.

Q: Does Amazon ElastiCache for Redis support Multi-AZ operation?

Yes, with Amazon ElastiCache for Redis you can create a read replica in another AWS Availability Zone. Upon a failure of a node, we will provision a new node. In scenarios where the primary node fails ElastCache for Redis will automatically promote an existing read replica to the primary role. For more details on how to handle node failures see here.

Q: What options does Amazon ElastiCache for Redis provide in case of node failures?

Amazon ElastiCache for Redis will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. Thus, the DNS name for a Redis node remains constant, but the IP address of a Redis node can change over time. If you have a replication group with one or more read replicas and Multi-AZ is enabled, then in case of primary node failure, Amazon ElastiCache will automatically detect the failure, select a replica and promote it to become the new primary. It will also propagate the DNS so that you can continue to use the primary endpoint and after the promotion it will point to the newly promoted primary. For more details see the Multi-AZ section of this FAQ.

When Redis replication option is selected with Multi-AZ disabled, in case of primary node failure, you will be given the option to initiate a failover to a read replica node. The failover target can be in the same zone or another zone. To failback to the original zone, promote the read replica in the original zone to be the primary. You may choose to architect your application to force the Redis client library to reconnect to the repaired Redis server node. This can help as some Redis libraries will stop using a server indefinitely when they encounter communication errors or timeouts.

Q: How does failover work?

When deploying ElastiCache for Redis with Cluster Mode Disabled, for Multi-AZ enabled replication groups, the failover behavior is described in the Multi-AZ section of this FAQ. If you choose not to enable Multi-AZ, then if Amazon ElastiCache monitors the primary node, and in case the node becomes unavailable or unresponsive, Amazon ElastiCache for Redis will repair the node by acquiring new service resources, and will then redirect the node's existing DNS name to point to the new service resources. Thus, the DNS name for a Redis node remains constant, but the IP address of a Redis node can change over time. However, if the primary node cannot be healed (and your Multi-AZ is disabled) you will have the choice to promote one of the read replicas to be the new primary. See here for how to select a new primary. The DNS record of the primary's endpoint will be updated to point to the promoted read replica node. A read replica node in the original primary's AZ will then be created to be a read replica in the replication group and will follow the new primary.

When deploying ElastiCache for Redis with Cluster Mode Enabled, you are spreading the cache key space across multiple shards. This means that your data and read/write access to that data is spread across multiple Redis nodes across Multiple-AZs (required with Cluster Mode Enabled). The role of primary node will automatically fail over to one of the read replicas. There is no need to create and provision a new primary node, because ElastiCache will handle this transparently. This failover and replica promotion ensure that you can resume writing to the new primary as soon as promotion is complete.

Q: Are my read replicas available during a primary node failure?

Yes, during a primary node failure, the read replicas continue to service requests. After the primary node is restored, either as a healed node or as a promoted read replica, there is a brief period during which the read replicas will not serve any requests as they sync the cache information from the primary.

Q: How do I configure parameters of my Amazon ElastiCache for Redis nodes?

You can configure your Redis installation using a cache parameter group, which must be specified for a Redis cluster. All read replica clusters use the parameter group of their primary cluster. A Redis parameter group acts as a “container” for Redis configuration values that can be applied to one or more Redis primary clusters. If you create a Redis primary cluster without specifying a cache parameter group, a default parameter group is used. This default group contains defaults for the node type you plan to run. However, if you want your Redis primary cluster to run with specified configuration values, you can simply create a new cache parameter group, modify the desired parameters, and modify the primary Redis cluster to use the new parameter group.

Q: Can I access Redis through the Amazon ElastiCache console?

Yes, Redis appears as an Engine option in the Amazon ElastiCache console. You can create a new Redis cache cluster with the Launch Wizard by choosing the Redis engine. You can also modify or delete an existing Redis cluster using the Amazon ElastiCache console.

Q: Can Amazon ElastiCache for Redis clusters be created in an Amazon VPC?

Yes. If your account is a VPC by default account, your Redis clusters will be created within the default VPC associated with your account. Using the Amazon ElastiCache console, you can specify a different VPC when you create your cluster.

Q: How can I secure my Redis cluster?

Amazon ElastiCache for Redis supports two methods to secure your Redis cluster. You can choose between Redis AUTH or Managed Role-Based Access Control (RBAC), which are both opt-in feature and require that encryption-in transit is enabled. Redis AUTH allows you to add a password to secure access to your Redis cluster and is supported in version 3.2.6 onwards. Starting with Redis 6, the RBAC feature enables you to create and manage users and user groups to secure your Redis cluster. You can assign the users to user groups aligned with a specific role (e.g. administrators, human resources, analytics, etc.) that are then deployed to one or more Amazon ElastiCache for Redis replication groups. By doing this, you can establish security boundaries between users using the same Redis replication group or groups and prevent clients from accessing each other’s data. Follow these links to learn more about Redis Auth and RBAC.

Q: How do I upgrade to a newer engine version?

You can easily upgrade to a newer engine version by using the ModifyCacheCluster or ModifyReplicationGroup APIs and specifying your preferred engine version for the EngineVersion parameter. On the Amazon ElastiCache console, you can select a cache cluster or replication group and click “Modify”. In the “Modify Cache Cluster” or “Modify Replication Group” window select your preferred engine version from the available options. The engine upgrade process is designed to make a best effort to retain your existing data and requires Redis replication to succeed. For more details on that see here.

Q: Can I downgrade to an earlier engine version?

No. Downgrading to an earlier engine version is not supported.

Q: How do I scale up to a larger node type or out to more nodes?

In Amazon ElastiCache for Redis you can easily scale up to larger instance types with cluster mode disabled and out to more instances, when using cluster mode enabled.

You can easily scale up to a larger node type by using the ModifyCacheCluster or ModifyReplicationGroup APIs and specifying your preferred node type for the CacheNodeType parameter. On the Amazon ElastiCache console, you can select a cache cluster or replication group and click “Modify”. In the “Modify Cache Cluster” or “Modify Replication Group” window select your preferred node type from the available options. The scale up process is designed to make a best effort to retain your existing data and requires Redis replication to succeed. For more details on that see here.

Cluster mode allows you to scale horizontally by adding or removing shards as opposed to vertically scaling a single node. Conceptually, horizontal scaling of the cluster is easy to understand on the server-side — a shard is simply added or removed. Each shard has a primary node and up to five read-only replica nodes. Once the new node is ready, the cluster will need to reallocate or balance the key space across the nodes as configured. With ElastiCache for Redis, the re-balance is automatic.

Q: What is the correct metric to use to measure Redis utilization?

Amazon ElastiCache provides two metrics to measure CPU utilization for Amazon ElastiCache for Redis workloads – EngineCPUUtilization and CPUUtilization. The CPUUtilization metric measures the CPU utilization for the instance (node), and EngineCPUUtilization metric measures the utilization at the Redis process level. You need the EngineCPUUtilization metric in addition to the CPUUtilization metric as the main Redis process is single threaded and uses just one CPU of the multiple CPU cores available on an instance. Therefore, the CPUUtilization metric does not provide precise visibility into the CPU utilization rates at the Redis process level.

We recommend that you use both the CPUUtilization and EngineCPUUtilization metrics together to get a detailed understanding of CPU Utilization for your Redis clusters. Both the metrics are available in all AWS regions, and you can access these metric using CloudWatch or via the AWS Management Console.

In addition to CPU utilization, Amazon ElastiCache for Redis adds dynamic network processing to enhanced I/O handling in Redis versions 5.0.3 and above. By utilizing the extra CPU power available in nodes with four or more vCPUs, ElastiCache transparently delivers up to 83% increase in throughput and up to 47% reduction in latency per node. See this blog: Boosting application performance & reducing costs with Amazon Elasticache for Redis

Q: Can I have cross-region replicas with Amazon ElastiCache for Redis?

Yes, you can create cross region replicas using the Global Datastore feature in Amazon ElastiCache for Redis. Global Datastore provides fully managed, fast, reliable and secure cross-region replication. It allows you to write to your Amazon ElastiCache for Redis cluster in one region and have the data available to be read from up to two other cross-region replica clusters, thereby enabling low-latency reads and disaster recovery across regions.

Read Replica

Q: What does it mean to run a Redis node as a Read Replica?

Read Replicas serve two purposes in Redis:

- Failure Handing
- Read Scaling

When you run a cache node with a read replica, the “primary” serves both writes and reads. The read replica acts as a “standby” which is “promoted” in failover scenarios. After failover, the

standby becomes the primary and accepts your cache operations. Read replicas also make it easy to elastically scale out beyond the capacity constraints of a single cache node for read-heavy cache workloads.

Q: When would I want to consider using a Redis read replica?

There are a variety of scenarios where deploying one or more read replicas for a given primary node may make sense. Common reasons for deploying a read replica include:

- Scaling beyond the compute or I/O capacity of a single primary node for read-heavy workloads. This excess read traffic can be directed to one or more read replicas.

- Serving read traffic while the primary is unavailable. If your primary node cannot take I/O requests (e.g. due to I/O suspension for backups or scheduled maintenance), you can direct read traffic to your read replicas. For this use case, keep in mind that the data on the read replica may be "stale" since the primary Instance is unavailable. The read replica can also be used to restart a failed primary warmed up.

- Data protection scenarios; in the unlikely event or primary node failure or that the Availability Zone in which your primary node resides becomes unavailable, you can promote a read replica in a different Availability Zone to become the new primary.

Q: How do I deploy a read replica node for a given primary node?

You can create a read replica in minutes using a CreateReplicationGroup API or a few clicks of the Amazon ElastiCache Management Console. When creating a replication group, you specify the MasterCacheClusterIdentifier. The MasterCacheClusterIdentifier is the cache cluster Identifier of the "primary" cache cluster from which you wish to replicate. You then create the read replica cluster within the replication group by calling the CreateCacheCluster API specifying the ReplicationGroupIdentifier and the CacheClusterIdentifier of the primary cluster. As with a standard cache cluster, you can also specify the Availability Zone. When you initiate the creation of a read replica, Amazon ElastiCache takes a snapshot of your primary cache cluster and begins replication. As a result, you will experience a brief I/O suspension on your primary cache cluster as the snapshot occurs. The I/O suspension typically lasts on the order of one minute.

The read replicas are as easy to delete as they are to create; simply use the Amazon ElastiCache Management Console or call the DeleteCacheCluster API (specifying the CacheClusterIdentifier for the read replica you wish to delete).

Q: Can I create both a primary and read replicas at the same time?

Yes. You can create a new cache cluster along with read replicas in minutes using the CreateReplicationGroup API or using the "Create" wizard at the Amazon ElastiCache Management Console and selecting "Multi-AZ Replication". When creating the cluster, specify an identifier, the total number of desired shard in a cluster and read replicas per shard, along with the creation parameters such as node type, engine version, etc. You can also specify the Availability Zone for each shard in the cluster.

Q: How do I connect to my read replica(s)?

You can connect to a read replica just as you would connect to a primary cache node, using the DescribeCacheClusters API or AWS Management Console to retrieve the endpoint(s) for you read replica(s). If you have multiple read replicas, it is up to your application to determine how read traffic will be distributed amongst them. Here are more details:

- Redis (cluster mode disabled) clusters, use the individual Node Endpoints for read operations (In the API/CLI these are referred to as Read Endpoints).

- Redis (cluster mode enabled) clusters, use the cluster's Configuration Endpoint for all operations. You must use a client that supports Redis Cluster (Redis 3.2). You can still read from individual node endpoints (In the API/CLI these are referred to as Read Endpoints).

Q: How many read replicas can I create for a given primary node?

At this time, Amazon ElastiCache allows you to create up to five (5) read replicas for a given primary cache node.

Q: What happens to read replicas if failover occurs?

In the event of a failover, any associated and available read replicas should automatically resume replication once failover has completed (acquiring updates from the newly promoted read replica).

Q: Can I create a read replica of another read replica?

Creating a read replica of another read replica is not supported.

Q: Can I promote my read replica into a “standalone” primary node?

No, this is not supported. Instead, you may snapshot your Amazon ElastiCache for Redis node (you may select the primary or any of the read-replicas). You can then use the snapshot to seed a new Amazon ElastiCache for Redis primary.

Q: Will my read replica be kept up-to-date with its primary node?

Updates to a primary cache node will automatically be replicated to any associated read replicas. However, with Redis’s asynchronous replication technology, a read replica can fall behind its primary cache node for a variety of reasons. Typical reasons include:

- Write I/O volume to the primary cache node exceeds the rate at which changes can be applied to the read replica

- Network partitions or latency between the primary cache node and a read replica

Read replicas are subject to the strengths and weaknesses of Redis replication. If you are using read replicas, you should be aware of the potential for lag between a read replica and its primary cache node, or “inconsistency.” Amazon ElastiCache emits a metric to help you understand the inconsistency.

Q: How do I gain visibility into active read replica(s)?

You can use the standard DescribeCacheClusters API to return a list of all the cache clusters you have deployed (including read replicas), or simply click on the "Cache Clusters" tab of the Amazon ElastiCache Management Console.

Amazon ElastiCache monitors the replication status of your read replicas and updates the Replication State field to Error if replication stops for any reason. You can review the details of the associated error thrown by the Redis engine by viewing the Replication Error field and take an appropriate action to recover from it. You can learn more about troubleshooting replication issues in the "Troubleshooting a Read Replica problem" section of the Amazon ElastiCache User Guide. If a replication error is fixed, the Replication State changes to Replicating.

Amazon ElastiCache allows you to gain visibility into how far a read replica has fallen behind its primary through the Amazon CloudWatch metric ("Replica Lag") available via the AWS Management Console or Amazon CloudWatch APIs.

Q: My read replica has fallen significantly behind its primary node. What should I do?

As discussed in the previous questions, “inconsistency” or lag between a read replica and its primary cache node is common with Redis asynchronous replication. If an existing read replica has fallen too far behind to meet your requirements, you can reboot it. Keep in mind that replica lag may naturally grow and shrink over time, depending on your primary cache node’s steady-state usage pattern.

Q: How do I delete a read replica? Will it be deleted automatically if its primary node is deleted?

You can easily delete a read replica with a few clicks of the AWS Management Console or by using DeleteCacheCluster, or DecreaseReplicaCount APIs. If you want to delete the read replica in addition to the primary cache node, you must use the DeleteReplicationGroup API or AWS Management Consol e.

Q: How much do read replicas cost? When does billing begin and end?

A read replica is billed as a standard cache node and at the same rates. Just like a standard cache node, the rate per “Cache Node hour” for a read replica is determined by the cache node class of the read replica – please see Amazon ElastiCache detail page for up-to-date pricing. You are not charged for the data transfer incurred in replicating data between your primary cache node and read replica. Billing for a read replica begins as soon as the read replica has been successfully created (i.e. when status is listed as “active”). The read replica will continue being billed at standard Amazon ElastiCache cache node hour rates until you issue a command to delete it.

Q: What happens during failover and how long does it take?

Initiated failover is supported by Amazon ElastiCache so that you can resume cache operations as quickly as possible. When failing over, Amazon ElastiCache simply flips the DNS record for your cache node to point at the read replica, which is in turn promoted to become the new primary. We encourage you to follow best practices and implement cache node connection retry at the application layer. Typically, start-to-finish, steps 1 to 5 below, complete within 6 minutes.

These are the automatic failover events, listed in the order of occurrence :

1. Replication group message: Test Failover API called for node group <node-group-id>

2. Cache cluster message: Failover from primary node <primary-node-id> to replica node <node-id> completed

3. Replication group message: Failover from primary node <primary-node-id> to replica node <node-id> completed

4. Cache cluster message: Recovering cache nodes <node-id>

5. Cache cluster message: Finished recovery for cache nodes <node-id>

Q: Can I create a read replica in another region as my primary?

No. Your read replica may only be provisioned in the same or different Availability Zone of the same Region as your cache node primary.

You can, however, use the Global Datastore for Redis to work with fully managed, fast, reliable, and secure replication across AWS Regions. Using this feature, you can create cross-Region read replica clusters for ElastiCache for Redis to enable low-latency reads and disaster recovery across AWS Regions.

Q: Can I see which Availability Zone my primary is currently located in?

Yes, you can gain visibility into the location of the current primary by using the AWS Management Console or DescribeCacheClusters API.

After failover, my primary is now located in a different Availability Zone than my other AWS resources (e.g. EC2 instances).

Q: Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same Region. In addition, you may want to consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone failure.

Q: Can I add and remove read replica nodes for my Redis Cluster environment?

Yes. You can add a remove replica across one or more shards in a Redis Cluster environment. The cluster continues to stay online and serve incoming I/O during this operation.

Multi-AZ

Q: What is Multi-AZ for an Amazon ElastiCache for Redis replication group?

An Amazon ElastiCache for Redis replication group consists of a primary and up to five read replicas. If Multi-AZ is enable then at least one replica is required per primary. See AutoFailover. Redis asynchronously replicates the data from the primary to the read replicas. During certain types of planned maintenance, or in the unlikely event of Amazon ElastiCache node failure or Availability Zone failure, Amazon ElastiCache will automatically detect the failure of a primary, select a read replica, and promote it to become the new primary. Amazon ElastiCache also propagates the DNS changes of the promoted read replica, so if your application is writing to the primary node endpoint, no endpoint change will be needed.

Q: What are the benefits of using Multi-AZ and when should I use it?

The main benefits of running your Amazon ElastiCache for Redis in Multi-AZ mode are enhanced availability and smaller need for administration. If an Amazon ElastiCache for Redis primary node failure occurs, the impact on your ability to read/write to the primary is limited to the time it takes for automatic failover to complete. When Multi-AZ is enabled, Amazon ElastiCache node failover is automatic and requires no administration. You no longer need to monitor your Redis nodes and manually initiate a recovery in the event of a primary node disruption.

Q: How does Multi-AZ work?

You can use Multi-AZ if you are using Amazon ElastiCache for Redis and have a replication group consisting of a primary node and one or more read replicas. If the primary node fails,

Amazon ElastiCache will automatically detect the failure, select one from the available read replicas, and promote it to become the new primary. Amazon ElastiCache will propagate the DNS changes of the promoted replica so that your application can keep writing to the primary endpoint. Amazon ElastiCache will also spin up a new node to replace the promoted read replica in the same Availability Zone of the failed primary. In case the primary failed due to temporary Availability Zone disruption, the new replica will be launched once that Availability Zone has recovered.

Q: Can I have replicas in the same Availability Zone as the primary?

Yes. Note that placing both the primary and the replica(s) in the same Availability Zone will not make your Amazon ElastiCache for Redis replication group resilient to an Availability Zone disruption. Additionally this well not be allowed if Multi-AZ is turned on.

Q: What events would cause Amazon ElastiCache to fail over to a read replica?

Amazon ElastiCache will failover to a read replica in the event of any of the following:

- Loss of availability in primary's Availability Zone
- Loss of network connectivity to primary
- Compute unit failure on primary

Q: When should I use Multi-AZ?

Using Redis replication in conjunction with Multi-AZ provides increased availability and fault tolerance. Such deployments are a natural fit for use in production environments.

Q: How do I create an Amazon ElastiCache for Redis replication group with Multi-AZ enabled?

You can create an Amazon ElastiCache for Redis primary and read replicas by clicking Launch Cache Cluster on the Amazon ElastiCache Management Console. You can also do so by calling the CreateReplicationGroup API. For existing replication groups (Redis 2.8.24, 2.8.23, 2.8.22, 2.8.21, 2.8.19, and 2.8.6), you can enable Multi-AZ by choosing a replication group and clicking Modify on the Amazon ElastiCache Management Console or by using the ModifyReplicationGroup API. Switching a replication group to Multi-AZ is not disruptive to your Redis data and does not interfere your nodes' ability to serve requests.

Q: Which read replica will be promoted in case of a primary node failure?
If there are more than one read replicas, the read replica with the smallest asynchronous replication lag to the primary will be promoted.

Q: How much does it cost to use Multi-AZ?

Multi-AZ is free of charge. You only pay for the Amazon ElastiCache nodes that you use.

Q: What are the performance implications of Multi-AZ?

Amazon ElastiCache currently uses the Redis engine's native, asynchronous replication and is subject to its strengths and limitations. In particular, when a read replica connects to a primary for the first time, or if the primary changes, the read replica does a full synchronization of the data from the primary, imposing load on itself and the primary. For additional details regarding Redis replication please see here.

Q: What node types support Multi-AZ?

All available cache node types in Amazon ElastiCache support Multi-AZ except T1 node family.

Q: Will I be alerted when automatic fail over occurs?

Yes, Amazon ElastiCache will create an event to inform you that automatic failover occurred. You can use the DescribeEvents API to return information about events related to your Amazon ElastiCache node, or click the Events section of the Amazon ElastiCache Management Console.

Q: After failover, my primary is now located in a different Availability Zone than my other AWS resources (for example, EC2 instances). Should I be concerned about latency?

Availability Zones are engineered to provide low latency network connectivity to other Availability Zones in the same region. You may consider architecting your application and other AWS resources with redundancy across multiple Availability Zones so your application will be resilient in the event of an Availability Zone disruption.

Q: Where can I get more information about Multi-AZ?

For more information about Multi-AZ, see Amazon ElastiCache documentation.

Backup and Restore

Q: What is Backup and Restore?

Backup and Restore is a feature that allows customers to create snapshots of their Amazon ElastiCache for Redis clusters. Amazon ElastiCache stores the snapshots, allowing users to subsequently use them to restore Redis clusters.

Q: What is a snapshot?

A snapshot is a copy of your entire Redis cluster at a specific moment.

Q: Why would I need snapshots?

Creating snapshots can be useful in case of data loss caused by node failure, as well as the unlikely event of a hardware failure. Another common reason to use backups is for archiving purposes. Snapshots are stored in Amazon S3, which is a durable storage, meaning that even a power failure won't erase your data.

Q: What can I do with a snapshot?

You can use snapshots to warm start an Amazon ElastiCache for Redis cluster with preloaded data.

Q: How does Backup and Restore work?

When a backup is initiated, Amazon ElastiCache will take a snapshot of a specified Redis cluster that can later be used for recovery or archiving. You can initiate a backup anytime you choose or set a recurring daily backup with retention period of up to 35 days.

When you choose a snapshot to restore, a new Amazon Amazon for Redis cluster will be created and populated with the snapshot's data. This way you can create multiple ElastiCache for Redis clusters from a specified snapshot.

Q: Where are my snapshots stored?

The snapshots are stored in S3.

Q: How can I get started using Backup and Restore?

You can select to use the Backup and Restore feature through the AWS Management Console, through the Amazon ElastiCache APIs (CreateCacheCluster, ModifyCacheCluster,

CreateReplicationGroup, and ModifyReplicationGroup API's) and CLI. You can deactivate and reactivate the feature anytime you choose.

Q: How do I specify which Redis cluster and node to backup?

Backup and Restore creates snapshots on a cluster basis. Users can specify which ElastiCache for Redis cluster to backup through the AWS Management Console, CLI or through the CreateSnapshot API. In a Replication Group, you can choose to backup the primary or any of the read-replica clusters. We recommend users enable backup on one of the read-replicas, mitigating any latency effect on the Redis primary.

Q: How can I specify when a backup will take place?

Through the AWS Management Console, CLI or APIs you can specify when to start a single backup or a recurring backup. Users are able to:

- Take a snapshot right now (through "Backup" console button in the "Redis" tab, or CreateSnapshot API)
- Set up an automatic daily backup. The backup will take place during your preferred backup window. You can set that up through Creating/Modifying cluster via console or the CreateCacheCluster, ModifyCacheCluster or ModifyReplicationGroup API's.

Q: What is a backup window and why do I need it?

The preferred backup window is the user-defined period of time during which your Amazon ElastiCache for Redis cluster backup will start. This is helpful if you want to backup at a certain time of day or to refrain from backups during a particularly high-utilization period.

Q: What is the performance impact of taking a snapshot?

While taking a snapshot, you may encounter increased latencies for a brief period at the node. Snapshots use Redis's built-in BGSAVE and are subject to its strengths and limitations. In particular, the Redis process forks and the parent continues to serve requests while the child saves the data on disk and then exits. The forking increases the memory usage for the duration of the snapshot generation. When this memory usage exceeds that of the available memory of the cache node, swapping can get triggered, further slowing down the node. For this reason, we recommend generating snapshots on one of the read replicas (instead of the primary). Also, we suggest setting the reserved-memory parameter to minimize swap usage. See here for more details.

Q: Can I create a snapshot from an Amazon ElastiCache for Redis read replica?

Yes. In a Replication Group with cluster mode disabled, creating a snapshot from a read replica is the best way to back up your data while minimizing performance impact. In a Replication Group with cluster mode enabled, you cannot choose to back up the primary or any of the read-replica clusters.

Q: In what regions is the Backup and Restore feature available?

Backup and Restore feature is available in all regions where ElastiCache service is available.

Q: Can I export Amazon ElastiCache for Redis snapshots to an S3 bucket owned by me?

Yes. You can export your Amazon ElastiCache for Redis snapshots to an authorized S3 bucket in the same region as your cluster. For more details on exporting snapshots and setting the required permissions, please refer to this.

Q: Can I copy snapshots from one region to another?

Yes. You must first copy your snapshot into an authorized S3 bucket of your choice in the same region and then use the S3 PUT object- Copy API to copy it to a bucket in another region. For more details on copying S3 objects, please see this.

Q: I have multiple AWS accounts using Amazon ElastiCache for Redis. Can I use Amazon ElastiCache snapshots from one account to warm start an Amazon ElastiCache for Redis cluster in a different one?

Yes. You must first copy your snapshot into an authorized S3 bucket of your choice in the same region and then grant cross-account bucket permissions to the other account. For more details on S3 cross-account permissions, please see this. Finally, specify the S3 location of your RDB file during cluster creation through the Launch Cache Cluster Wizard in the console or through the CreateCacheCluster API.

Q: How much does it cost to use Backup and Restore?

Amazon ElastiCache provides storage space for one snapshot free of charge for each active Amazon ElastiCache for Redis cluster. Additional storage will be charged based on the space used by the snapshots with $0.085/GB every month (same price in all regions). Data transfer for using the snapshots is free of charge.

Q: What is the retention period?

Retention period is the time span during which the automatic snapshots are retained. For example, if a retention period is set for 5, a snapshot that was taken today will be retained for 5 days before being deleted. You can choose to copy one or more automatic snapshots to store them as manual so that they won't be deleted after the retention period is over.

Q: How do I manage the retention of my automated snapshots?

You can use the AWS Management Console or ModifyCreateCluster or ModifyReplicationGroup API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you desire to turn off automated backups altogether, you can do so by setting the retention period to 0 (not recommended).

Q: What happens to my snapshots if I delete my Amazon ElastiCache for Redis cluster?

When you delete an Amazon ElastiCache for Redis cluster, your manual snapshots are retained. You will also have an option to create a final snapshot before the cluster is deleted. Automatic cache snapshots are not retained.

Q: What cache nodes types support backup and restore capability?

All Amazon ElastiCache for Redis instance node types besides t1.micro support backup and restore:

Current Generation Cache Nodes:

- cache.m4.large
- cache.m4.xlarge

- cache.m4.2xlarge
- cache.m4.4xlarge
- cache.m4.10xlarge
- cache.m5.large
- cache.m5.xlarge
- cache.m5.2xlarge
- cache.m5.4xlarge
- cache.m5.12xlarge
- cache.m5.24xlarge
- cache.m6g.large
- cache.m6g.xlarge
- cache.m6g.2xlarge
- cache.m6g.4xlarge
- cache.m6g.8xlarge
- cache.m6g.12xlarge
- cache.m6g.16xlarge
- cache.r4.large
- cache.r4.xlarge
- cache.r4.2xlarge
- cache.r4.4xlarge
- cache.r4.8xlarge
- cache.r4.16xlarge
- cache.r5.large
- cache.r5.xlarge
- cache.r5.2xlarge
- cache.r5.4xlarge
- cache.r5.12xlarge
- cache.r5.24xlarge
- cache.r6g.large

- cache.r6g.xlarge
- cache.r6g.2xlarge
- cache.r6g.4xlarge
- cache.r6g.8xlarge
- cache.r6g.12xlarge
- cache.r6g.16xlarge
- cache.t2.medium
- cache.t2.small
- cache.t2.micro
- cache.t3.medium
- cache.t3.small
- cache.t4g.micro
- cache.t4g.small
- cache.t4g.medium

Current Generation Cache Nodes with data tiering:

- cache.r6gd.xlarge
- cache.r6gd.2xlarge
- cache.r6gd.4xlarge
- cache.r6gd.8xlarge
- cache.r6gd.12xlarge
- cache.r6gd.16xlarge

Previous Generation Nodes:

- cache.m1.small
- cache.m1.medium
- cache.m1.large
- cache.m1.xlarge
- cache.m2.xlarge
- cache.m2.2xlarge
- cache.m2.4xlarge

- cache.m3.medium
- cache.m3.large
- cache.m3.xlarge
- cache.m3.2xlarge
- cache.r3.large
- cache.r3.xlarge
- cache.r3.2xlarge
- cache.r3.4xlarge
- cache.r3.8xlarge
- cache.c1.xlarge

Q: Can I use my own RDB snapshots stored in S3 to warm start an Amazon ElastiCache for Redis cluster?

Yes. You can specify the S3 location of your RDB file during cluster creation through the "Create Cluster" Wizard in the console or through the CreateCacheCluster or CreateReplicationGroup API.

Q: Can I use the Backup and Restore feature if I am running Amazon ElastiCache in a VPC?

Yes.

Redis Cluster

Q: What is ElastiCache for Redis Cluster?

ElastiCache for Redis Cluster allows customers to create and run managed Redis Clusters with multiple shards. It is compatible with open source Redis 3.2.4 onwards and comes with a number of enhancements for a more stable and robust experience (see the "enhanced engine" section below for additional details on these enhancements).

Q: Why would I need a scale out Redis environment?

There are three main scenarios for running a scale out Redis environment. First, if the total memory size of your Redis data exceeds or is projected to exceed the memory capacity of a single VM. Second, if the write throughput of your application to Redis exceeds the capacity of a single VM. Third, if you would like to spread the data across multiple shards so that any potential issue that comes up with a single node will have a smaller impact on the overall Redis environment.

Q: Why would I run my Redis Cluster workload on Amazon ElastiCache?

Amazon ElastiCache provides a fully managed distributed in-memory Redis environment, from provisioning server resources to installing the engine software and applying any configuration parameters you choose. It uses enhancements to the Redis engine developed by Amazon, which results in a more robust and stable experience (see "enhanced engine" section for more details). Once your Redis environment is up and running, the service automates common administrative tasks such as failure detection and recovery, backups and software patching. It also provides a robust Multi-AZ solution with automatic failover. In case of a failure of one or

more primary nodes in your cluster, Amazon ElastiCache will automatically detect the failure and respond by promoting the most up to date replica to primary. This process is automated and does not mandate any manual work on your behalf. Amazon ElastiCache also provides detailed monitoring metrics associated with your ElastiCache nodes, enabling you to diagnose and respond to issues very quickly.

Q: Is ElastiCache for Redis Cluster compatible with open source Redis?

Yes, Amazon ElastiCache for Redis Cluster is compatible with open source Redis 3.2.4 onwards. You can use the open source Redis Cluster clients to access scale-out clusters on ElastiCache for Redis.

Q: What is the upgrade path from current ElastiCache for Redis 2.8.x to ElastiCache for Redis Cluster (version 3.2.4)?

If you are using Redis 3.2 with cluster_mode parameter disabled, you can simply choose the node or cluster you wish to upgrade and modify the engine version. ElastiCache will provision a Redis 3.2.4 cluster and migrate your data to it, while maintaining the endpoint.

If you are using Redis 3.2 with cluster_mode enabled, you can migrate to Redis Cluster by first creating a snapshot of your data using the backup and restore feature. Then, select the created snapshot and click on “Restore Snapshot” to create a Redis 3.2 cluster using the snapshotted data. Finally, update the new endpoint in your client. Note that to use Redis 3.2 in cluster mode you would need to switch to a Redis Cluster client.

Q: Is the pricing for clustered configuration different from non-clustered configuration?

No. Amazon ElastiCache for Redis provides the flexibility of clustered and non-clustered configuration at the same price. Customers can now enjoy enhanced engine functionality within Amazon ElastiCache for Redis and use full feature support for clustered configuration and scalability at the same price.

Q: What is Multi-AZ for ElastiCache for Redis Cluster?

Each shard of an ElastiCache for Redis cluster consists of a primary and up to five read replicas. Redis asynchronously replicates the data from the primary to the read replicas. During certain types of planned maintenance, or in the unlikely event of ElastiCache node failure or Availability Zone failure, Amazon ElastiCache will automatically detect the failure of a primary, select a read-replica, and promote it to become the new primary.

ElastiCache for Redis Cluster provides enhancements and management for Redis 3.x and onwards environments. When running an unmanaged Redis environment, in a case of primary node failure, the cluster relies on a majority of masters to determine and start a failover. If such majority doesn’t exist, the cluster will go into failed state, rejecting any further reads and writes. This could lead to major availability impact on the application, as well as requiring human intervention to manually salvage the cluster. ElastiCache for Redis Multi-AZ capability is built to handle any failover case for Redis Cluster with robustness and efficiency.

Q: How is Multi-AZ in ElastiCache for Redis Cluster different than in ElastiCache for Redis versions 2.8.x?

Redis 3.x and onwards work with intelligent clients that store a node map with all the cluster nodes’ endpoints. During a failover, the client updates the node map with the IP endpoint for the new primary. This provides up to 4x faster failover time than with ElastiCache for Redis 2.8.x.

Q: How does Multi-AZ work for Redis Cluster?

You can use Multi-AZ if you are using an ElastiCache for Redis Cluster with each shard having one or more read-replicas. If a primary node of a shard fails, ElastiCache will automatically detect the failure, select one of the available read-replicas, and promote it to become the new primary. The Redis 3.x and onwards client will update the promoted replica as primary. No application change is required. ElastiCache will also spin up a new node to replace the promoted read-replica in the same Availability Zone of the failed primary. In case the primary failed due to a temporary Availability Zone failure, the new replica will be launched once that Availability Zone has recovered.

Q: What is a backup in ElastiCache for Redis Cluster?

An ElastiCache for Redis Cluster backup is a series of snapshots of the cluster's shards, stored together to keep a copy of your entire Redis data around a certain time frame.

Q: How is a backup in ElastiCache for Redis Cluster different from a snapshot in ElastiCache for Redis?

Since a non-clustered ElastiCache for Redis environment has a single primary node, a backup is a single file which contains a copy of the Redis data. ElastiCache for Redis Cluster can have one or more shards, thus a backup might contain multiple files.

Q: How do I specify which ElastiCache for Redis nodes to backup in each shard?

You cannot manually specify a node to backup within each shard. When initiating a backup, ElastiCache will automatically select the most up-to-date read replica in each shard and take a snapshot of its data.

Q: How does ElastiCache for Redis Cluster Backup and Restore work?

When a backup is initiated, ElastiCache will take a backup of a specified cluster; that backup can later be used for recovery or archiving. The backup will include a copy of each of the cluster's shards, thus a full backup contains a series of files. You can initiate a backup anytime you choose or set a recurring daily backup with retention period of up to 35 days.

When you choose a backup to restore, a new ElastiCache for Redis cluster will be created and populated with the backup's data. Also, you can use this feature for an easy migration path to a managed Redis Cluster experience on ElastiCache. If you are running self-managed Redis on EC2, you can take RDB snapshots or your existing workloads (both Redis Cluster and single-shard Redis) and store them in S3. Then simply provide them as input for creating a sharded Redis Cluster on ElastiCache, and the desired number of shards. ElastiCache will do the rest.

Currently, ElastiCache uses Redis' native mechanism to create and store an RDB file for each shard as the backup.

Q: Is the backup in ElastiCache for Redis Cluster a point-in-time snapshot?

When you initiate a backup, ElastiCache will trigger backups of all of the shards of your cluster at the same time. In rare cases there might be a need to retake a snapshot of one or more nodes that did not complete successfully the first time. ElastiCache does that automatically and no user intervention is required. But in such a case, while each individual snapshot is a point-in-time representation of the node it was taken from, not all the cluster's snapshots would be taken at the same time.

Q: How can I specify when a backup will take place?

Through the AWS Management Console, CLI or APIs you can specify when to start a single backup or a recurring backup. Users are able to:

- Take a backup right now (through “Create Snapshot” console button or CreateSnapshot API)
- Set up an automatic daily backup. The backup will take place during your preferred backup window. You can set that up through Creating/Modifying cluster via console or the CreateCacheCluster, ModifyCacheCluster, CreateReplicationGroup or ModifyReplicationGroup API’s.

Q: Can I use my own RDB snapshots stored in S3 to pre-seed a scale out ElastiCache for Redis Cluster environment?

Yes. You can specify the S3 location of your RDB files during cluster creation through the Create Cluster Wizard in the console or through the CreateReplicationGroup API. ElastiCache will automatically parse the Redis key-space of the RDB snapshot and redistribute it among the shards of the new cluster.

Enhanced Engine

Q: How is the engine within ElastiCache for Redis different from open-source Redis?

The engine within ElastiCache for Redis is fully compatible with open source Redis but also comes with enhancements that improve robustness and stability. Some of the enhancements are:

- More usable memory: You can now safely allocate more memory for your application without risking increased swap usage during syncs and snapshots.
- Improved synchronization: More robust synchronization under heavy load and when recovering from network disconnections. Additionally, syncs are faster as both the primary and replicas no longer use the disk for this operation.
- Smoother failovers: In the event of a failover, your shard now recovers faster as replicas no longer flush their data to do a full re-sync with the primary.

Q: How do I use the enhanced engine?

To use the enhanced engine from the Amazon ElastiCache management console, just select an engine compatible with Redis engine version 2.8.22 or higher when creating a cluster. From that point on you will be using the enhanced engine. You can also use the enhanced engine through the ElastiCache API or AWS CLI by specifying the engine version when running the CreateCacheCluster API.

Q: Do I need to change my application code to use the enhanced engine on ElastiCache?

No. The enhanced engine is fully compatible with open-source Redis, thus you can enjoy its improved robustness and stability without the need to make any changes to your application code.

Q: How much does it cost to use the enhanced engine?

There is no additional charge for using the enhanced engine. As always, you will only be charged for the nodes you use.

Online Cluster Resizing

Q: What is Online Cluster Resizing?

Amazon ElastiCache for Redis provides the ability to add and remove shards from running cluster-mode enabled Redis Cluster. You can dynamically scale-out or scale-in your Redis cluster workloads to adapt to changes in demand. Amazon ElastiCache will resize the cluster by

adding or removing shards and redistributing hash slots uniformly across the new shard configuration, all while the cluster continues to stay online and serve requests.

Q: What are the benefits of using Online Cluster Resizing?

The ability to dynamically scale-out and scale-in a cluster can help you manage application variability and meet oscillating demands. You can right-size your clusters by adding or removing shards to scale performance and in-memory capacity. This feature eliminates the need to overprovision clusters based on peak demand, helps improve efficiency, and reduces cost.

Q: How can I use Online Cluster Resizing?

Online Cluster Resizing is available for cluster-mode enabled Redis Clusters on version 3.2.10 or higher. To reshard your cluster, select the cluster and specify whether you want to add or remove shards. When you resize the cluster to scale-out, Amazon ElastiCache adds shards and migrates slots from existing shards to new shards, in a way such that the slots are uniformly distributed (by count) across shards. Similarly, when resizing the cluster to scale-in, Amazon ElastiCache migrates slots to the remaining shards to uniformly distribute the slots and deletes specified shards.

Q: How long does the Online Cluster Resizing take?

The time taken to resize a cluster depends on multiple factors, such as number of slots that need to be migrated across shards, size of data and incoming request rate on the cluster. The workflow is optimized to parallelize slot migration for faster scale out.

Q: Can the cluster be used while cluster resizing is in progress?

Yes, the cluster continues to stay online and serve incoming requests, while resharding is in progress. However, snapshotting a cluster is not supported when resharding is in progress.

Q: Is there any performance impact of this operation on the cluster?

While Online Cluster Resizing provides the benefits to scale out/in with zero downtime, it is a compute-intensive operation and can increase the latency of your client connection. To reduce the load on the cluster during the operation, we recommend that you follow the best practices (described in the documentation).

Q: How can I track the progress of an online resharding operation?

You can track the progress of resharding by viewing the status of the cluster, shards and nodes. During the operation, the cluster, shards and nodes will stay in “modifying” status. Similarly, when shards are being created, deleted or participating in slot migration, the individual shard status will reflect these statuses to show progress. Additionally, the status of end-to-end operation can also be tracked using the progress indicator for the resharding operation, which indicates percentage completed and provides insight into the remaining time for the operation. Lastly, event messages indicate the progress by describing actions being taken (shard creation, slot migration, etc.) during this operation.

Q: What is the rebalance operation for Amazon ElastiCache for Redis cluster?

The rebalance operation can be used to redistribute slots amongst existing shards to achieve a uniform distribution. This is useful when a cluster is created with manually specified uneven slot distribution or a scale-out/in operation leaves the cluster with uneven distribution. Assuming the slots are identical in their memory and I/O requirements, uniform slot distribution by count is an easy way to load balance across shards.

Q: How does tagging work when a cluster scales-out?

When new nodes are added to scale-out a cluster, the nodes carry the same set of tags that are common across all existing nodes. Additionally, users can modify tags on all nodes and continue to use tagging as before.

Q: Are there any client or application side changes needed to use online cluster resizing?

No. The enhanced slot distribution used in cluster resizing is compliant with Redis cluster client behavior and does not require any application changes. Amazon ElastiCache retains cluster endpoints, enabling you to continue using existing clients without any changes.

Encryption

Q: What does encryption at-rest for Amazon ElastiCache ElastiCache for Redis provide?

Encryption-at-rest provides mechanisms to guard against unauthorized access of your data on the server. When enabled on a replication group, it encrypts the following aspects:

- Disk during sync, backup and swap operations
- Backups stored in Amazon S3

Amazon ElastiCache for Redis offers default (service managed) encryption at rest, as well as ability to use your own symmetric customer managed customer primary keys in AWS Key Management Service (KMS). At-rest encryption can be enabled on a replication group only when it is created. You can read more here.

Q: What does encryption in-transit for Amazon ElastiCache for Redis provide?

The encryption in-transit feature enables you to encrypt all communications between clients and Redis server as well as between the Redis servers (primary and read replica nodes). It is an optional feature and can only be enabled on Redis replication groups when they are created. You can read more here.

Q: How can I use encryption in-transit, at-rest, and Redis AUTH?

Encryption in-transit, encryption at-rest, and Redis AUTH and Managed Role-Based Access Control (RBAC) are all opt-in features. At the time of Redis cluster creation via the console or command line interface, you can specify if you want to enable at-rest and/or in-transit encryption. If you enabled in-transit encryption you can choose to use Redis AUTH or RBAC for added security and access control. Once the cluster is setup with encryption enabled, Amazon ElastiCache seamlessly manages certificate expiration and renewal without requiring any additional action from the application. Additionally, the Redis clients need to support TLS to avail of the encrypted in-transit traffic. If you choose to use Redis AUTH you will need to have Redis 3.2.6 onward, while RBAC requires that you use Redis 6.

Q: Is there an Amazon ElastiCache for Redis client that I need to use when using encryption in-transit, or at-rest?

No. Encryption in-transit requires clients to support TLS. Most of the popular Redis clients (such as Lettuce, Predis, go-Redis) provide support for TLS with some configuration settings. You have to make sure that your Redis client of choice is configured to support TLS and continue to use Amazon ElastiCache for Redis as before.

Q: Can I enable encryption in-transit and encryption at-rest on my existing Amazon ElastiCache for Redis clusters?

No. Encryption in-transit and encryption at-rest support is only available for new clusters and is not supported on existing Amazon ElastiCache for Redis clusters. Amazon ElastiCache for Redis versions 6.2.5, 6.0.5, 5.0.0, 4.0.10, and 3.2.6 support these features.

Q: Is there any action needed to renew certificates?

No. Amazon ElastiCache manages certification expiration and renewal behind the scene. No user action is necessary for ongoing certificate maintenance.

Q: Can I use my certificates for encryption?

No. Currently, Amazon ElastiCache does not provide the ability for you to use your certificates. Amazon ElastiCache manages certificates transparently for you.

Q: Which instance types are supported for encryption in transit and encryption at rest?

All current generation instances are supported for encryption in transit and encryption at rest. For a full list of in-transit encryption conditions see here and at-rest encryption conditions see here.

Q: Are there additional costs for using encryption?

There are no additional costs for using encryption.

Compliance

Q: Which compliance programs does Amazon ElastiCache for Redis support?

Amazon ElastiCache for Redis supports compliance programs such as SOC 1, SOC 2, SOC 3, ISO, MTCS, C5, PCI, HIPAA, and FedRAMP. See AWS Services in Scope by Compliance Program for current list of supported compliance programs.

Q: Is Amazon ElastiCache for Redis PCI compliant?

Yes, the AWS PCI compliance program includes Amazon ElastiCache for Redis as a PCI compliant Service. To learn more, see the following resources:

- Amazon ElastiCache for Redis Compliance page
- AWS PCI Compliance page

To see the current list of compliance programs that Amazon ElastiCache for Redis is in scope for, see AWS Services in Scope by Compliance Program.

Q: Is Amazon ElastiCache for Redis HIPAA eligible?

Yes, Amazon ElastiCache for Redis is a HIPAA Eligible Service and has been added to the AWS Business Associate Addendum (BAA). This means you can use Amazon ElastiCache for Redis to help you process, maintain, and store protected health information (PHI) and power healthcare applications.

Q: What do I have to do to use HIPAA eligible Amazon ElastiCache for Redis?

If you have an executed Business Associate Agreement (BAA) with AWS, you can use ElastiCache for Redis to build HIPAA-compliant applications. If you do not have a BAA or have other questions about using AWS for your HIPAA-compliant applications, contact us for more information. See Architecting for HIPAA Security and Compliance on Amazon Web Services for information about how to configure Amazon HIPAA Eligible Services to store, process, and transmit PHI.

Q: Is Amazon ElastiCache for Redis FedRAMP authorized?

The AWS FedRAMP compliance program includes Amazon ElastiCache for Redis as a FedRAMP authorized service. United States government customers and their partners can now use the latest version of Amazon ElastiCache for Redis to process and store their FedRAMP systems, data, and mission-critical, high-impact workloads in the AWS GovCloud (US) Region, and at moderate impact level in AWS US East/West Regions.

To learn more, see the following resources:

- Amazon ElastiCache for Redis Compliance page
- AWS FedRAMP Compliance page

To see the current list of compliance programs that Amazon ElastiCache for Redis is in scope for, see AWS Services in Scope by Compliance Program.

Q: Does it cost extra to use compliance features?

No, there is no additional cost for using compliance features.

Global Datastore

Q: What is Global Datastore for Redis?

Global Datastore is a feature of Amazon ElastiCache for Redis that provides fully managed, fast, reliable and secure cross-region replication. With Global Datastore, you can write to your Amazon ElastiCache for Redis cluster in one region, and have the data available for read in up to two other cross-region replica clusters, thereby enabling low-latency reads and disaster recovery across regions.

Designed for real-time applications with a global footprint, Global Datastore for Redis supports cross-region replication latency of typically under one second, increasing the responsiveness of your applications by providing geo-local reads closer to the end users. In the unlikely event of regional degradation, one of the healthy cross-region replica clusters can be promoted to become the primary cluster with full read/write capabilities. Once initiated, the promotion typically completes in less than a minute, allowing your applications to remain available.

Q: How many AWS regions can I replicate to?

You can replicate to up to two secondary regions within a Global Datastore for Redis. The clusters in secondary regions can be used to serve low-latency local reads and for disaster recovery, in the unlikely event of a regional degradation.

Q: Which engine versions support Global Datastore for Redis?

Global Datastore is supported on Amazon ElastiCache for Redis 5.0.6 onward. Customers can upgrade engine version to 5.0.6 and use Global Datastore.

Q: How can I create Global Datastore for Redis?

You can setup a Global Datastore by using an existing cluster or creating a new cluster to be used as a primary. You can create a Global Datastore for Redis with just a few clicks on the Amazon ElastiCache Management Console or by downloading the latest AWS SDK or CLI. There is support for Global Datastore in AWS CloudFormation.

Q: Does Amazon ElastiCache automatically failover a Global Datastore for Redis to promote a secondary cluster in the event when primary cluster (region) is degraded?

No, Amazon ElastiCache doesn't automatically promote a secondary cluster in the event when primary cluster (region) is degraded. You can manually initiate the failover by promoting a

secondary cluster to become a primary. The failover and promotion of secondary cluster typically completes in less than one minute.

Q: How do I perform application failover for disaster recovery if my primary cluster experiences degradation of service?

In case your primary cluster in a Global Datastore for Redis experiences degradation of service, you can assign a secondary cluster as your new primary cluster, and then remove the old primary cluster from your Global Datastore. Once the secondary cluster is promoted to primary, Amazon ElastiCache will reconfigure the old primary (if reachable) as secondary, and setup replication to synchronize all secondary regions with the new primary. If your entire application stack is replicated to another AWS region, you may failover the entire application stack (including your compute resources) to that AWS region. If the rest of your application stack does not require failover, make sure your application has access to the secondary cluster endpoint.

Q: How is my data secured when using Global Datastore for Redis?

Global Datastore for Redis uses encryption in-transit for cross-region traffic to keep your data secure. Additionally, you can also encrypt your primary and secondary clusters using encryption at-rest to keep your end-to-end data secure. Each primary and secondary cluster can have a separate customer managed Customer Master Key (CMK) in AWS Key Management Service (KMS) for encryption at rest.

Q: What Recovery Point Objective (RPO) and Recovery Time Objective (RTO) can I expect with Global Datastore for Redis?

Amazon ElastiCache doesn't provide an SLA for RPO and RTO. The RPO varies based on replication lag between regions, and depends on network latency between regions and cross-region network traffic congestion. The RPO of Global Datastore is typically under one second, so the data written in primary region is available in secondary regions within one second. The RTO of Global Datastore for Redis is typically under a minute. Once a failover to a secondary cluster is initiated, Amazon ElastiCache typically promotes the secondary to full read/write capabilities in under a minute.

Q: What is the pricing for Global Datastore for Redis?

Amazon ElastiCache does not charge any premium to use Global Datastore for Redis. You pay for the primary and secondary clusters in your Global Datastore, and the cross-region data transfer traffic.

Data Tiering

Q: What is data tiering for ElastiCache for Redis?

Data tiering provides a new price-performance option for Redis workloads by utilizing lower-cost solid state drives (SSDs) in each cluster node in addition to storing data in memory. It is ideal for workloads that access up to 20% of their overall dataset regularly, and for applications that can tolerate additional latency when accessing data on SSD. Amazon ElastiCache R6gd nodes with memory and solid state drives have nearly 5x more total storage capacity and can help customers achieve over 60% savings in price when running at maximum utilization compared to ElastiCache R6g nodes with memory only."

Q: How does data tiering for ElastiCache for Redis work?

Data tiering works by automatically and transparently moving the least recently used items from memory to locally attached NVMe SSDs when available memory capacity is completely consumed. When an item that moves to SSD is subsequently accessed, ElastiCache moves it back to memory asynchronously before serving the request.

Q: What performance can I expect when using clusters with data tiering?

Data tiering is designed to have minimal impact on application performance. Assuming 500-byte String values, you can expect an additional 300μs latency on average for requests to data stored on SSD compared to requests to data in memory.

Q: Which engine versions support data tiering?

ElastiCache for Redis supports data tiering for Redis versions 6.2 and above.

Q: Which node types support data tiering?

ElastiCache for Redis supports data tiering on Redis clusters using R6gd nodes.

Q: Which ElastiCache features are supported for clusters using data tiering?

All Redis commands and most ElastiCache features are supported when using data tiering. For a list of features that are not supported on clusters using data tiering, see the documentation.

Q: What is the price for data tiering for ElastiCache for Redis?

There are no additional costs for using data tiering besides the node's hourly cost. Nodes with data tiering are available with on-demand pricing and as reserved nodes.

Amazon Timestream

Fast, scalable, and serverless time-series database.

With Amazon Timestream, you can easily store and analyze sensor data for IoT applications, metrics for DevOps use cases, and telemetry for application monitoring scenarios such as clickstream data analysis.

How it works

Amazon Timestream is a fast, scalable, and serverless time-series database service that makes it easier to store and analyze trillions of events per day up to 1,000 times faster. Amazon Timestream automatically scales up or down to adjust capacity and performance, so that you don't have to manage the underlying infrastructure.

IoT applications

DevOps applications

Analytics applications

Why Timestream?

Timestream saves you time and cost in managing the lifecycle of time-series data by keeping recent data in memory and moving historical data to a cost-optimized storage tier based on user-defined policies.

Use cases

IoT applications
Quickly analyze time-series data generated by IoT applications using built-in analytic functions to help you identify trends and patterns.

DevOps applications
Collect and analyze operational metrics to monitor health and usage, and analyze data in real time to improve performance and availability.

Analytics applications
Store and process incoming and outgoing web traffic data for your applications, with additional aggregate functions for analysis and insights.

What is Amazon Timestream?

Amazon Timestream is a fast, scalable, fully managed, purpose-built time series database that makes it easy to store and analyze trillions of time series data points per day. Timestream saves you time and cost in managing the lifecycle of time series data by keeping recent data in memory and moving historical data to a cost optimized storage tier based upon user defined policies. Timestream's purpose-built query engine lets you access and analyze recent and historical data together, without having to specify its location. Amazon Timestream has built-in time series analytics functions, helping you identify trends and patterns in your data in near real-time. Timestream is serverless and automatically scales up or down to adjust capacity and performance. Because you don't need to manage the underlying infrastructure, you can focus on optimizing and building your applications.

Timestream also integrates with commonly used services for data collection, visualization, and machine learning. You can send data to Amazon Timestream using AWS IoT Core, Amazon Kinesis, Amazon MSK, and open source Telegraf. You can visualize data using Amazon QuickSight, Grafana, and business intelligence tools through JDBC. You can also use Amazon SageMaker with Timestream for machine learning.

Timestream key benefits

The key benefits of Amazon Timestream are:

- *Serverless with auto-scaling* - With Amazon Timestream, there are no servers to manage and no capacity to provision. As the needs of your application change, Timestream automatically scales to adjust capacity.
- *Data lifecycle management* - Amazon Timestream simplifies the complex process of data lifecycle management. It offers storage tiering, with a memory store for recent data and a magnetic store for historical data. Amazon Timestream automates the transfer of data from the memory store to the magnetic store based upon user configurable policies.
- *Simplified data access* - With Amazon Timestream, you no longer need to use disparate tools to access recent and historical data. Amazon Timestream's purpose-built query engine transparently accesses and combines data across storage tiers without you having to specify the data location.

- *Purpose-built for time series* - You can quickly analyze time series data using SQL, with built-in time series functions for smoothing, approximation, and interpolation. Timestream also supports advanced aggregates, window functions, and complex data types such as arrays and rows.
- *Always encrypted* - Amazon Timestream ensures that your time series data is always encrypted, whether at rest or in transit. Amazon Timestream also enables you to specify an AWS KMS customer managed key (CMK) for encrypting data in the magnetic store.
- *High availability* - Amazon Timestream ensures high availability of your write and read requests by automatically replicating data and allocating resources across at least 3 different Availability Zones within a single AWS Region. For more information, see the Timestream Service Level Agreement.
- *Durability* - Amazon Timestream ensures durability of your data by automatically replicating your memory and magnetic store data across different Availability Zones within a single AWS Region. All of your data is written to disk before acknowledging your write request as complete.

Timestream use cases

Examples of a growing list of use cases for Timestream include:

- Monitoring metrics to improve the performance and availability of your applications.
- Storage and analysis of industrial telemetry to streamline equipment management and maintenance.
- Tracking user interaction with an application over time.
- Storage and analysis of IoT sensor data.

Amazon Timestream FAQ

General

What is time series data?

Time series data is a sequence of data points recorded over a time interval for measuring events that change over time. Examples are stock prices over time, temperature measurements over time, and the CPU utilization of an EC2 instance over time. With time-series data, each data point consists of a timestamp, one or more attributes, and the event that changes over time. This data is used to derive insights into the performance and health of an application, detect anomalies, and identify optimization opportunities. For example, DevOps engineers might want to view data that measures changes in infrastructure performance metrics, manufacturers might want to track IoT sensor data that measures changes in equipment across a facility, and online marketers might want to analyze clickstream data that captures how a user navigates a website

over time. Time series data is generated from multiple sources in extremely high volumes, it needs to be cost-effectively collected in near real time, and it requires efficient storage that helps organize and analyze the data.

What is Amazon Timestream?

Amazon Timestream is a fast, scalable, and serverless time series database service for IoT and operational applications that makes it easy to store and analyze trillions of events per day up to 1,000 times faster and at as little as 1/10th the cost of relational databases. Amazon Timestream saves you time and cost in managing the lifecycle of time series data by keeping recent data in memory and moving historical data to a cost optimized storage tier based upon user defined policies. Amazon Timestream's adaptive query engine lets you access and analyze recent and historical data together, without having to specify its location. Amazon Timestream has built-in time series analytics functions, helping you identify trends and patterns in your data in near real-time. Amazon Timestream is serverless and automatically scales up or down to adjust capacity and performance, so you don't need to manage the underlying infrastructure, freeing you to focus on building your applications.

How does Amazon Timestream work?

Amazon Timestream has been designed from the ground up to collect, store, and process time series data. Its serverless architecture supports fully decoupled data ingestion, storage, and query processing systems that can scale independently, enabling Amazon Timestream to offer virtually infinite scale for your application's needs. Rather than pre-defining the schema at table creation time, a Timestream table's schema is dynamically created based on the attributes of the incoming time series data, allowing for flexible and incremental schema definition. When stored, Amazon Timestream partitions the data by time and attributes of the data, accelerating data access using a purpose-built index. In addition, Amazon Timestream automates data lifecycle management by offering an in-memory store for recent data, a magnetic store for historical data, and by supporting configurable rules to automatically move data from the memory store to the magnetic store as it reaches a certain age. Amazon Timestream also simplifies data access through its purpose-built adaptive query engine that can seamlessly access and combine data across storage tiers without having to specify the data location, so you can quickly and easily derive insights from your data using SQL. Lastly, Amazon Timestream works seamlessly with your preferred data collection, visualization, analytics, and machine learning services, making it easy for you to include Amazon Timestream in your time series solutions.

How do I get started with Amazon Timestream?

You can get started with Amazon Timestream using the AWS management console, CLI, or SDKs. For more information, including tutorials and other getting started content, please see the developer guide.

Performance and scale

What performance can I expect from Amazon Timestream?

Amazon Timestream offers near real-time latencies for data ingestion. Amazon Timestream's built-in memory store is optimized for rapid point-in-time queries, and the magnetic store is optimized to support fast analytical queries. With Amazon Timestream, you can run queries that analyze tens of gigabytes of time-series data from the memory store within milliseconds, and analytical queries that analyze terabytes of time-series data from the magnetic store within seconds. Scheduled queries further improve query performance by calculating and storing the

aggregates, rollups, and other real-time analytics used to power frequently accessed operational dashboards, business reports, applications, and device-monitoring systems.

As your applications continue to send more data, Amazon Timestream automatically scales to accommodate their data ingestion and query needs. You can store exabytes of data in a single table. As your data grows over time, Amazon Timestream uses its distributed architecture and massive amounts of parallelism to process larger and larger volumes of data while keeping query latencies almost unchanged.

How does Amazon Timestream scale?

Amazon Timestream's serverless architecture supports fully decoupled data ingestion, storage, and query processing systems that can scale independently, enabling Amazon Timestream to offer virtually infinite scale for your application's needs.

What are the current limits and quotas?

For current limits and quotas, see the documentation.

Data ingestion

How can I send data to Amazon Timestream?

You can collect time series data from connected devices, IT systems, and industrial equipment, and write it into Amazon Timestream. You can send data to Amazon Timestream using data collection services such as AWS IoT Core, Amazon Kinesis Data Analytics for Apache Flink, or Telegraf, or through the AWS SDKs. For more information, see the documentation.

Do I need to define a schema before sending data to Amazon Timestream?

No. Amazon Timestream dynamically creates a table's schema based on a set of dimensional attributes and measures. This offers flexible and incremental schema definition that may be adjusted at any time without affecting availability.

How do I deal with late or future arrival data?

Late arrival data is data that has a timestamp in the past. Future data is data that has a timestamp in the future. Amazon Timestream lets you store and access both kinds.

To store late arrival data, you simply write the data into Amazon Timestream, and the service will automatically determine whether it gets written to the memory store or to the magnetic store, based on the timestamp of the data and on the configured data retention for the memory and magnetic stores.

To store future data, model your data as a multi-measure record, and represent the future timestamp as a measure within the record.

What is the time stamp used by Amazon Timestream when ingesting data?

Amazon Timestream uses the timestamp of the time series event being written into the database. It supports timestamps with nanosecond granularity.

How can I send data to Amazon Timestream using AWS IoT Core?

You can collect data from your IoT devices and store that data in Amazon Timestream using AWS IoT Core rule actions. For more detailed information, see the documentation.

How can I send data to Amazon Timestream using AWS Lambda?

You can create AWS Lambda functions that interact with Amazon Timestream. For more detailed information, see the documentation.

How can I send data to Amazon Timestream using Amazon Kinesis?

You can use Apache Flink to send your time series data from Amazon Kinesis directly into Amazon Timestream. For more detailed information, see the documentation.

How can I send data to Amazon Timestream using Amazon Kinesis Data Analytics?

You can use Apache Flink to transfer your time series data from Amazon Kinesis Data Analytics directly into Amazon Timestream. For more detailed information, see the documentation.

How can I send data to Amazon Timestream using Amazon MSK?

You can use Apache Flink to send your time series data from Amazon MSK directly into Amazon Timestream. For more detailed information, see the documentation.

How can I send data to Amazon Timestream using open source Telegraf?

You can send time series data collected using open source Telegraf directly into Amazon Timestream using the Telegraf connector. For more detailed information, see the documentation.

Data storage

How does Amazon Timestream store data?

Amazon Timestream organizes and stores time-series data using its timestamp and organizes data across time based on its dimensional attributes. See Werner Vogels blog for more details. Using Amazon Timestream, you can automate data lifecycle management by simply configuring data retention policies to automatically move data from the memory store to the magnetic store as it reaches the configured age.

What are the benefits of the memory store?

Amazon Timestream's memory store is a write-optimized store that accepts and deduplicates the incoming time series data. It also accepts and processes late arriving data from devices and applications with intermittent connectivity. The memory store is also optimized for latency sensitive point-in-time queries.

What are the benefits of the magnetic store?

Amazon Timestream's magnetic store is a read-optimized store that contains historical data. The magnetic store is also optimized for fast analytical queries that scan hundreds of terabytes of data.

Data access, analytics, and machine learning

How can I query data with Amazon Timestream?

You can use SQL to query your time series data stored in Amazon Timestream. You can also conduct time series analytics functions for interpolation, regression, and smoothing using SQL. For more information, see the documentation.

Can I automatically rollup or aggregate or preprocess my data with Amazon Timestream?

Amazon Timestream's scheduled queries offer a fully managed, serverless, and scalable solution for calculating and storing aggregates, rollups, and other real-time analytics used to power frequently accessed operational dashboards, business reports, applications, and device monitoring systems.

With scheduled queries, you simply define the real-time analytics queries that calculate aggregates, rollups, and other real-time analytics on your incoming data, and Amazon Timestream periodically and automatically runs these queries and reliably writes the query results into a separate table. You can then point your dashboards, reports, applications, and monitoring systems to simply query the destination tables, instead of querying the considerably larger source tables containing the incoming time-series data. This leads to performance and cost reductions by an order of magnitude because the destination tables contain much less data than the source tables, thereby offering faster and cheaper data access and storage.

Can I use a JDBC driver with Amazon Timestream?

You can use a JDBC deriver to connect Amazon Timestream to your preferred business intelligence tools and other applications. See the documentation for additional details.

What visualization, analytics, and machine learning (ML) tools can I use with Amazon Timestream?

You can visualize and analyze time-series data in Amazon Timestream using Amazon QuickSight and Grafana. You can also use Amazon SageMaker with Amazon Timestream for your ML needs.

How do I use Amazon Timestream with Amazon QuickSight?

You can create rich and interactive dashboards for your Amazon Timestream time-series data using Amazon QuickSight. For more information, see the documentation.

How do I use Amazon Timestream with Grafana?

You can visualize your Amazon Timestream time-series data and create alerts using Grafana, a multi-platform, open-source analytics and interactive visualization tool. To learn more and find sample applications, see the documentation.

How do I use Amazon Timestream with Amazon SageMaker?

You can use Amazon SageMaker notebooks to integrate your ML models with Amazon Timestream. For more information, see the documentation.

Availability

What availability does Amazon Timestream have?

Amazon Timestream provides 99.99% availability. Please refer to the service level agreement (SLA).

Security and compliance

Does Amazon Timestream support data encryption?

In Amazon Timestream, data is always encrypted, whether at rest or in transit. Amazon Timestream also enables you to specify an AWS KMS customer managed key (CMK) for encrypting data in the magnetic store.

What compliance certification readiness does Amazon Timestream meet?

Amazon Timestream is HIPAA eligible, ISO (9001, 27001, 27017, and 27018), PCI DSS, FedRAMP (Moderate), and Health Information Trust Alliance (HITRUST) Common Security Framework (CSF) compliant. Furthermore, Amazon Timestream is in scope for AWS's SOC reports SOC 1, SOC 2, and SOC 3.

Can I use Amazon Timestream in an Amazon Virtual Private Cloud (VPC)?

You can access Amazon Timestream from your Amazon VPC using VPC endpoints. Amazon VPC endpoints are easy to configure and provide reliable connectivity to Amazon Timestream APIs without requiring an internet gateway or a Network Address Translation (NAT) instance.

Amazon Quantum Ledger Database (QLDB)

Maintain an immutable, cryptographically verifiable log of data changes.

Amazon Quantum Ledger Database (Amazon QLDB) is a fully managed ledger database that provides a transparent, immutable, and cryptographically verifiable transaction log owned by a central trusted authority. You can use Amazon QLDB to track all application data changes, and maintain a complete and verifiable history of changes over time.

How it works

Amazon Quantum Ledger Database (QLDB) is a fully managed ledger database that provides a transparent, immutable, and cryptographically verifiable transaction log.

Use cases

Store financial transactions
Create a complete and accurate record of all financial transactions, such as credit and debit transactions.

Reconcile supply chain systems
Record the history of each transaction and provide details of every batch manufactured, shipped, stored, and sold from facility to store.

Maintain claims history
Track a claim over its lifetime, and cryptographically verify data integrity to make the application resilient against data entry errors and manipulation.

Centralize digital records
Implement a system-of-record application to create a complete, centralized record of employee details such as payroll, bonus, and benefits.

What is Amazon QLDB?

Amazon Quantum Ledger Database (Amazon QLDB) is a fully managed ledger database that provides a transparent, immutable, and cryptographically verifiable transaction log owned by a central trusted authority. You can use Amazon QLDB to track all application data changes, and maintain a complete and verifiable history of changes over time.

Ledgers are typically used to record a history of economic and financial activity in an organization. Many organizations build applications with ledger-like functionality because they want to maintain an accurate history of their applications' data. For example, they might want to track the history of credits and debits in banking transactions, verify the data lineage of an insurance claim, or trace the movement of an item in a supply chain network. Ledger applications are often implemented using custom audit tables or audit trails created in relational databases.

Amazon QLDB is a new class of database that helps eliminate the need to engage in the complex development effort of building your own ledger-like applications. With QLDB, the history of changes to your data is immutable—it can't be overwritten or altered in place. And using cryptography, you can verify that there have been no unintended changes to your application's data. QLDB uses an immutable transactional log, known as a *journal*. The journal is append-only and is composed of a sequenced and hash-chained set of *blocks* that contain your committed data.

Amazon Quantum Ledger Database (QLDB) FAQs

General

What is Amazon Quantum Ledger Database?

Amazon Quantum Ledger Database (QLDB) is a purpose-built ledger database that provides a complete and cryptographically verifiable history of all changes made to your application data.

How is a ledger database different from other databases?

Traditional databases allow you to overwrite or delete data, so developers use techniques such as audit tables and audit trails to help track data lineage. While these approaches can work, they require custom development, can be difficult to scale, and put the onus on the application developer to ensure all the right data is being recorded. Data in Amazon QLDB is written to an append-only journal, providing the developer with full data lineage. Moreover, data in Amazon QLDB's journal is immutable and verifiable, meaning you can trust the data in your ledger.

What data should I store in a ledger database?

Amazon QLDB's features make it a natural fit for system-of-record applications – those for which data integrity, completeness, and verifiability are critical. For example, in the supply chain and logistics space, an application built on Amazon QLDB would have the entire history of changes, such as movement between carriers and across borders, available for query and analysis. In finance, system-of-record applications track critical data, such as credit and debit transactions. Instead of building complex record keeping functionality within their application, banks can use QLDB to easily store a permanent and complete record of all financial transactions.

Is Amazon Quantum Ledger Database a distributed ledger or blockchain service?

Amazon QLDB is not a blockchain or distributed ledger technology. Blockchain and distributed ledger technologies focus on solving the problem of decentralized applications involving multiple parties where there can be no single entity that owns the application, and the parties do not necessarily trust each other fully. On the other hand, QLDB is a ledger database purpose-built for customers who need to maintain a complete and verifiable history of data changes in an application that they own. Amazon QLDB offers history, immutability and verifiability combined with the familiarity, scalability and ease of use of a fully managed AWS database. If your application requires decentralization and involves multiple, untrusted parties, a blockchain solution may be appropriate. If your application requires a complete and verifiable history of all application data changes, but does not involve multiple, untrusted parties, Amazon QLDB is a great fit. If you have a use case for distributed ledgers or blockchain, please see Amazon Managed Blockchain.

What kind of functionality does Amazon QLDB support?

In addition to providing a complete and verifiable history of application data changes, Amazon QLDB supports transactions with ACID semantics, a flexible document data model, and a familiar SQL-like API. QLDB is also fully managed and automatically scales to meet the needs of your application with no provisioning required.

How do I connect to Amazon QLDB from my application?

In order to connect to Amazon QLDB and transact with the data in the ledger, you need to use the AWS-provided QLDB driver. Follow the steps in this link to download the driver and configure a connection.

How do I try Amazon QLDB?

Getting started with Amazon QLDB is easy as there are no servers to manage or capacity to provision. You can create a new ledger in minutes using the AWS Management Console, AWS Command Line Interface (CLI), an AWS CloudFormation template, or by making calls to the QLDB API.

Performance

What type of performance can I expect from Amazon QLDB?

Amazon QLDB can execute 2 – 3X as many transactions than ledgers in common blockchain frameworks. Blockchain frameworks are decentralized so to execute a transaction, they require a majority of members of the network to reach consensus on the validity of the transaction. On the other hand, QLDB has a centralized design, allowing its transactions to execute without the need for multi-party consensus.

Querying

What is PartiQL? How does Amazon QLDB support it?

Amazon QLDB allows you to access and manipulate your data using PartiQL, which is a new open standard query language that supports SQL-compatible access to QLDB's document-oriented data model that includes semi-structured and nested data while remaining independent of any particular data source. To learn more about PartiQL read here.

In which AWS regions is Amazon QLDB available?

Amazon QLDB is available in the following AWS regions: US East (N. Virginia, Ohio), US West (Oregon), Europe (Ireland, Frankfurt), and Asia Pacific (Singapore, Sydney, Seoul, Tokyo), with additional regions coming soon.

Scalability

How does Amazon QLDB scale?

With Amazon QLDB, you don't have to worry about provisioning capacity or configuring read and write limits. You create a ledger, define your tables, and QLDB automatically scales to support the demands of your application.

What are the limits associated with an Amazon QLDB?

You can see the limits associated with Amazon QLDB on the AWS Service Limits page.

Backup and Restore

Can I take a snapshot or backup of my ledger?

Amazon QLDB does not support a backup and restore feature as of now. At present, an export to S3 functionality is available. Using this functionality you can export the contents of your QLDB journal to S3.

Can I restore my ledger to a particular point in time?

Amazon QLDB does not support a point-in-time restore feature as of now.

Availability, Durability, and Replication

Is Amazon QLDB durable?

Amazon QLDB's ledger is deployed across multiple AZs with multiple copies per AZ. We maintain redundancy within the region and ensure full recovery from availability zone failures. A write is acknowledged only after being written to a durable storage in multiple AZs, and hence, QLDB is strongly durable.

How does high availability work in Amazon QLDB?

Amazon QLDB is a highly available service. By default, multiple copies of your QLDB ledger are replicated across availability zones in a region. So, in the case of a zone failure you can still continue to operate QLDB.

Does Amazon QLDB have cross-region replication?

Amazon QLDB does not support cross-region replication as of now. QLDB's export to S3 feature enables customers to export the contents of the QLDB journal to a S3 bucket. The S3 buckets can be configured for cross-region replication.

Security

Can I use Amazon QLDB in Amazon Virtual Private Cloud (Amazon VPC)?

Amazon QLDB is integrated with AWS Private Link. Customers can create a VPC endpoint, which enables them to privately connect a VPC to supported AWS services powered by PrivateLink without requiring an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection.

How does authentication work with Amazon QLDB?

Amazon QLDB uses the same authentication mechanism as other AWS services. The mechanism requires a request signature to be attached to the HTTP requests (header or query string). The signature is computed using other requests fields and the AWS credentials (Access Key ID and Secret Access Key).

How does encryption work in Amazon QLDB?

By default, all data in transit and at rest is encrypted. Today, Amazon QLDB does not support customer managed CMKs (Customer Master Keys). Amazon QLDB uses AWS-owned keys to encrypt customer data.

Streaming Capability

How do customers get started using Amazon QLDB Streaming Capability?

1/ Create an Amazon Kinesis Data Stream (KDS): Customers create a KDS stream either by using Amazon Kinesis console or AWS SDK. They can also use an existing stream.

2/ Create an Amazon QLDB Stream: Customers create an Amazon QLDB Stream using Amazon QLDB console or AWS SDK. The following key aspects need to be configured for creating the stream:

- Amazon QLDB Ledger name
- Target KDS stream
- IAM role that Amazon QLDB will assume, which has permissions to write the KDS stream
- Start time, that specifies the clock time from which to start streaming
- End time, that specifies the clock time to end streaming. Omitting this optional parameter implies a continuous stream.

3/ Consume the KDS Stream: Customers can write Kinesis applications using Kinesis services, such as Kinesis Data Firehose and Kinesis Data Analytics, to consume and process the stream. Customers can also use AWS Lambda to consume the Kinesis Data Stream records.

What is QLDB Streams delivery guarantee and ordering behavior?

Amazon QLDB Streaming capability provide at-least-once delivery guarantee. Each control, block summary and revision details record generated from the journal by a QLDB Stream, is streamed to Amazon Kinesis at least once. Since duplicate records can appear in the Kinesis stream under some circumstances, customers will need to have deduplication logic in the application layer, if it is needed by the application. There are no ordering guarantees, as previous blocks and revisions can be published in the Kinesis stream out-of-order.

Amazon Redshift

Best price-performance for cloud data warehousing.

Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse service that makes it simple and cost-effective to efficiently analyze all your data using your existing business intelligence tools. It is optimized for datasets ranging from a few hundred gigabytes to a petabyte or more and costs less than $1,000 per terabyte per year, a tenth the cost of most traditional data warehousing solutions.

How it works

Amazon Redshift uses SQL to analyze structured and semi-structured data across data warehouses, operational databases, and data lakes, using AWS-designed hardware and machine learning to deliver the best price performance at any scale.

Introduction to Amazon Redshift

Tens of thousands of customers today rely on Amazon Redshift to analyze exabytes of data and run complex analytical queries, making it a widely used cloud data warehouse. Run and scale analytics in seconds on all your data without having to manage your data warehouse infrastructure.

Use cases

Improve financial and demand forecasts
Automatically create, train, and deploy machine learning models for predictive insights.

Collaborate and share data
Securely share data among accounts, organizations, and partners while building applications on top of third-party data.

Optimize your business intelligence
Build insight-driven reports and dashboards using Amazon QuickSight, Tableau, Microsoft PowerBI, or other business intelligence tools.

Increase developer productivity
Get simplified data access, ingest, and egress from numerous programming languages and platforms without configuring drivers and managing database connections.

What is Amazon Redshift?

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. You can start with just a few hundred gigabytes of data and scale to a petabyte or more. This allows you to use your data to gain new insights for your business and customers.

The first step to create a data warehouse is to launch a set of nodes, called an Amazon Redshift cluster. After you provision your cluster, you can upload your data set and then perform data analysis queries. Regardless of the size of the data set, Amazon Redshift offers fast query performance using the same SQL-based tools and business intelligence applications that you use today.

Amazon Redshift Serverless

Amazon Redshift Serverless provides data warehouse capacity and intelligent scaling with simple configuration and management. You can access and analyze data without setting up a provisioned data warehouse. For more information, see What is Amazon Redshift Serverless?.

Amazon Redshift FAQs

General

To find out what's new with Amazon Redshift, visit the What's New page.
To view more detailed information and usage guidance, visit the Documentation.

Q: What is Amazon Redshift?

Amazon Redshift is a fully managed, scalable cloud data warehouse that accelerates your time to insights with fast, easy, and secure analytics at scale. Thousands of customers rely on Amazon Redshift to analyze data from terabytes to petabytes and run complex analytical queries. You can get real-time insights and predictive analytics on all your data across your operational databases, data lake, data warehouse, and third-party datasets. Amazon Redshift delivers all this at a price performance that's up to 3x better than other cloud data warehouses out of the box, helping you keep your costs predictable.

Amazon Redshift Serverless makes it easy for you to run petabyte-scale analytics in seconds to get rapid insights without having to configure and manage your data warehouse clusters. Amazon Redshift Serverless automatically provisions and scales the data warehouse capacity to deliver high performance for demanding and unpredictable workloads, and you pay only for the resources you use.
Q: What are the top reasons customers choose Amazon Redshift?

Thousands of customers choose Amazon Redshift to accelerate their time to insights because it's easy to use, it delivers performance at any scale, and it lets you analyze all your data. Amazon Redshift is a fully managed service and offers both provisioned and serverless options, making it easy for you to run and scale analytics without having to manage your data warehouse. You can choose the provisioned option for predictable workloads or go with the Amazon Redshift Serverless option to automatically provision and scale the data warehouse capacity to deliver high performance for demanding and unpredictable workloads. It delivers performance at any scale with up to 3x better price performance than other cloud data warehouses out of the box, helping you keep your costs predictable. Amazon Redshift lets you get insights from running real-time and predictive analytics on all your data across your operational databases, data lake, data warehouse, and thousands of third-party datasets. Amazon Redshift keeps your data secure at rest and in transit and meets internal and external compliance requirements. It supports industry-leading security to protect your data in transit and at rest and is compliant with SOC1, SOC2, SOC3, and PCI DSS Level 1 requirements. All Redshift security and compliance features are included at no additional cost.
Q: How does Amazon Redshift simplify data warehouse management?

Amazon Redshift is fully managed by AWS so you no longer need to worry about data warehouse management tasks such as hardware provisioning, software patching, setup, configuration, monitoring nodes and drives to recover from failures, or backups. AWS manages the work needed to set up, operate, and scale a data warehouse on your behalf, freeing you to focus on building your applications. Amazon Redshift also has automatic tuning capabilities, and surfaces recommendations for managing your warehouse in Redshift Advisor. For Redshift Spectrum, Amazon Redshift manages all the computing infrastructure, load balancing, planning, scheduling, and execution of your queries on data stored in Amazon S3. The serverless option automatically provisions and scales the data warehouse capacity to deliver high performance for demanding and unpredictable workloads, and you pay only for the resources you use.

Q: How does the performance of Amazon Redshift compare to that of other data warehouses?

TPC-DS benchmark results show that Amazon Redshift provides the best price performance out of the box, even for a comparatively small 3 TB dataset. Amazon Redshift delivers up to 3x better price performance than other cloud data warehouses. This means that you can benefit from Amazon Redshift's leading price performance from the start without manual tuning. Get up to 3x better price performance with Amazon Redshift than with other cloud data warehouses | AWS Big Data Blog.

Amazon Redshift uses a variety of innovations to achieve up to 10x better performance than traditional databases for data warehousing and analytics workloads, including efficient read-optimized columnar compressed data storage with massively parallel processing (MPP) compute clusters that scale linearly to hundreds of nodes. Instead of storing data as a series of rows, Amazon Redshift organizes the data by column. When loading data into an empty table, Amazon Redshift automatically samples your data and selects the most appropriate compression scheme.

Redshift Spectrum lets you run queries against exabytes of data in Amazon S3. There is no loading or extract, transform, and load (ETL) required. Even if you don't store any of your data in Amazon Redshift, you can still use Redshift Spectrum to query datasets as large as an exabyte in Amazon S3. Materialized views provide significantly faster query performance for repeated and predictable analytical workloads such as dashboards, queries from business intelligence (BI) tools, and ETL data processing. Using materialized views, you can store the precomputed results of queries and efficiently maintain them by incrementally processing the latest changes made to the source tables. Subsequent queries referencing the materialized views use the precomputed results to run much faster, and automatic refresh and query rewrite capabilities simplify and automate the use of materialized views. Amazon Redshift also creates materialized views automatically, providing the same benefits of user-created materialized views.

The compute and storage capacity of on-premises data warehouses are limited by the constraints of the on-premises hardware. Amazon Redshift gives you the ability to scale compute and storage independently as needed to meet changing workloads. With Redshift Managed Storage (RMS), you now have the ability to scale your storage to petabytes using Amazon S3 storage.

Automatic Table Optimization (ATO) is a self-tuning capability that helps you achieve the performance benefits of creating optimal sort and distribution keys without manual effort. ATO observes how queries interact with tables and uses machine learning (ML) to select the best sort and distribution keys to optimize performance for the cluster's workload. ATO optimizations have shown to increase cluster performance by 24% and 34% using the 3 TB and 30 TB TPC-DS benchmarks, respectively, versus a cluster without ATO. Additional features such as Automatic Vacuum Delete, Automatic Table Sort, and Automatic Analyze eliminate the need for manual maintenance and tuning of Redshift clusters to get the best performance for new clusters and production workloads.

Workload management allows you to route queries to a set of defined queues to manage the concurrency and resource utilization of the cluster. Today, Amazon Redshift has both automatic and manual configuration types. With manual WLM configurations, you're responsible for defining the amount of memory allocated to each queue and the maximum number of queries, each of which gets a fraction of that memory, which can run in each of their queues. Manual WLM configurations don't adapt to changes in your workload and require an intimate knowledge of your queries' resource utilization to get right. Amazon Redshift Auto WLM doesn't require you to define the memory utilization or concurrency for queues. Instead, it adjusts the concurrency dynamically to optimize for throughput. Optionally, you can define query priorities to provide queries preferential resource allocation based on your business priority. Auto WLM also provides powerful tools to let you manage your workload. Query priorities let you define priorities for workloads so they can get preferential treatment in Amazon Redshift, including more resources during busy times for consistent query performance, and query monitoring

rules offer ways to manage unexpected situations such as detecting and preventing runaway or expensive queries from consuming system resources. The following are key areas of Auto WLM with adaptive concurrency performance improvements: proper allocation of memory, elimination of static partitioning of memory between queues, and improved throughput.

Amazon Redshift Advisor develops customized recommendations to increase performance and optimize costs by analyzing your workload and usage metrics for your cluster. Sign in to the Amazon Redshift console to view Advisor recommendations. For more information, see Working with recommendations from Amazon Redshift Advisor.

Q: How do I get started with Amazon Redshift?

With just a few clicks in the AWS Management Console, you can start querying data. You can take advantage of pre-loaded sample data sets, including benchmark datasets TPC-H, TPC-DS, and other sample queries to kick start analytics immediately. You can create databases, schemas, tables and load data from Amazon S3, Amazon Redshift data shares, or restore from an existing Amazon Redshift provisioned cluster snapshot. You can also directly query data in open formats, such as Parquet or ORC in Amazon S3 data lake, or query data in operational databases, such as Amazon Aurora, Amazon RDS PostgreSQL and MySQL.

To get started with Amazon Redshift Serverless, choose “Try Amazon Redshift Serverless” and start querying data. Amazon Redshift Serverless automatically scales to meet any increase in workloads.

Q: Can I get help to learn more about and onboard to Amazon Redshift?

A: Yes, Amazon Redshift specialists are available to answer questions and provide support. Contact Us and you’ll hear back from us in one business day to discuss how AWS can help your organization.

Q: What is Amazon Redshift managed storage?

Amazon Redshift managed storage is available with serverless and RA3 node types and lets you scale and pay for compute and storage independently so you can size your cluster based only on your compute needs. It automatically uses high-performance SSD-based local storage as tier-1 cache and takes advantage of optimizations such as data block temperature, data block age, and workload patterns to deliver high performance while scaling storage automatically to Amazon S3 when needed without requiring any action.

Q: How do I use Amazon Redshift’s managed storage?

If you are already using Amazon Redshift Dense Storage or Dense Compute nodes, you can use Elastic Resize to upgrade your existing clusters to the new compute instance RA3. Amazon Redshift Serverless and clusters using the RA3 instance automatically use Redshift-managed storage to store data. No other action outside of using Amazon Redshift Serverless or RA3 instances is required to use this capability.

Q: What is Amazon Redshift Spectrum?

Amazon Redshift Spectrum is a feature of Amazon Redshift that lets you run queries against your data lake in Amazon S3, with no data loading or ETL required. When you issue an SQL query, it goes to the Amazon Redshift endpoint, which generates and optimizes a query plan. Amazon Redshift determines what data is local and what is in Amazon S3, generates a plan to minimize the amount of S3 data that needs to be read, and requests Amazon Redshift Spectrum workers out of a shared resource pool to read and process data from S3.
Q: When should I consider using RA3 instances?

Consider choosing RA3 node types in these cases:

- You need the flexibility to scale and pay for compute separate from storage.
- You query a fraction of your total data.
- Your data volume is growing rapidly or is expected to grow rapidly.
- You want the flexibility to size the cluster based only on your performance needs.

As the scale of data continues to grow, reaching petabytes, the amount of data you ingest into your Amazon Redshift data warehouse is also growing. You may be looking for ways to cost-effectively analyze all your data.

With new Amazon Redshift RA3 instances with managed storage, you can choose the number of nodes based on your performance requirements, and pay only for the managed storage that you use. This gives you the flexibility to size your RA3 cluster based on the amount of data you process daily without increasing your storage costs. Built on the AWS Nitro System, RA3 instances with managed storage use high performance SSDs for your hot data and Amazon S3 for your cold data, providing ease of use, cost-effective storage, and fast query performance.
Q: When would I use Amazon Redshift vs. Amazon RDS?

Both Amazon Redshift and Amazon Relational Database Service (RDS) let you run traditional relational databases in the cloud while off-loading database administration. Customers use Amazon RDS databases primarily for online-transaction processing (OLTP) workloads, while Amazon Redshift is used primarily for reporting and analytics. OLTP workloads require quickly querying specific information, and support for transactions such as insert, update, and delete are best handled by Amazon RDS. Amazon Redshift harnesses the scale and resources of multiple nodes and uses a variety of optimizations to provide order of magnitude improvements over traditional databases for analytic and reporting workloads against very large datasets. Amazon Redshift provides an excellent scale-out option as your data and query complexity grows if you want to prevent your reporting and analytic processing from interfering with the performance of your OLTP workload. Now, with the new Federated Query feature, you can easily query data across your Amazon RDS or Aurora database services with Amazon Redshift.
Q: When would I use Amazon Redshift or Redshift Spectrum vs. Amazon EMR?

You should use Amazon EMR if you use custom code to process and analyze extremely large datasets with big data processing frameworks such as Apache Spark, Hadoop, Presto, or Hbase. Amazon EMR gives you full control over the configuration of your clusters and the software you install on them.

Data warehouses like Amazon Redshift are designed for a different type of analytics altogether. Data warehouses are designed to pull together data from lots of different sources, like inventory, financial, and retail sales systems. In order to ensure that reporting is consistently accurate across the entire company, data warehouses store data in a highly structured fashion. This structure builds data consistency rules directly into the tables of the database. Amazon Redshift is the best service to use when you need to perform complex queries on massive collections of structured and semi-structured data and get fast performance.

While the Redshift Spectrum feature is great for running queries against data in Amazon Redshift and S3, it really isn't a fit for the types of use cases that enterprises typically ask from processing frameworks like Amazon EMR. Amazon EMR goes far beyond just running SQL queries. Amazon EMR is a managed service that lets you process and analyze extremely large data sets using the latest versions of popular big data processing frameworks, such as Spark, Hadoop, and Presto, on fully customizable clusters. With Amazon EMR, you can run a wide variety of scale-out data processing tasks for applications such as machine learning, graph analytics, data transformation, streaming data, and virtually anything you can code.

You can use Redshift Spectrum with EMR. Redshift Spectrum uses the same approach to store table definitions as Amazon EMR. Redshift Spectrum can support the same Apache Hive Metastore used by Amazon EMR to locate data and table definitions. If you're using Amazon EMR and have a Hive Metastore already, you just have to configure your Amazon Redshift cluster to use it. You can then start querying that data right away along with your Amazon EMR jobs. Therefore, if you're already using EMR to process a large data store, you can use Redshift Spectrum to query that data at the same time without interfering with your Amazon EMR jobs.

Query services, data warehouses, and complex data processing frameworks all have their place, and they are used for different things. You just need to choose the right tool for the job.
Q: When should I use Amazon Athena vs. Amazon Redshift Spectrum?

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is easy to use. Simply point to your data in S3, define the schema, and start querying using standard SQL.

Redshift Spectrum is a feature of Amazon Redshift. If you need to analyze frequently accessed data with highest-performance, strict Service Level Agreement (SLA), you should use Amazon Redshift. You can use Redshift Spectrum to extend your Amazon Redshift queries out to less frequently accessed data in your Amazon S3 data lake. This gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need it.
Q: Why should I use Amazon Redshift instead of running my own MPP data warehouse cluster on Amazon EC2?

Amazon Redshift automatically handles many of the time-consuming tasks associated with managing your own data warehouse, including:

- *Setup:* With Amazon Redshift, you simply create a data warehouse cluster, define your schema, and begin loading and querying your data. You don't have to manage provisioning, configuration or patching.
- *Data Durability:* Amazon Redshift replicates your data within your data warehouse cluster and continuously backs up your data to Amazon S3, which is designed for eleven nines of durability. Amazon Redshift mirrors each drive's data to other nodes within your cluster. If a drive fails, your queries will continue with a slight latency increase while Redshift rebuilds your drive from replicas. In case of node failure(s), Amazon Redshift automatically provisions new node(s) and begins restoring data from other drives within the cluster or from Amazon S3. It prioritizes restoring your most frequently queried data so your most frequently executed queries will become performant quickly.
- *Scaling:* You can add or remove nodes from your Amazon Redshift data warehouse cluster with a single API call or via a few clicks in the AWS Management Console as your capacity and performance needs change. You can also schedule your scaling and resize operations by using the scheduler capability in Amazon Redshift.
- *Automatic Updates and Patching:* Amazon Redshift automatically applies upgrades and patches your data warehouse so you can focus on your application and not on its administration.
- *Exabyte Scale Query Capability:* Amazon Redshift Spectrum enables you to run queries against exabytes of data in Amazon S3. There is no loading or ETL required. Even if you don't store any of your data in Amazon Redshift, you can still use Redshift Spectrum to query datasets as large as an exabyte in Amazon S3.

Q: How do I create and access an Amazon Redshift data warehouse cluster?

You can easily create an Amazon Redshift data warehouse cluster by using the AWS Management Console or the Amazon Redshift APIs. You can start with a single node, 160 GB data warehouse and scale all the way to petabytes or more with a few clicks in the AWS Console or a single API call.

The single node configuration, which is best suited for evaluation or development/test workloads, lets you get started with Amazon Redshift quickly and cost effectively and scale up to a multi-node configuration as your needs grow. A Redshift data warehouse cluster can contain 1–128 compute nodes, depending on the node type. For the latest generation node type, RA3, the minimum number of nodes is two. For details, see the documentation.

The multi-node configuration requires a leader node that manages client connections and receives queries, and two compute nodes that store data and perform queries and computations. The leader node, which is the same size as the compute node, is provisioned for you automatically and you are not charged for it.

Simply specify your preferred Availability Zone (optional), the number of nodes, node types, a primary name and password, security groups, your preferences for backup retention, and other system settings. Once you've chosen your desired configuration, Amazon Redshift will provision the required resources and set up your data warehouse cluster.

Once your data warehouse cluster is available, you can retrieve its endpoint and JDBC and ODBC connection string from the AWS Management Console or by using the Redshift APIs. You can then use this connection string with your favorite database tool, programming language, or Business Intelligence (BI) tool. You will need to authorize network requests to your running data warehouse cluster. For a detailed explanation, please refer to our Getting Started Guide.
Q: Why should I use Amazon Redshift Spatial?

Amazon Redshift spatial provides location-based analytics for rich insights into your data. It seamlessly integrates spatial and business data to provide analytics for decision making. Amazon Redshift launched native spatial data processing support in November 2019, with a polymorphic data type GEOMETRY and several key SQL spatial functions. We now support GEOGRAPHY data type, and our library of SQL spatial functions has grown to 80. We support all the common spatial datatypes and standards, including Shapefiles, GeoJSON, WKT, WKB, eWKT, and eWKB. To learn more, visit the documentation page or the Amazon Redshift spatial tutorial page.
Q: What is cold query performance enhancement, and what does Amazon Redshift do to enhance cold query performance?

Amazon Redshift can process queries up to 2x faster when they need to be compiled. This improvement gives you better query performance when you create a new Redshift cluster, onboard a new workload on an existing cluster, or after a software update of an existing cluster. These query performance improvements are available at no additional cost, and no action is needed to enable it on your clusters.

With cold query performance enhancement, query compilations are scaled to a serverless compilation service beyond the compute resources of the leader node of your cluster. Amazon Redshift supports an unlimited cache to store compiled objects to increase cache hits, from 99.60% to 99.95%, when your mission-critical queries are submitted to Amazon Redshift.

When queries are sent to Amazon Redshift, the query execution engine compiles the query into machine code and distributes it to the cluster nodes. The compiled code runs faster because it eliminates the overhead of using an interpreter. For a new cluster with no code cache or after an existing cluster is upgraded with the latest release, code cache is flushed, and queries must undergo query compilation. As a result, the latency of a query may vary, which may not meet the requirements of some workloads. With this update, unlimited cache minimizes the need to compile code, and when compilation is needed, a scalable compilation farm compiles it in parallel to speed up your workloads. The magnitude of increased speed depends on the workload's complexity and concurrency. To learn more about code compilation, see Query Processing in the Database developer guide.

Serverless

Q: What is Amazon Redshift Serverless (preview)?

Amazon Redshift Serverless (preview) is a serverless option of Amazon Redshift that makes it easy to run and scale analytics in seconds without the need to set up and manage data warehouse infrastructure. With Redshift Serverless, any user—including data analysts, developers, business professionals, and data scientists—can get insights from data by simply loading and querying data in the data warehouse.
Q: How do I get started with Amazon Redshift Serverless (preview)?

With just a few clicks in the AWS Management Console, you can choose "configure Amazon Redshift Serverless" and begin querying data. You can take advantage of preloaded sample datasets, such as weather data, census data, and benchmark datasets, along with sample queries to kick start analytics immediately. You can create databases, schemas, tables, and load data from Amazon S3, Amazon Redshift data shares, or restore from an existing Redshift provisioned cluster snapshot. You can also directly query data in open formats (such as Parquet or ORC) in the Amazon S3 data lake, or query data in operational databases, such as Amazon Aurora and Amazon RDS PostgreSQL and MySQL.
Q: What capabilities does Amazon Redshift Serverless (preview) provide?

Amazon Redshift Serverless offers you numerous benefits, including:

- The ability to gain insights rapidly without provisioning and managing clusters.
- Intelligent and automatic scaling based on workload demands without having to over-provision resources.
- Continuous service availability for scaling and version updates.
- Fast, out-of-the box query performance for both data loaded in the data warehouse, open formats in Amazon S3 data lake, and data in operational databases without requiring database tuning.
- Rich SQL analytics, durability, and transactional guarantees of Amazon Redshift.
- Cost efficiency by paying only for the capacity used and reduced data warehouse complexity.

Q: What are the benefits of using Amazon Redshift Serverless (preview)?

If you don't have data warehouse management experience, you don't have to worry about setting up, configuring, managing clusters or tuning the warehouse. You can focus on deriving meaningful insights from your data or delivering on your core business outcomes through data. You pay only for what you use, keeping costs manageable. You continue to benefit from all of Amazon Redshift's top-notch performance, rich SQL features, seamless integration with data lakes and operational data warehouses, and built-in predictive analytics and data sharing

capabilities. If you need fine-grained control of your data warehouse, you can provision Redshift clusters.
Q: How does Amazon Redshift Serverless (preview) work with other AWS services?

You can continue to use all the rich analytics functionality of Amazon Redshift, such as complex joins, direct queries to data in the Amazon S3 data lake and operational databases, materialized views, stored procedures, semi-structured data support, and ML, as well as high performance at scale. All the related services that Amazon Redshift integrates with (such as Amazon Kinesis, AWS Lambda, Amazon QuickSight, Amazon SageMaker, Amazon EMR, AWS Lake Formation, and AWS Glue) continue to work with Amazon Redshift Serverless.
Q: What use cases can I handle with Amazon Redshift Serverless (preview)?

You can continue to run all analytics use cases. With a simple getting started workflow, automatic scaling, and the ability to pay for use, the Amazon Redshift Serverless experience now makes it even easier and more cost-effective to run development and test environments that need to get started quickly, ad-hoc business analytics, workloads with varying and unpredictable compute needs, and intermittent or sporadic workloads.
Q: How is Amazon Athena different than Amazon Redshift Serverless?

Amazon Athena and Amazon Redshift address different needs and use cases even if both services are serverless. A data warehouse such as Amazon Redshift is the best choice if you need the best price performance for complex BI and analytics workloads that require high performance at any scale. Amazon Redshift also provides the capability to query data stored in Amazon S3 and combine with data stored in the data warehouse. By comparison, Athena is better suited for interactive analysis on any data store without worrying about ingesting and formatting data. Athena analysis is decoupled from storage, so it gives you the flexibility to use other tools and services such as Spark, Flink, and Kafka to further enrich analysis and data processing on the same data analyzed by Athena.

Data sharing

Q: What is Amazon Redshift data sharing?

Amazon Redshift data sharing lets you share live data in Amazon Redshift to securely and easily share data for read purposes with other Redshift clusters within and across AWS accounts and with AWS analytic services using the data lake. With data sharing, you can instantly query live data from any Redshift cluster as long as they have permissions to access without the complexity and delays associated with data copies and data movement. Amazon Redshift lets you share and query live data across the organization, accounts, and even Regions.
Q: What are the use cases for data sharing?

Key use cases include:

- A central ETL cluster sharing data with many BI/analytics clusters to provide read workload isolation and optional charge-ability.
- A data provider sharing data to external consumers.
- Sharing common datasets such as customers, products across different business groups and collaborating for broad analytics and data science.
- Decentralizing a data warehouse to simplify management.
- Sharing data between development, test, and production environments.
- Accessing Redshift data from other AWS analytic services.

Q: What are cross-database queries in Amazon Redshift?

With cross-database queries, you can seamlessly query and join data from any Redshift database that you have access to, regardless of which database you are connected to. This can include databases local on the cluster and also shared datasets made available from remote clusters. Cross-database queries give you flexibility to organize data as separate databases to support multi-tenant configurations.
Q: What is AWS Data Exchange for Amazon Redshift?

AWS Data Exchange for Amazon Redshift lets you find and subscribe to third-party data in AWS Data Exchange that you can query in a Redshift data warehouse in minutes. You can also easily license your data in Amazon Redshift through AWS Data Exchange. Access is automatically granted when a customer subscribes to your data and automatically revoked when their subscription ends, invoices are automatically generated, and payments are automatically collected and disbursed through AWS. This feature empowers you to quickly query, analyze, and build applications with third-party data.
Q: Who are the primary users of AWS Data Exchange?

AWS Data Exchange makes it easy for AWS customers to securely exchange and use third-party data in AWS. Data analysts, product managers, portfolio managers, data scientists, quants, clinical trial technicians, and developers in nearly every industry would like access to more data to drive analytics, train ML models, and make data-driven decisions. But there is no one place to find data from multiple providers and no consistency in how providers deliver data, leaving them to deal with a mix of shipped physical media, FTP credentials, and bespoke API calls. Conversely, many organizations would like to make their data available for research or commercial purposes, but it's too hard and expensive to build and maintain data delivery, entitlement, and billing technology, which further depresses the supply of valuable data.
Q: What AWS Regions is AWS Data Exchange available in?

AWS Data Exchange has a single, globally available product catalog offered by providers. You can see the same catalog regardless of which Region you are using. The resources underlying the product (datasets, revisions, and assets) are regional resources that you manage programmatically or through the AWS Data Exchange console in specific AWS Regions. See the AWS Regional Availability Table for a list of AWS Regions in which AWS Data Exchange is available today.
Q: What is the difference between AWS Data Exchange and the Registry of Open Data on AWS?

There are five key differences between AWS Data Exchange and the Registry of Open Data on AWS:

- First, AWS Data Exchange supports both free and commercial data products, with any applicable commercial fees applied to your AWS invoice. The Registry of Open Data on AWS gives you access to a curated list of free and open datasets.
- Second, AWS Data Exchange requires you to explicitly agree to the Data Subscription Agreement outlining the terms that the data provider set when publishing their product. The data on the Registry of Open Data on AWS does not have terms of use.
- Third, you must use the AWS Data Exchange API to copy data from AWS Data Exchange to your desired Amazon S3 location. The Registry of Open Data on AWS datasets are accessed via S3 APIs.
- Fourth, AWS Data Exchange delivers data providers access to daily, weekly, and monthly reports detailing subscription activity. With the Registry of Open Data on AWS, data providers must analyze their own logs to track usage of data.
- Finally, to become a data provider on AWS Data Exchange, qualified customers must register as a data provider on AWS Marketplace to be eligible to list both free and

commercial products. However, any customer can add free data to the Registry of Open Data on AWS through GitHub and may apply to the AWS Public Dataset Program for AWS to sponsor the costs of storage and bandwidth for select open datasets.

Q: What is Amazon Redshift Query Editor V2?

Amazon Redshift Query Editor v2 is a web-based SQL client application that you can use to author and run queries on your Redshift data warehouse. You can visualize query results with charts and collaborate by sharing queries with members of your team. Query Editor v2 provides several capabilities, such as the ability to browse and explore multiple databases, external tables, views, stored procedures, and user-defined functions. It provides wizards to create schemas, tables, and user-defined functions. You can also load data in Amazon Redshift from Amazon S3 using a visual wizard. It simplifies the management and collaboration of saved queries. You can also gain faster insights by visualizing the results with a single click. With the latest preview release, data analysts can share their queries and collaborate through a common interface called the Query Doc that lets them embed code/SQL queries, annotations, results, and visualizations.

Q: Why should I use Query Editor V2?

If you are a data analyst, data scientist, or data engineer, you can now use Query Editor V2 to browse, create schemas and tables, load data, and author SQL queries, stored procedures, and UDFs through a web-based interface. You can also perform a visual analysis of data in-place without having to leave the tool. You can also schedule your long-running queries or queries for a simple reporting purpose such as daily reporting.

Q: What are the features included in Query Editor v2?

Query Editor v2 allows you to:

- Visually create schemas, tables, and load data from Amazon S3.
- Author queries and gain faster insights with an intuitive editor for authoring SQL queries.
- Perform analysis of results and download results in JSON/CSV formats to your desktop.
- Automatically manage different versions of queries.
- Collaborate with other users to share queries, analysis, and results.
- Run queries in the background even if the browser is closed.

Scalability and concurrency

Q: How do I scale the size and performance of my Amazon Redshift data warehouse cluster?

If you would like to increase query performance or respond to CPU, memory, or I/O overutilization, you can increase the number of nodes within your data warehouse cluster using Elastic Resize through the AWS Management Console or the ModifyCluster API. When you modify your data warehouse cluster, your requested changes will be applied immediately. Metrics for compute utilization, storage utilization, and read/write traffic to your Redshift data warehouse cluster are available free of charge through the AWS Management Console or Amazon CloudWatch APIs. You can also add user-defined metrics via Amazon CloudWatch custom metric functionality.

With the Concurrency Scaling feature, you can support virtually unlimited concurrent users and concurrent queries, with consistently fast query performance. When concurrency scaling is enabled, Amazon Redshift automatically adds cluster capacity when your cluster experiences increase in query queueing.

With Amazon Redshift Spectrum, you can run multiple Redshift clusters accessing the same data in Amazon S3. You can use different clusters for different use cases. For example, you can use one cluster for standard reporting and another for data science queries. Your marketing team can use their own clusters different from your operations team. Redshift Spectrum automatically distributes the execution of your query to several Redshift Spectrum workers out of a shared resource pool to read and process data from Amazon S3, and pulls results back into your Redshift cluster for any remaining processing.
Q: Will my data warehouse cluster remain available during scaling?

It depends. When you using the Concurrency Scaling feature, the cluster is fully available for read and write during concurrency scaling. With Elastic resize, the cluster is unavailable for four to eight minutes of the resize period. With the Redshift RA3 storage elasticity in managed storage, the cluster is fully available and data is automatically moved between managed storage and compute nodes.
Q: When should I use concurrency scaling and when should I use data sharing?

Data sharing and concurrency scaling are complementary features. With concurrency scaling, Amazon Redshift allows you to auto-scale one or more workloads in a single cluster to handle high concurrency and query spikes. Amazon Redshift elastically and automatically spins up the capacity in seconds to deal with the bursts of user activity and brings it down when activity subsides. Applications continue to interact with Amazon Redshift using a single application endpoint. Data sharing allows you to scale to diverse workloads with multi-cluster, multi-account deployments. This allows for workload isolation and charge-ability, cross-group collaboration in decentralized environments, and the ability to offer data as a service to internal and external stakeholders. You can enable concurrency scaling on both data sharing producer clusters and consumer clusters.
Q: How do I manage resources to ensure that my Amazon Redshift cluster can provide consistently fast performance during periods of high concurrency?

A typical data warehouse has significant variance in concurrent query usage over the course of a day. It is more cost-effective to add resources just for the period during which they are required rather than provisioning to peak demand. Amazon Redshift handles this automatically on your behalf.

Concurrency Scaling is a feature in Amazon Redshift that provides consistently fast query performance, even with thousands of concurrent queries. With this feature, Amazon Redshift automatically adds transient capacity when needed to handle heavy demand. Amazon Redshift automatically routes queries to scaling clusters, which are provisioned in seconds and begin processing queries immediately.

This feature is free for most customers. Each Amazon Redshift cluster earns up to one hour of free Concurrency Scaling credits per day. This gives you predictability in your month-to-month cost, even during periods of fluctuating analytical demand.
Q: What is Elastic Resize and how is it different from Concurrency Scaling?

Elastic Resize adds or removes nodes from a single Redshift cluster within minutes to manage its query throughput. For example, an ETL workload for certain hours in a day or month-end reporting may need additional Amazon Redshift resources to complete on time. Concurrency Scaling adds additional cluster resources to increase the overall query concurrency.
Q: Can I access the Concurrency Scaling clusters directly?

No. Concurrency Scaling is a massively scalable pool of Amazon Redshift resources and customers do not have direct access.

Data integration and loading

Q: How do I load data into my Amazon Redshift data warehouse?

You can load data into Amazon Redshift from a range of data sources including Amazon S3, Amazon RDS, Amazon DynamoDB, Amazon EMR, AWS Glue, AWS Data Pipeline and or any SSH-enabled host on Amazon EC2 or on-premises. Amazon Redshift attempts to load your data in parallel into each compute node to maximize the rate at which you can ingest data into your data warehouse cluster. Clients can connect to Amazon Redshift using ODBC or JDBC and issue 'insert' SQL commands to insert the data. Please note this is slower than using S3 or DynamoDB since those methods load data in parallel to each compute node while SQL insert statements load via the single leader node. For more details on loading data into Amazon Redshift, please view our Getting Started Guide.
Q: How do I load data from my existing Amazon RDS, Amazon EMR, Amazon DynamoDB, and Amazon EC2 data sources to Amazon Redshift?

You can use our COPY command to load data in parallel directly to Amazon Redshift from Amazon EMR, Amazon DynamoDB, or any SSH-enabled host. Amazon Redshift Spectrum also lets you load data from Amazon S3 into your cluster with a simple INSERT INTO command. With this, you could load data from various formats such as Parquet and ORC into your cluster. Note that if you use this approach, you will accrue Redshift Spectrum charges for the data scanned from Amazon S3.

AWS Data Pipeline provides a high performance, reliable, fault tolerant solution to load data from a variety of AWS data sources like Amazon RDS to Redshift. You can use AWS Data Pipeline to specify the data source and desired data transformations, and then run a prewritten import script to load your data into Amazon Redshift. In addition, AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy to prepare and load data for analytics. You can create and run an AWS Glue ETL job with a few clicks in the AWS Management Console. Further, many ETL companies have certified Amazon Redshift for use with their tools, and a number are offering free trials to help you get started loading your data. Some of these features have also implemented deeper integration with the Redshift console for easier discovery and monitoring of data pipelines into Amazon Redshift from a large variety of third-party sources.
Q: I have a lot of data for initial loading into Amazon Redshift. Transferring via the Internet would take a long time. How do I load this data?

You can use AWS Snowball to transfer the data to Amazon S3 using portable storage devices. In addition, you can use AWS Direct Connect to establish a private network connection between your network or data center and AWS. You can choose 1Gbit/sec or 10Gbit/sec connection ports to transfer your data.

Security

Q: How does Amazon Redshift keep my data secure?

Amazon Redshift supports industry-leading security with built-in AWS IAM integration, identity federation for single-sign on (SSO), multi-factor authentication, column-level access control, row-level security, Amazon Virtual Private Cloud (Amazon VPC), and provides built-in AWS KMS integration to protect your data in transit and at rest. Amazon Redshift encrypts and keeps your data secure in transit and at rest using industry-standard encryption techniques. To keep data secure in transit, Amazon Redshift supports SSL-enabled connections between your client application and your Redshift data warehouse cluster. To keep your data secure at rest, Amazon Redshift encrypts each block using hardware-accelerated AES-256 as it is written to disk. This takes place at a low level in the I/O subsystem, which encrypts everything written to disk, including intermediate query results. The blocks are backed up as is, which means that

backups are encrypted as well. By default, Amazon Redshift takes care of key management but you can choose to manage your keys through AWS Key Management Service. All Amazon Redshift security features are offered at no additional costs. Redshift Spectrum supports Amazon S3's Server Side Encryption (SSE) using your account's default key managed used by the AWS Key Management Service (KMS).
Q: Does Redshift support granular access controls like column level security?

Yes, Amazon Redshift provides support for role-based access control. Row-level access control allows you to assign one or more roles to a user, and assign system and object privileges by role. You can use out-of-the-box system roles–root user, dba, operator, and security admins, or create your own roles.
Q: Does Amazon Redshift support data masking or data tokenization?

AWS Lambda user-defined functions (UDFs) enable you to use an AWS Lambda function as a UDF in Amazon Redshift and invoke it from Redshift SQL queries. This functionality enables you to write custom extensions for your SQL query to achieve tighter integration with other services or third-party products. You can write Lambda UDFs to enable external tokenization, data masking, identification or de-identification of data by integrating with vendors like Protegrity, and protect or unprotect sensitive data based on a user's permissions and groups, in query time.
Q: Does Amazon Redshift support single sign-on?

Yes. Customers who want to use their corporate identity providers such as Microsoft Azure Active Directory, Active Directory Federation Services, Okta, Ping Federate, or other SAML compliant identity providers can configure Amazon Redshift to provide single-sign on.
Q: How does Amazon Redshift support single sign-on with Microsoft Azure Active Directory?

You can sign-on to Amazon Redshift cluster with Microsoft Azure Active Directory (AD) identities. This allows you to be able to sign-on to Redshift without duplicating Azure Active Directory identities in Redshift.
Q: Does Amazon Redshift support multi-factor authentication (MFA)?

Yes. You can use multi-factor authentication (MFA) for additional security when authenticating to your Amazon Redshift cluster.
Q: Can I use Amazon Redshift in Amazon Virtual Private Cloud (Amazon VPC)?

Yes. You can use Amazon Redshift as part of your VPC configuration. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own data center. This gives you complete control over who can access your Redshift data warehouse cluster. You can use Redshift Spectrum with a Redshift cluster that is part of your Amazon VPC.

Amazon Redshift supports managed VPC endpoints (powered by AWS PrivateLink) to connect to your Redshift cluster in a VPC. With an Amazon Redshift–managed endpoint, you can privately access your Redshift data warehouse within your VPC from your client applications in another VPC within the same or another AWS account and running on premises without using public IPs or requiring traffic to traverse the Internet.
Q: Can I access my Amazon Redshift compute nodes directly?

No. Your Amazon Redshift compute nodes are in a private network space and can only be accessed from your data warehouse cluster's leader node. This provides an additional layer of security for your data.
Q: Does Redshift support role-based access control in the database? (Pre-announcement)

Amazon Redshift will provide support for role-based access control soon.

Availability and durability

Q: What happens to my data warehouse cluster availability and data durability if a drive on one of my nodes fails?

Amazon Redshift will detect a drive or node failure in these cases and replace the cluster node automatically. On Dense Compute (DC) and Dense Storage (DS2) clusters, the data is stored on the compute nodes to ensure high data durability. When a node is replaced, the data is refreshed from the mirror copy on the other node.

RA3 clusters and Redshift serverless are not impacted the same way since the data is stored in Amazon S3 and the local drive is just used as a data cache. In the event of a node replacement, the data is retrieved from Amazon S3. Amazon S3 provides a 99.9999% data durability guarantee. In the event of a multi-node or complete cluster failure, an up-to-date copy of the data is available in S3 and the cluster can be recovered in the same AZ or another AZ without any data loss.

The data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the database. Amazon Redshift makes your replacement node available immediately and loads your most frequently accessed data from Amazon S3 on RA3 and serverless, and from the mirror on DS2 and Amazon Dense Compute (DC2). Single node DC2 and DS2 clusters do not support data replication. In the event of a drive failure, you will need to restore the cluster from a snapshot on S3. Single node RA3.XLPLUS clusters can be recreated without any data loss using the data stored in S3 with the assistance of AWS Support. We recommend using at least two nodes for production to maximize availability.
Q: What happens to my data warehouse cluster availability and data durability in the event of individual node failure?

Amazon Redshift will automatically detect and replace a failed node in your data warehouse cluster. The data warehouse cluster will be unavailable for queries and updates until a replacement node is provisioned and added to the DB. Amazon Redshift makes your replacement node available immediately and loads your most frequently accessed data from S3 first to allow you to resume querying your data as quickly as possible. Single node clusters do not support data replication. In the event of a drive failure, you will need to restore the cluster from snapshot on S3. We recommend using at least two nodes for production.
Q: What happens to my data warehouse cluster availability and data durability if my data warehouse cluster's Availability Zone (AZ) has an outage?

If your Amazon Redshift data warehouse cluster's Availability Zone becomes unavailable, Amazon Redshift will automatically move your cluster to another AWS Availability Zone (AZ) without any data loss or application changes. To activate this, you must enable the relocation capability in your cluster configuration settings.
Q: Does Amazon Redshift support Multi-AZ Deployments?

Currently, Amazon Redshift only supports Single-Region deployments. To set up a disaster recovery (DR) configuration, you can enable cross-Region snapshot copy on your cluster. This will replicate all snapshots from your cluster to another AWS Region. In the event of a DR event, the snapshots in the replica Region can be restored to create a new cluster. Amazon Redshift also supports cross-Region data sharing, where a consumer cluster can access live data in a producer cluster in another region. This is supported only with Amazon Redshift Serverless and RA3.

Querying and analytics

Q: Are Amazon Redshift and Redshift Spectrum compatible with my preferred business intelligence software package and ETL tools?

Yes, Amazon Redshift uses industry-standard SQL and is accessed using standard JDBC and ODBC drivers. You can download Amazon Redshift custom JDBC and ODBC drivers from the Connect Client tab of the Redshift Console. We have validated integrations with popular BI and ETL vendors, a number of which are offering free trials to help you get started loading and analyzing your data. You can also go to the AWS Marketplace to deploy and configure solutions designed to work with Amazon Redshift in minutes.

Amazon Redshift Spectrum supports all Amazon Redshift client tools. The client tools can continue to connect to the Amazon Redshift cluster endpoint using ODBC or JDBC connections. No changes are required.

You use exactly the same query syntax and have the same query capabilities to access tables in Redshift Spectrum as you have for tables in the local storage of your Redshift cluster. External tables are referenced using the schema name defined in the CREATE EXTERNAL SCHEMA command where they were registered.
Q: What data formats and compression formats does Amazon Redshift Spectrum support?

Amazon Redshift Spectrum currently supports many open source data formats, including Avro, CSV, Grok, Amazon Ion, JSON, ORC, Parquet, RCFile, RegexSerDe, Sequence, Text, and TSV.

Amazon Redshift Spectrum currently supports Gzip and Snappy compression.
Q: What happens if a table in my local storage has the same name as an external table?

Just like with local tables, you can use the schema name to pick exactly which one you mean by using schema_name.table_name in your query.
Q: I use a Hive Metastore to store metadata about my S3 data lake. Can I use Redshift Spectrum?

Yes. The CREATE EXTERNAL SCHEMA command supports Hive Metastores. We do not currently support DDL against the Hive Metastore.
Q: How do I get a list of all external database tables created in my cluster?

You can query the system table SVV_EXTERNAL_TABLES to get that information.
Q: Does Redshift support the ability to use Machine Learning with SQL?

Yes, the Amazon Redshift ML feature makes it easy for SQL users to create, train, and deploy machine learning (ML) models using familiar SQL commands. Amazon Redshift ML allows you to leverage your data in Amazon Redshift with Amazon SageMaker, a fully managed ML service. Amazon Redshift supports both unsupervised learning (K-Means) and supervised learning (Autopilot, XGBoost, MLP algorithms). You can also use AWS Language AI services to translate, redact, and analyze text fields in SQL queries with prebuilt Lambda UDF functions - see blog post.
Q: Does Amazon Redshift provide an API to query data?

Amazon Redshift provides a Data API that you can use to painlessly access data from Amazon Redshift with all types of traditional, cloud-native, and containerized, serverless web services-based and event-driven applications. The Data API simplifies access to Amazon Redshift because you don't need to configure drivers and manage database connections. Instead, you can run SQL commands to an Amazon Redshift cluster by simply calling a secured API endpoint provided by the Data API. The Data API takes care of managing database connections and buffering data. The Data API is asynchronous, so you can retrieve your results later. Your query results are stored for 24 hours.
Q: What types of credentials can I use with Amazon Redshift Data API?

The Data API supports both IAM credentials and using a secret key from AWS Secrets Manager. The Data API federates AWS Identity and Access Management (IAM) credentials so

you can use identity providers like Okta or Azure Active Directory or database credentials stored in Secrets Manager without passing database credentials in API calls.
Q: Can I use Amazon Redshift Data API from AWS CLI?

Yes, you can use the Data API from AWS CLI using the aws redshift-data command line option.
Q: Is the Redshift Data API integrated with other AWS services?

You can use the Data API from other services such as AWS Lambda, AWS Cloud9, AWS AppSync and Amazon EventBridge.
Q: Do I have to pay separately for using the Amazon Redshift Data API?

No, there is no separate charge for using the Data API.

Backup and restore

Q: How does Amazon Redshift backup my data? How do I restore my cluster from a backup?

Amazon Redshift RA3 clusters and Amazon Redshift Serverless use Redshift Managed Storage, which always has the latest copy of the data available. DS2 and DC2 clusters mirror the data on the cluster to ensure the latest copy is available in the event of a failure. Backups are automatically created on all Redshift cluster types and retained for 24 hours, and on serverless recovery points are provided for the past 24 hours.

You can also create your own backups that can be retained indefinitely. These backups can be created at any time, and the Amazon Redshift automated backups or Amazon Redshift Serverless recovery points can be converted into a user backup for longer retention.

Amazon Redshift can also asynchronously replicate your snapshots or recovery points to Amazon S3 in another Region for disaster recovery.

On a DS2 or DC2 cluster, free backup storage is limited to the total size of storage on the nodes in the data warehouse cluster and only applies to active data warehouse clusters.

For example, if you have total data warehouse storage of 8 TB, we will provide at most 8 TB of backup storage at no additional charge. If you would like to extend your backup retention period beyond one day, you can do so using the AWS Management Console or the Amazon Redshift APIs. For more information on automated snapshots, please refer to the Amazon Redshift Management Guide.

Amazon Redshift only backs up data that has changed, so most snapshots use only a small amount of your free backup storage. When you need to restore a backup, you have access to all the automated backups within your backup retention window. Once you choose a backup from which to restore, we will provision a new data warehouse cluster and restore your data to it.
Q: How do I manage the retention of my automated backups and snapshots?

You can use the AWS Management Console or ModifyCluster API to manage the period of time your automated backups are retained by modifying the RetentionPeriod parameter. If you wish to turn off automated backups altogether, you can set up the retention period to 0 (not recommended).
Q: What happens to my backups if I delete my data warehouse cluster?

When you delete a data warehouse cluster you have the ability to specify whether a final snapshot is created upon deletion. This enables a restore of the deleted data warehouse cluster at a later date. All previously created manual snapshots of your data warehouse cluster will be retained and billed at standard Amazon S3 rates, unless you choose to delete them.

Monitoring and maintenance

Q: How do I monitor the performance of my Amazon Redshift data warehouse cluster?

Metrics for compute utilization, storage utilization, and read/write traffic to your Amazon Redshift data warehouse cluster are available free of charge via the AWS Management Console or Amazon CloudWatch APIs. You can also add additional, user-defined metrics via Amazon CloudWatch's custom metric functionality. The AWS Management Console provides a monitoring dashboard that helps you monitor the health and performance of all your clusters. Amazon Redshift also provides information on query and cluster performance via the AWS Management Console. This information enables you to see which users and queries are consuming the most system resources to diagnose performance issues by viewing query plans and execution statistics. In addition, you can see the resource utilization on each of your compute nodes to ensure that you have data and queries that are well-balanced across all nodes.
Q: What is a maintenance window? Will my data warehouse cluster be available during software maintenance?

Amazon Redshift periodically performs maintenance to apply fixes, enhancements and new features to your cluster. You can change the scheduled maintenance windows by modifying the cluster, either programmatically or by using the Redshift Console. During these maintenance windows, your Amazon Redshift cluster is not available for normal operations. For more information about maintenance windows and schedules by region, see Maintenance Windows in the Amazon Redshift Management Guide.

AWS Lake Formation

Build, manage, and secure data lakes in days.

AWS Lake Formation is a managed service that makes it easy to set up, secure, and manage your data lakes. Lake Formation helps you discover your data sources and then catalog, cleanse, and transform the data. You can use Lake Formation to secure and ingest the data in an Amazon Simple Storage Service (Amazon S3) data lake.

How it works

AWS Lake Formation easily creates secure data lakes, making data available for wide-ranging analytics.

Use cases

Build your data lake in days
Quickly import data from all your data sources, and then describe and manage them in a centralized data catalog.

Secure and govern your data lake at scale
Scale permissions more easily with fine-grained security capabilities, including row- and cell-level permissions and tag-based access control.

Enable self-service analytics across your organization
Centrally manage access to available datasets and apply fine-grained permissions for all data users.

Build data mesh with minimal data movement
Deploy a data mesh or data fabric, or simplify cross-account data sharing in your organization.

What is AWS Lake Formation?

AWS Lake Formation is a fully managed service that makes it easy to build, secure, and manage data lakes. Lake Formation simplifies and automates many of the complex manual steps that are usually required to create data lakes. These steps include collecting, cleansing, moving, and cataloging data, and securely making that data available for analytics and machine learning.

Lake Formation provides its own permissions model that augments the IAM permissions model. This centrally defined permissions model enables fine-grained access to data stored in data lakes through a simple grant or revoke mechanism, much like a relational database management system (RDMS). Lake Formation permissions are enforced using granular controls at the column, row, and cell-levels across AWS analytics and machine learning services, including Amazon Athena, Amazon QuickSight, and Amazon Redshift.

Lake Formation features

Lake Formation helps you break down data silos and combine different types of structured and unstructured data into a centralized repository. First, identify existing data stores in Amazon S3 or relational and NoSQL databases, and move the data into your data lake. Then crawl, catalog, and prepare the data for analytics. Next, provide your users with secure self-service access to the data through their choice of analytics services.

Setup and data management

Import data from databases already in AWS

Once you specify where your existing databases are and provide your access credentials, Lake Formation reads the data and its metadata (schema) to understand the contents of the data source. It then imports the data to your new data lake and records the metadata in a central catalog. With Lake Formation, you can import data from MySQL, PostgreSQL, SQL Server, MariaDB, and Oracle databases running in Amazon RDS or hosted in Amazon EC2. Both bulk and incremental data loading are supported.

Import data from other external sources

You can use Lake Formation to move data from on-premises databases by connecting with Java Database Connectivity (JDBC). Identify your target sources and provide access credentials in the console, and Lake Formation reads and loads your data into the data lake. To import data from databases other than the ones listed above, you can create custom ETL jobs with AWS Glue.

Catalog and label your data

Lake Formation crawls and reads your data sources to extract technical metadata and creates a searchable catalog to describe this information for users so they can discover available datasets. You can also add your own custom labels to your data (at the table and column level) to define attributes, such as “sensitive information” and “European sales data.” Lake Formation provides a text-based search over this metadata so your users can quickly find the data they need to analyze. For more information about adding tables to the Data Catalog, see Managing Data Catalog Tables and Databases.

Transform data

Lake Formation can perform transformations on your data, such as rewriting various date formats for consistency to ensure that the data is stored in an analytics-friendly fashion. Lake Formation creates transformation templates and schedules jobs to prepare your data for analysis. Your data is transformed with AWS Glue and written in columnar formats, such as Parquet and ORC, for better performance.

Clean and deduplicate data

Lake Formation helps clean and prepare your data for analysis by providing a machine learning transform called FindMatches for deduplication and finding matching records. For example, use FindMatches to find duplicate records in your database of restaurants, such as when one record lists “Joe's Pizza” at “121 Main

St.” and another shows “Joseph's Pizzeria” at “121 Main.” FindMatches will simply ask you to label sets of records as either “matching” or “not matching.” The system will then learn your criteria for calling a pair of records a match and will build an machine learning transform that you can use to find duplicate records within a database or matching records across two databases.For more information about FindMatches, see Matching Records with AWS Lake Formation FindMatches in the *AWS Glue Developer Guide*.

Storage optimizations

Analytics performance can be impacted by inefficient storage of many small files that are automatically created as new data is written to the data lake. Processing these many small files creates additional overhead for analytics services and causes slower query responses. Lake Formation includes a storage optimizer that automatically combines small files into larger files to speed up queries by up to 7x. This process, commonly known as compaction, is performed in the background so that there is no performance impact on your production workloads while this is taking place. For more information about the storage optimization features of Lake Formation, see Storage optimizations for governed tables.

Row and cell-level security

Lake Formation provides data filters that allow you to restrict access to a combination of columns and rows. Use row and cell-level security to protect sensitive data like Personal Identifiable Information (PII). For more information about row-level security, see Overview of data filtering.

Security management

Define and manage access controls

Lake Formation provides a single place to manage access controls for data in your data lake. You can define security policies that restrict access to data at the database, table, column, row, and cell levels. These policies apply to IAM users and roles, and to users and groups when federating through an external identity provider. You can use fine-grained controls to access data secured by Lake Formation within Amazon Redshift Spectrum, Athena, AWS Glue ETL, and Amazon EMR for Apache Spark.

Implement audit logging

Lake Formation provides comprehensive audit logs with CloudTrail to monitor access and show compliance with centrally defined policies. You can audit data access history across analytics and machine learning services that read the data in your data lake via Lake Formation. This lets you see which users or roles have attempted to access what data, with which services, and when. You can access audit logs in the same way you access any other CloudTrail logs using the CloudTrail APIs and console. For more information about CloudTrail logs see Logging AWS Lake Formation API Calls Using AWS CloudTrail.

Tag-based access control

You can classify your data and limit access to sensitive information.You can also add your own custom labels (LF-tags) to the data at the table- and column-level to define attributes, like “sensitive information” or “European sales data.” Lake Formation provides a text-based search over this metadata, so your users can quickly find the data they need to analyze. You can grant access to the data based on these LF-tags. For more information about tag-based access control, see Lake Formation Tag-based access control.

Cross account access

Lake Formation permission management capabilities simplify securing and managing distributed data lakes across multiple AWS accounts through a centralized approach, providing fine-grained access control to the Data Catalog and Amazon S3locations.

Governed tables

Data lakes need to show users the correct view of data at all times, even while there are simultaneous real-time or frequent updates to the data. Loading streaming data or incorporating changes from multiple source data systems requires processing inserts and deletes across multiple tables in parallel. Today, developers write custom application code or use open source tools to manage these updates. These solutions are complex and difficult to scale because writing application code that maintains consistency when concurrently reading and writing the same data is tedious, brittle, and error prone.

Lake Formation introduces new APIs that support atomic, consistent, isolated, and durable (ACID) transactions using a new data lake table type, called a *governed table*. A governed table allows multiple users to concurrently insert and delete data across tables using manifests, while still allowing other users to simultaneously run

analytical queries and ML models on the same data sets that return consistent and up-to-date results.

For more information about using transactions with Lake Formation, see the following topics:

- Governed tables in Lake Formation
- Reading from and writing to the data lake within transactions

AWS service integrations with Lake Formation

The following AWS services integrate with AWS Lake Formation and honor Lake Formation permissions.

AWS Service	How Integrated
AWS Glue	AWS Glue and Lake Formation share the same Data Catalog. For console operations (such as viewing a list of tables) and all API operations, AWS Glue users can access only the databases and tables on which they have Lake Formation permissions. **Note** AWS Glue does not support Lake Formation column permissions.
Amazon Athena	When Amazon Athena users select the AWS Glue catalog in the query editor, they can query only the databases, tables, and columns that they have Lake Formation permissions on. Queries using manifests are not supported. In addition to principals who authenticate with Athena through AWS Identity and Access Management (IAM), Lake Formation supports Athena users who connect through the JDBC or ODBC driver and authenticate through SAML. Supported SAML providers include Okta and Microsoft Active Directory Federation Service (AD FS). For more information, see Using Lake Formation and

AWS Service	How Integrated
	the Athena JDBC and ODBC Drivers for Federated Access to Athena in the *Amazon Athena User Guide*. **Note** Currently, authorizing access to SAML identities in Lake Formation is not supported in the following regions: • Middle East (Bahrain) - me-south-1 • Asia Pacific (Hong Kong) - ap-east-1 • Africa (Cape Town) - af-south-1 • China (Ningxia) - cn-northwest-1 • Asia Pacific (Osaka) - ap-northeast-3
Amazon Redshift Spectrum	When Amazon Redshift users create an external schema on a database in the AWS Glue catalog, they can query only the tables and columns in that schema on which they have Lake Formation permissions. Queries using manifests are not supported.
Amazon QuickSight Enterprise Edition	When an Amazon QuickSight Enterprise Edition user queries a dataset in an Amazon S3 location that is registered with Lake Formation, the user must have the Lake Formation SELECT permission on the data.
Amazon EMR	Lake Formation permissions are enforced when Apache Spark applications are submitted using Apache Zeppelin or EMR Notebooks.

Lake Formation also works with AWS Key Management Service (AWS KMS) to enable you to more easily set up these integrated services to encrypt and decrypt data in Amazon Simple Storage Service (Amazon S3) locations.

AWS Lake Formation FAQs

General

Q: What is a data lake?

A data lake is a scalable central repository of large quantities and varieties of data, both structured and unstructured. Data lakes let you manage the full lifecycle of your data. The first

step of building a data lake is ingesting and cataloging data from a variety of sources. The data is then enriched, combined, and cleaned before analysis. This makes it easy to discover and analyze the data with direct queries, visualization, and machine learning (ML). Data lakes complement traditional data warehouses, providing more flexibility, cost-effectiveness, and scalability for ingestion, storage, transformation, and analysis of your data. The traditional challenges around the construction and maintenance of data warehouses and limitations in the types of analysis can be overcome using data lakes.

Read more about "What is a data lake?"
Q: What is AWS Lake Formation?

Lake Formation is an integrated data lake service that makes it easy for you to ingest, clean, catalog, transform, and secure your data and make it available for analysis and ML. Lake Formation gives you a central console where you can discover data sources, set up transformation jobs to move data to an Amazon Simple Storage Service (S3) data lake, remove duplicates and match records, catalog data for access by analytic tools, configure data access and security policies, and audit and control access from AWS analytic and ML services. Lake Formation automatically manages access to the registered data in Amazon S3 through services including AWS Glue, Amazon Athena, Amazon Redshift, Amazon QuickSight, and Amazon EMR using Zeppelin notebooks with Apache Spark to ensure compliance with your defined policies. If you've set up transformation jobs spanning AWS services, Lake Formation configures the flows, centralizes their orchestration, and lets you monitor the jobs. With Lake Formation, you can configure and manage your data lake without manually integrating multiple underlying AWS services.
Q: Why should I use Lake Formation to build my data lake?

Lake Formation makes it easy to build, secure, and manage your AWS data lake. Lake Formation integrates with underlying AWS security, storage, analysis, and ML services and automatically configures them to comply with your centrally defined access policies. It also gives you a single console to monitor your jobs and data transformation and analytic workflows.

Lake Formation can manage data ingestion through AWS Glue. Data is automatically classified, and relevant data definitions, schema, and metadata are stored in the central data catalog. AWS Glue also converts your data to your choice of open data formats to be stored in Amazon S3 and cleans your data to remove duplicates and link records across datasets. Once your data is in your S3 data lake, you can define access policies, including table-and-column-level access controls, and enforce encryption for data at rest. You can then use a wide variety of AWS analytic and ML services to access your data lake. All access is secured, governed, and auditable.
Q: Can I see a presentation on AWS Lake Formation?

Yes. You can watch the full recording of the "Intro to AWS Lake Formation" session from re:Invent.
Q: What kind of problems does the FindMatches ML Transform solve?

FindMatches generally solves record linkage and data deduplication problems. Deduplication is necessary when you're trying to identify records in a database that are conceptually the same but for which you have separate records. This problem is trivial if duplicate records can be identified by a unique key (for instance, if products can be uniquely identified by a UPC Code), but it becomes very challenging when you have to do a "fuzzy match."

Record linkage is basically the same problem as data deduplication, but this term usually means that you're doing a "fuzzy join" of two databases that don't share a unique key rather than deduplicating a single database. As an example, consider the problem of matching a large database of customers to a small database of known fraudsters. FindMatches can be used on both record linkage and deduplication problems.

For instance, Lake Formation's FindMatches ML Transform can help you with the following problems:

Link patient records: Link patient records between hospitals so doctors have more background information and are better able to treat patients. Use FindMatches on separate databases that both contain common fields such as name, birthday, home address, and phone number.

Deduplicate data: Deduplicate a database of movies containing columns such as “Title,” “Plot synopsis,” “Year of release,” “Run time,” and “Cast.” For example, there might be variations in how the title or the cast names are listed, resulting in duplicates rather than a clean dataset.

Group products: Automatically group all related products together in your storefront by identifying equivalent items in an apparel product catalog, where you want to define “equivalent” to mean that they’re the same when ignoring certain differences. For instance, you might consider all pants to be equivalent despite differences in size and color.
Q: How does Lake Formation deduplicate my data?

Lake Formation's FindMatches ML Transform makes it easy to find and link records that refer to the same entity but don’t share a reliable identifier. Before FindMatches, developers would commonly solve data-matching problems deterministically, by writing huge numbers of hand-tuned rules. FindMatches uses ML algorithms behind the scenes to learn how to match records according to each developer's business criteria. FindMatches first identifies records for you to label as to whether they match or don’t match and then uses ML to create an ML Transform. You can then run this Transform on your database to find matching records, or you can ask FindMatches to give you additional records to label to push your ML Transform to higher levels of accuracy.
Q: What are ML Transforms?

ML Transforms provide a destination for creating and managing machine-learned transforms. Once created and trained, these ML Transforms can be run in standard AWS Glue scripts. You select an algorithm (for example, the FindMatches ML Transform) and input datasets and training examples, and the tuning parameters needed by that algorithm. AWS Lake Formation uses those inputs to build an ML Transform that can be incorporated into a normal ETL job workflow.
Q: Can I see a presentation on using AWS Lake Formation to find matches and deduplicate records?

Yes. The full recording of the AWS Online Tech Talk "Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation" is available here.
Q: How does Lake Formation relate to other AWS services?

Lake Formation manages data access for registered data that is stored in Amazon S3 and manages query access from AWS Glue, Athena, Redshift, Amazon QuickSight, and EMR using Zeppelin notebooks with Apache Spark through a unified security model and permissions. Lake Formation can ingest data from S3, Amazon RDS databases, and AWS CloudTrail logs, understand their formats, and make data clean and able to be queried. Lake Formation configures the flows, centralizes their orchestration, and lets you monitor the jobs.

Read more about "Data Lakes and Analytics on AWS" including how to build a customized data lake.
Q: How does Lake Formation relate to AWS Glue?

Lake Formation uses a shared infrastructure with AWS Glue, including console controls, ETL code creation and job monitoring, blueprints to create workflows for data ingest, the same data catalog, and a serverless architecture. Although AWS Glue focuses on these types of functions, Lake Formation encompasses all AWS Glue features and provides additional capabilities

designed to help build, secure, and manage a data lake. See the AWS Glue features page for more details.

ETL and catalog

Q: How does Lake Formation help me discover the data I can move into my data lake?

Lake Formation automatically discovers all AWS data sources to which it is provided access by your AWS IAM policies. It crawls Amazon S3, Amazon RDS, and AWS CloudTrail sources, and through blueprints it identifies them to you as data that can be ingested into your data lake. No data is ever moved or made accessible to analytic services without your permission. You can also use AWS Glue to ingest data from other sources, including S3 and Amazon DynamoDB.

You can also define JDBC connections to allow Lake Formation to access your AWS databases and on-premises databases including Oracle, MySQL, Postgres, SQL Server, and MariaDB.

Lake Formation ensures that all your data is described in a central data catalog, giving you one location to browse the data that you have permission to view and query. The permissions are defined in your data access policy and can be set at the table and column level.

In addition to the properties automatically populated by the crawlers, you can add labels (including business attributes such as data sensitivity) at the table or column level, and add field-level comments.
Q: How does Lake Formation organize my data in a data lake?

You can use one of the blueprints available in Lake Formation to ingest data into your data lake. Lake Formation creates Glue workflows that crawl source tables, extract the data, and load it to Amazon S3. In S3, Lake Formation organizes the data for you, setting up partitions and data formats for optimized performance and cost. For data already in S3, you can register those buckets with Lake Formation to manage them.

Lake Formation also crawls your data lake to maintain a data catalog and provides an intuitive user interface for you to search entities (by type, classification, attribute, or free-form text).
Q: How does Lake Formation use machine learning to clean my data?

Lake Formation provides jobs that run ML algorithms to perform deduplication and link matching records. Creating ML Transforms is as easy as selecting your source, selecting a desired transform, and providing training data for the desired changes. Once trained to your satisfaction, the ML Transforms can be run as part of your regular data movement workflows, with no ML expertise required.
Q: What are other ways I can ingest data to AWS for use with Lake Formation?

You can move petabytes to exabytes of data from your data centers to AWS using physical appliances with AWS Snowball, AWS Snowball Edge, and AWS Snowmobile. You can also connect your on-premises applications directly to AWS with AWS Storage Gateway. You can accelerate data transfer using a dedicated network connection between your network and AWS with AWS Direct Connect, or boost long-distance global data transfers using Amazon's globally distributed edge locations with Amazon S3 Transfer Acceleration. Amazon Kinesis also provides a useful way to load streaming data to S3. Lake Formation Data Importers can be set up to perform ongoing ETL jobs and prepare ingested data for analysis.
Q: Can I use my existing data catalog or Hive Metastore with Lake Formation?

Lake Formation provides a way for you to import your existing catalog and metastore into the data catalog. However, Lake Formation requires your metadata to reside in the data catalog to ensure governed access to your data.

Security and governance

Q: How does Lake Formation protect my data?

Lake Formation protects your data by giving you a central location where you can configure granular data access policies that protect your data, regardless of which services are used to access it.

To centralize data access policy controls using Lake Formation, first shut down direct access to your buckets in Amazon S3 so all data access is managed by Lake Formation. Next, configure data protection and access policies using Lake Formation, which enforces those policies across all the AWS services accessing data in your lake. You can configure users and roles and define the data these roles can access, down to the table and column level.

Lake Formation currently supports server-side encryption on S3 (SSE-S3, AES-256). Lake Formation also supports private endpoints in your Amazon Virtual Private Cloud (VPC) and records all activity in AWS CloudTrail, so you have network isolation and auditability.
Q: How does Lake Formation work with AWS IAM?

Lake Formation integrates with IAM so authenticated users and roles can be automatically mapped to data protection policies that are stored in the data catalog. The IAM integration also lets you use Microsoft Active Directory or LDAP to federate into IAM using SAML.
Q: How do I convert an existing table in Amazon S3 to a governed table?

If you have existing Amazon S3–based tables cataloged in the AWS Glue Data Catalog, you can convert them to governed tables by running the AWS Glue blueprint available on the AWS Labs Github page. Additionally, you can create a new governed table and update the manifest information in Lake Formation using the AWS SDK and CLI. The manifest information contains a list of S3 objects and associated metadata that represent the current state of your table. You can also use AWS Glue ETL to read from an existing table and create a copy of it as a Governed Table. This allows you to migrate your applications and users to the Governed Table at your own pace.

Enabling data access

Q: How does Lake Formation help an analyst or data scientist discover what data they can access?

Lake Formation ensures that all your data is described in the data catalog, giving you a central location to browse the data that you have permission to view and query. The permissions are defined in your data access policy and can be set at the table and column level.
Q: Can I use third party business intelligence tools with Lake Formation?

Yes. You can use third-party business applications, such as Tableau and Looker, to connect to your AWS data sources through services such as Athena or Redshift. Access to data is managed by the underlying data catalog, so regardless of which application you use, you’re assured that access to your data is governed and controlled.
Q: Does Lake Formation provide APIs or a CLI?

Yes. Lake Formation provides APIs and a CLI to integrate Lake Formation functionality into your custom applications. Java and C++ SDKs are also available to let you integrate your own data engines with Lake Formation.
Q: What is the AWS Lake Formation Storage API, and why should I use it?

The Lake Formation Storage API, provides a single interface for AWS services, ISV solutions and applications developers to securely and reliably read and write data in the data lake. To

write data, the Storage API exposes ACID (atomic, consistent, isolated, and durable) transactions that allows you to write data into Governed Tables, a new type of Amazon S3 table, in a reliable and consistent manner. To read data, the Storage API allows you to query data in Governed Tables and standard S3 tables secured with Lake Formation fine-grained permissions. The Storage API will automatically enforce permissions before returning the filtered results to the calling application. Access permissions are enforced consistently across a wide range of services and tools.

AWS Lake Formation (AMS SSPS)

AWS Lake Formation is a service that makes it easy to set up a secure data lake in days. A data lake is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. A data lake enables you to break down data silos and combine different types of analytics to gain insights and guide better business decisions.

Creating a data lake with Lake Formation is as simple as defining data sources and what data access and security policies you want to apply. Lake Formation then helps you collect and catalog data from databases and object storage, move the data into your new Amazon S3 data lake, clean and classify your data using machine learning algorithms, and secure access to your sensitive data. Your users can access a centralized data catalog (for details, see AWS Glue FAQs) that describes available data sets and their appropriate usage. Your users then leverage these data sets with their choice of analytics and machine learning services, like Amazon Redshift, Amazon Athena, and (in beta) Amazon EMR for Apache Spark. Lake Formation builds on the capabilities available in AWS Glue.

Lake Formation in AWS Managed Services FAQs

Q: How do I request access to AWS Lake Formation in my AMS account?

Request access by submitting a Management | AWS service | Self-provisioned service | Add (review required) (ct-3qe6io8t6jtny) change type. This RFC provisions the following IAM role to your account: customer_lakeformation_data_analyst_role. After it's provisioned in your account, you must onboard the roles in your federation solution.

Q: What are the restrictions to using AWS Lake Formation in my AMS account?

Full functionality of Lake Formation is available in AMS.

Q: What are the prerequisites or dependencies to using AWS Lake Formation in my AMS account?

Lake Formation integrates with the AWS Glue service, therefore AWS Glue users can access only the databases and tables on which they have Lake Formation permissions. Additionally AWS Athena and Amazon Redshift users can only query the AWS Glue databases and tables on which they have Lake Formation permissions.

Amazon Kinesis

Easily collect, process, and analyze video and data streams in real time.

Kinesis Video Streams

Kinesis Data Streams

Kinesis Data Firehose

Kinesis Data Analytics

Amazon Kinesis

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information. Amazon Kinesis offers key capabilities to cost-effectively process streaming data at any scale, along with the flexibility to choose the tools that best suit the requirements of your application. With Amazon Kinesis, you can ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications. Amazon Kinesis enables you to process and analyze data as it arrives and respond instantly instead of having to wait until all your data is collected before the processing can begin.

Benefits

Real-time
Amazon Kinesis enables you to ingest, buffer, and process streaming data in real-time, so you can derive insights in seconds or minutes instead of hours or days.

Fully managed
Amazon Kinesis is fully managed and runs your streaming applications without requiring you to manage any infrastructure.

Scalable
Amazon Kinesis can handle any amount of streaming data and process data from hundreds of thousands of sources with very low latencies.

Amazon Kinesis capabilities

Kinesis Video Streams

Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), and other processing.

Kinesis Data Streams

Amazon Kinesis Data Streams is a scalable and durable real-time data streaming service that can continuously capture gigabytes of data per second from hundreds of thousands of sources.

Kinesis Data Firehose

Amazon Kinesis Data Firehose is the easiest way to capture, transform, and load data streams into AWS data stores for near real-time analytics with existing business intelligence tools.

Kinesis Data Analytics

Amazon Kinesis Data Analytics is the easiest way to process data streams in real time with SQL or Apache Flink without having to learn new programming languages or processing frameworks.

How it works

Amazon Kinesis Video Streams

Amazon Kinesis Data Streams

Amazon Kinesis Data Firehose

Amazon Kinesis Data Analytics

Use cases

Build video analytics applications

You can use Amazon Kinesis to securely stream video from camera-equipped devices in homes, offices, factories, and public places to AWS. You can then use these video streams for video playback, security monitoring, face detection, machine learning, and other analytics.

Example: Amber alert system

Evolve from batch to real-time analytics

With Amazon Kinesis, you can perform real-time analytics on data that has been traditionally analyzed using batch processing. Common streaming use cases include sharing data between different applications, streaming extract-transform-load, and real-time analytics. For example, you can use Kinesis Data Firehose to continuously load streaming data into your S3 data lake or analytics services.

Example: Clickstream analytics

Build real-time applications

You can use Amazon Kinesis for real-time applications such as application monitoring, fraud detection, and live leader-boards. You can ingest streaming data using Kinesis Data Streams, process it using Kinesis Data Analytics, and emit the results to any data store or application using Kinesis Data Streams with millisecond end-to-end latency. This can help you learn about what your customers, applications, and products are doing right now and react promptly.

Example: Analysis of streaming social media data

Analyze IoT device data

You can use Amazon Kinesis to process streaming data from IoT devices such as consumer appliances, embedded sensors, and TV set-top boxes. You can then use the data to send real-time alerts or take other actions programmatically when a sensor exceeds certain operating thresholds. Use our sample IoT analytics code to build your application. No need to start from scratch.

Example: Sensors in tractor detect need for a spare part and automatically place order

Amazon Kinesis Video Streams

Capture, process, and store media streams for playback, analytics, and machine learning.

Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), playback, and other processing. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest streaming video data from millions of devices. It durably stores, encrypts, and indexes video data in your streams, and allows you to access your data through easy-to-use APIs. Kinesis Video Streams enables you to playback video for live and on-demand viewing, and quickly build applications that take advantage of computer vision and video analytics through integration with Amazon Rekognition Video, and libraries for ML frameworks such as Apache MxNet, TensorFlow, and OpenCV. Kinesis Video Streams also supports WebRTC, an open-source project that enables real-time media streaming and interaction between web browsers, mobile applications, and connected devices via simple APIs. Typical uses include video chat and peer-to-peer media streaming.

Benefits

Stream video from millions of devices
Amazon Kinesis Video Streams provides SDKs that make it easy for devices to securely stream media to AWS for playback, storage, analytics, machine learning, and other processing. Kinesis Video Streams can ingest data from edge devices, smartphones, security cameras, and other data sources such as RADARs, LIDARs, drones, satellites, dash cams, and depth-sensors.

Build real-time vision and video-enabled apps
Easily build applications with real-time computer vision capabilities through integration with Amazon Rekognition Video, and with real-time video analytics capabilities using popular open source machine learning frameworks.

Playback live and recorded video streams
Easily stream live and recorded media from your Kinesis video streams to your browser or mobile application using the Kinesis Video Streams HTTP Live Streaming (HLS) capability.

Extract images for enhanced playback or Machine Learning applications
Amazon Kinesis Video Streams offers APIs and SDKs to help you extract images from your video streams. You can use these images for enhanced playback applications such as thumbnails or enhanced scrubbing, or for use in Machine Learning pipelines. Kinesis Video Streams offers on-demand image extraction via APIs or automated image extraction from metadata tags in ingested video.

Build apps with two-way, real-time media streaming
Amazon Kinesis Video Streams supports the open-source project WebRTC for two-way, real-time media streaming between web browsers, mobile applications, and connected devices. With support for WebRTC, you can use simple APIs to build rich applications like video chat and peer-to-peer data sharing with ultra-low latency and two-way communication between your applications and connected devices.

Secure
Amazon Kinesis Video Streams allows you to control access to your streams using AWS Identity and Access Management (IAM). It helps you protect your data by automatically encrypting the data at rest using AWS Key Management Service (KMS) and in transit using the industry-standard Transport Layer Security (TLS) protocol.

Durable, searchable storage
Amazon Kinesis Video Streams uses Amazon S3 as the underlying data store, which means your data is stored durably and reliably. Kinesis Video Streams enables you to quickly search and retrieve video fragments based on device and service generated timestamps.

No infrastructure to manage
Amazon Kinesis Video Streams manages all the infrastructure for you. You don’t have to worry about configuration, software updates, failures, or scaling infrastructure as the number of streams and consuming applications grows. Kinesis Video Streams handles all the administration and maintenance required to manage your streams, so you can focus your time on building innovative applications.

How it works

Capture, process, and store media streams for playback, analytics, and machine learning.

Build applications with ultra-low latency live streaming and two-way real-time communication.

Use cases

Smart home
With Amazon Kinesis Video Streams, you can easily stream live video and audio from camera-equipped home devices such as doorbells, baby monitors, webcams, and home surveillance systems to AWS. You can then use the streams to build a variety of smart home applications ranging from simple video playback to intelligent lighting, climate control systems, and security monitoring. You can use WebRTC capabilities for two-way, real-time media streaming and interaction for use cases like talking with the person at your doorbell or remotely controlling your camera-enabled robot vacuum from your mobile phone.

Example: Interact with a camera-enabled doorbell from your mobile phone

Smart city

Many cities have installed large numbers of cameras at traffic lights, parking lots, shopping malls, and just about every public venue, capturing video 24/7. You can use Amazon Kinesis Video Streams to securely and cost-effectively ingest, store, and analyze this massive volume of video data to help solve traffic problems, help prevent crime, dispatch emergency responders, and much more.

Example: Amber alert system

Industrial automation

You can use Amazon Kinesis Video Streams to collect a variety of time-encoded data such as RADAR and LIDAR signals, temperature profiles, and depth data from industrial equipment. You can then analyze the data using your favorite machine learning framework including Apache MxNet, TensorFlow, and OpenCV for industrial automation use cases like predictive maintenance. For example, you can predict the lifetime of a gasket or valve and schedule part replacement in advance, reducing downtime and defects in a manufacturing line.

Example: Equipment preventive maintenance

Amazon Kinesis Data Streams

Easily stream data at any scale.

How it works

Amazon Kinesis Data Streams is a serverless streaming data service that makes it easy to capture, process, and store data streams at any scale.

Use cases

Stream log and event data
Ingest and collect terabytes of data per day from application and service logs, clickstream data, sensor data, and in-app user events to power live dashboards, generate metrics, and deliver data into data lakes.

Run real-time analytics
Build applications for high-frequency event data such as clickstream data, and gain access to insights in seconds, not days, using AWS Lambda or Amazon Kinesis Data Analytics.

Power event-driven applications
Quickly pair with AWS Lambda to respond to or adjust immediate occurrences within the event-driven applications in your environment, at any scale.

Amazon Kinesis Data Firehose

Reliably load real-time streams into data lakes, warehouses, and analytics services.

How it works

Amazon Kinesis Data Firehose is an extract, transform, and load (ETL) service that reliably captures, transforms, and delivers streaming data to data lakes, data stores, and analytics services.

Use cases

Stream into data lakes and warehouses
Stream data into Amazon S3 and convert data into required formats for analysis without building processing pipelines.

Boost security
Monitor network security in real time and create alerts when potential threats arise using supported Security Information and Event Management (SIEM) tools.

Build ML streaming applications
Enrich your data streams with machine learning (ML) models to analyze data and predict inference endpoints as streams move to their destination.

Amazon Kinesis Data Analytics

Gain actionable insights from streaming data with serverless, fully managed Apache Flink.

How it works

Amazon Kinesis Data Analytics is the easiest way to transform and analyze streaming data in real time using Apache Flink.

Use cases

Deliver streaming data in seconds
Develop applications that transform and deliver data to Amazon Simple Storage Service (Amazon S3), Amazon OpenSearch Service, and more.

Create real-time analytics
Interactively query and analyze data in real time and continuously produce insights for time-sensitive use cases.

Perform stateful processing
Use long-running, stateful computations to trigger real-time actions like anomaly detection based on historical data trends.

What Is Amazon Kinesis Video Streams?

Amazon Kinesis Video Streams is a fully managed AWS service that you can use to stream live video from devices to the AWS Cloud, or build applications for real-time video processing or batch-oriented video analytics.

Kinesis Video Streams isn't just storage for video data. You can use it to watch your video streams in real time as they are received in the cloud. You can either monitor your live streams in the AWS Management Console, or develop your own monitoring application that uses the Kinesis Video Streams API library to display live video.

You can use Kinesis Video Streams to capture massive amounts of live video data from millions of sources, including smartphones, security cameras, webcams,

cameras embedded in cars, drones, and other sources. You can also send non-video time-serialized data such as audio data, thermal imagery, depth data, RADAR data, and more. As live video streams from these sources into a Kinesis video stream, you can build applications that can access the data, frame-by-frame, in real time for low-latency processing. Kinesis Video Streams is source-agnostic; you can stream video from a computer's webcam using the GStreamer library, or from a camera on your network using real-time streaming protocol (RTSP)".

You can also configure your Kinesis video stream to durably store media data for the specified retention period. Kinesis Video Streams automatically stores this data and encrypts it at rest. Additionally, Kinesis Video Streams time-indexes stored data based on both the producer time stamps and ingestion time stamps. You can build applications that periodically batch-process the video data, or you can create applications that require ad hoc access to historical data for different use cases.

Your custom applications, real-time or batch-oriented, can run on Amazon EC2 instances. These applications might process data using open source deep-learning algorithms, or use third-party applications that integrate with Kinesis Video Streams.

Benefits of using Kinesis Video Streams include the following:

- **Connect and stream from millions of devices** – Kinesis Video Streams enables you to connect and stream video, audio, and other data from millions of devices ranging from consumer smartphones, drones, dash cams, and more. You can use the Kinesis Video Streams producer libraries to configure your devices and reliably stream in real time, or as after-the-fact media uploads.
- **Durably store, encrypt, and index data** – You can configure your Kinesis video stream to durably store media data for custom retention periods. Kinesis Video Streams also generates an index over the stored data based on producer-generated or service-side time stamps. Your applications can easily retrieve specified data in a stream using the time-index.
- **Focus on managing applications instead of infrastructure** – Kinesis Video Streams is serverless, so there is no infrastructure to set up or manage. You don't need to worry about the deployment, configuration, or elastic scaling of the underlying infrastructure as your data streams and number of consuming applications grow and shrink. Kinesis Video Streams automatically does all

the administration and maintenance required to manage streams, so you can focus on the applications, not the infrastructure.

- **Build real-time and batch applications on data streams** – You can use Kinesis Video Streams to build custom real-time applications that operate on live data streams, and create batch or ad hoc applications that operate on durably persisted data without strict latency requirements. You can build, deploy, and manage custom applications: open source (Apache MXNet, OpenCV), homegrown, or third-party solutions via the AWS Marketplace to process and analyze your streams. Kinesis Video Streams `Get` APIs enable you to build multiple concurrent applications processing data in a real-time or batch-oriented basis.
- **Stream data more securely** – Kinesis Video Streams encrypts all data as it flows through the service and when it persists the data. Kinesis Video Streams enforces Transport Layer Security (TLS)-based encryption on data streaming from devices, and encrypts all data at rest using AWS Key Management Service (AWS KMS). Additionally, you can manage access to your data using AWS Identity and Access Management (IAM).
- **Pay as you go** – For more information, see AWS Pricing.

What Is Amazon Kinesis Data Streams?

You can use Amazon Kinesis Data Streams to collect and process large streams of data records in real time. You can create data-processing applications, known as *Kinesis Data Streams applications*. A typical Kinesis Data Streams application reads data from a *data stream* as data records. These applications can use the Kinesis Client Library, and they can run on Amazon EC2 instances. You can send the processed records to dashboards, use them to generate alerts, dynamically change pricing and advertising strategies, or send data to a variety of other AWS services.

Kinesis Data Streams is part of the Kinesis streaming data platform, along with Kinesis Data Firehose, Kinesis Video Streams, and Kinesis Data Analytics.

For more information about AWS big data solutions, see Big Data on AWS. For more information about AWS streaming data solutions, see What is Streaming Data?.

What Can I Do with Kinesis Data Streams?

You can use Kinesis Data Streams for rapid and continuous data intake and aggregation. The type of data used can include IT infrastructure log data, application logs, social media, market data feeds, and web clickstream data. Because the response time for the data intake and processing is in real time, the processing is typically lightweight.

The following are typical scenarios for using Kinesis Data Streams:

Accelerated log and data feed intake and processing

You can have producers push data directly into a stream. For example, push system and application logs and they are available for processing in seconds. This prevents the log data from being lost if the front end or application server fails. Kinesis Data Streams provides accelerated data feed intake because you don't batch the data on the servers before you submit it for intake.

Real-time metrics and reporting

You can use data collected into Kinesis Data Streams for simple data analysis and reporting in real time. For example, your data-processing application can work on metrics and reporting for system and application logs as the data is streaming in, rather than wait to receive batches of data.

Real-time data analytics

This combines the power of parallel processing with the value of real-time data. For example, process website clickstreams in real time, and then analyze site usability engagement using multiple different Kinesis Data Streams applications running in parallel.

Complex stream processing

You can create Directed Acyclic Graphs (DAGs) of Kinesis Data Streams applications and data streams. This typically involves putting data from multiple Kinesis Data Streams applications into another stream for downstream processing by a different Kinesis Data Streams application.

Benefits of Using Kinesis Data Streams

Although you can use Kinesis Data Streams to solve a variety of streaming data problems, a common use is the real-time aggregation of data followed by loading the aggregate data into a data warehouse or map-reduce cluster.

Data is put into Kinesis data streams, which ensures durability and elasticity. The delay between the time a record is put into the stream and the time it can be retrieved (put-to-get delay) is typically less than 1 second. In other words, a Kinesis Data Streams application can start consuming the data from the stream almost immediately after the data is added. The managed service aspect of Kinesis Data Streams relieves you of the operational burden of creating and running a data intake pipeline. You can create streaming map-reduce–type applications. The elasticity of Kinesis Data Streams enables you to scale the stream up or down, so that you never lose data records before they expire.

Multiple Kinesis Data Streams applications can consume data from a stream, so that multiple actions, like archiving and processing, can take place concurrently and independently. For example, two applications can read data from the same stream. The first application calculates running aggregates and updates an Amazon DynamoDB table, and the second application compresses and archives data to a data store like Amazon Simple Storage Service (Amazon S3). The DynamoDB table with running aggregates is then read by a dashboard for up-to-the-minute reports.

The Kinesis Client Library enables fault-tolerant consumption of data from streams and provides scaling support for Kinesis Data Streams applications.

Related Services

For information about using Amazon EMR clusters to read and process Kinesis data streams directly, see Kinesis Connector.

What Is Amazon Kinesis Data Firehose?

Amazon Kinesis Data Firehose is a fully managed service for delivering real-time streaming data to destinations such as Amazon Simple Storage Service (Amazon S3), Amazon Redshift, Amazon OpenSearch Service, Splunk, and any custom HTTP endpoint or HTTP endpoints owned by supported third-party service providers, including Datadog, Dynatrace, LogicMonitor, MongoDB, New Relic, and Sumo Logic. Kinesis Data Firehose is part of the Kinesis streaming data platform, along with Kinesis Data Streams, Kinesis Video Streams, and Amazon Kinesis Data Analytics. With Kinesis Data Firehose, you don't need to write applications or manage resources. You configure your data producers to send data to Kinesis Data Firehose, and it automatically delivers the data to the destination that you specified.

You can also configure Kinesis Data Firehose to transform your data before delivering it.

For more information about AWS big data solutions, see Big Data on AWS. For more information about AWS streaming data solutions, see What is Streaming Data?

Note

Note the latest AWS Streaming Data Solution for Amazon MSK that provides AWS CloudFormation templates where data flows through producers, streaming storage, consumers, and destinations.

Key Concepts

As you get started with Kinesis Data Firehose, you can benefit from understanding the following concepts:

Kinesis Data Firehose delivery stream

The underlying entity of Kinesis Data Firehose. You use Kinesis Data Firehose by creating a Kinesis Data Firehose delivery stream and then sending data to it. For more information, see Creating an Amazon Kinesis Data Firehose Delivery Stream and Sending Data to an Amazon Kinesis Data Firehose Delivery Stream.

record

The data of interest that your data producer sends to a Kinesis Data Firehose delivery stream. A record can be as large as 1,000 KB.

data producer

Producers send records to Kinesis Data Firehose delivery streams. For example, a web server that sends log data to a delivery stream is a data producer. You can also configure your Kinesis Data Firehose delivery stream to automatically read data from an existing Kinesis data stream, and load it into destinations. For more information, see Sending Data to an Amazon Kinesis Data Firehose Delivery Stream.

buffer size and buffer interval

Kinesis Data Firehose buffers incoming streaming data to a certain size or for a certain period of time before delivering it to destinations. **Buffer Size** is in MBs and **Buffer Interval** is in seconds.

Data Flow

For Amazon S3 destinations, streaming data is delivered to your S3 bucket. If data transformation is enabled, you can optionally back up source data to another Amazon S3 bucket.

For Amazon Redshift destinations, streaming data is delivered to your S3 bucket first. Kinesis Data Firehose then issues an Amazon Redshift **COPY** command to load data from your S3 bucket to your Amazon Redshift cluster. If data transformation is enabled, you can optionally back up source data to another Amazon S3 bucket.

For OpenSearch Service destinations, streaming data is delivered to your OpenSearch Service cluster, and it can optionally be backed up to your S3 bucket concurrently.

For Splunk destinations, streaming data is delivered to Splunk, and it can optionally be backed up to your S3 bucket concurrently.

Using a Studio notebook with Kinesis Data Analytics for Apache Flink

Studio notebooks for Kinesis Data Analytics allows you to interactively query data streams in real time, and easily build and run stream processing applications using standard SQL, Python, and Scala. With a few clicks in the AWS Management console, you can launch a serverless notebook to query data streams and get results in seconds.

A notebook is a web-based development environment. With notebooks, you get a simple interactive development experience combined with the advanced

capabilities provided by Apache Flink. Studio notebooks uses notebooks powered by Apache Zeppelin, and uses Apache Flink as the stream processing engine. Studio notebooks seamlessly combines these technologies to make advanced analytics on data streams accessible to developers of all skill sets.

Apache Zeppelin provides your Studio notebooks with a complete suite of analytics tools, including the following:

- Data Visualization
- Exporting data to files
- Controlling the output format for easier analysis

To get started using Kinesis Data Analytics and Apache Zeppelin, see Creating a Studio notebook Tutorial. For more information about Apache Zeppelin, see the Apache Zeppelin documentation.

With a notebook, you model queries using the Apache Flink Table API & SQL in SQL, Python, or Scala, or DataStream API in Scala. With a few clicks, you can then promote the Studio notebook to a continuously-running, non-interactive, Kinesis Data Analytics stream-processing application for your production workloads.

What Is Amazon Kinesis Data Analytics for SQL Applications?

For new projects, it is recommended to use the new Kinesis Data Analytics Studio over Kinesis Data Analytics for SQL Applications. Kinesis Data Analytics Studio combines ease of use with advanced analytical capabilities, enabling you to build sophisticated stream processing applications in minutes.

With Amazon Kinesis Data Analytics for SQL Applications, you can process and analyze streaming data using standard SQL. The service enables you to quickly author and run powerful SQL code against streaming sources to perform time series analytics, feed real-time dashboards, and create real-time metrics.

To get started with Kinesis Data Analytics, you create a Kinesis data analytics application that continuously reads and processes streaming data. The service supports ingesting data from Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose streaming sources. Then, you author your SQL code using the interactive editor and test it with live streaming data. You can also configure destinations where you want Kinesis Data Analytics to send the results.

Kinesis Data Analytics supports Amazon Kinesis Data Firehose (Amazon S3, Amazon Redshift, Amazon OpenSearch Service, and Splunk), AWS Lambda, and Amazon Kinesis Data Streams as destinations.

When Should I Use Amazon Kinesis Data Analytics?

Amazon Kinesis Data Analytics enables you to quickly author SQL code that continuously reads, processes, and stores data in near real time. Using standard SQL queries on the streaming data, you can construct applications that transform and provide insights into your data. Following are some of example scenarios for using Kinesis Data Analytics:

- **Generate time-series analytics** – You can calculate metrics over time windows, and then stream values to Amazon S3 or Amazon Redshift through a Kinesis data delivery stream.
- **Feed real-time dashboards** – You can send aggregated and processed streaming data results downstream to feed real-time dashboards.
- **Create real-time metrics** – You can create custom metrics and triggers for use in real-time monitoring, notifications, and alarms.

For information about the SQL language elements that are supported by Kinesis Data Analytics, see Amazon Kinesis Data Analytics SQL Reference.

Amazon Kinesis Video Streams FAQs

General

Q: What is Amazon Kinesis Video Streams?

Amazon Kinesis Video Streams makes it easy to securely stream media from connected devices to AWS for storage, analytics, machine learning (ML), playback, and other processing. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest streaming media from millions of devices. It durably stores, encrypts, and indexes media in your streams, and allows you to access your media through easy-to-use APIs. Kinesis Video Streams enables you to quickly build computer vision and ML applications through integration with Amazon Rekognition Video, Amazon SageMaker, and libraries for ML frameworks such as Apache MxNet, TensorFlow, and OpenCV. For live and on-demand playback, Kinesis Video Streams provides fully-managed capabilities for HTTP Live Streaming (HLS) and Dynamic Adaptive Streaming over HTTP (DASH). Kinesis Video Streams also supports ultra-low latency two-way media streaming with WebRTC, as a fully managed capability.

Q: What is time-encoded data?

Time-encoded data is any data in which the records are in a time series, and each record is related to its previous and next records. Video is an example of time-encoded data, where each frame is related to the previous and next frames through spatial transformations. Other

examples of time-encoded data include audio, RADAR, and LIDAR signals. Amazon Kinesis Video Streams is designed specifically for cost-effective, efficient ingestion, and storage of all kinds of time-encoded data for analytics and ML use cases.

Q: What are common use cases for Kinesis Video Streams?

Kinesis Video Streams is ideal for building media streaming applications for camera-enabled IoT devices and for building real-time computer vision-enabled ML applications that are becoming prevalent in a wide range of use cases such as the following:

Smart Home

With Kinesis Video Streams, you can easily stream video and audio from camera-equipped home devices such as baby monitors, webcams, and home surveillance systems to AWS. You can then use the streams to build a variety of smart home applications ranging from simple media playback to intelligent lighting, climate control systems, and security solutions.

Smart City

Many cities have installed large numbers of cameras at traffic lights, parking lots, shopping malls, and just about every public venue, capturing video 24/7. You can use Kinesis Video Streams to securely and cost-effectively ingest, store, playback, and analyze this massive volume of media data to help solve traffic problems, help prevent crime, dispatch emergency responders, and much more.

Industrial Automation

You can use Kinesis Video Streams to collect a variety of time-encoded data such as RADAR and LIDAR signals, temperature profiles, and depth data from industrial equipment. You can then analyze the data using your favorite machine learning framework including Apache MxNet, TensorFlow, and OpenCV for industrial automation use cases like predictive maintenance. For example, you can predict the lifetime of a gasket or valve and schedule part replacement in advance, reducing downtime and defects in a manufacturing line.

Q: What does Amazon Kinesis Video Streams manage on my behalf?

Amazon Kinesis Video Streams is a fully managed service for media ingestion, storage, and processing. It enables you to securely ingest, process, and store video at any scale for applications that power robots, smart cities, industrial automation, security monitoring, machine learning (ML), and more. Kinesis Video Streams also ingests other kinds of time-encoded data like audio, RADAR, and LIDAR signals. Kinesis Video Streams provides you SDKs to install on your devices to make it easy to securely stream media to AWS. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest media streams from millions of devices. It also durably stores, encrypts, and indexes the media streams and provides easy-to-use APIs so that applications can retrieve and process indexed media fragments based on tags and timestamps. Kinesis Video Streams provides a library to integrate ML frameworks such as Apache MxNet, TensorFlow, and OpenCV with video streams to build machine learning applications. Kinesis Video Streams is integrated with Amazon Rekognition Video, enabling you to build computer vision applications that detect objects, events, and people.

Key concepts

Q: What is a video stream?

A video stream is a resource that enables you to capture live video and other time-encoded data, optionally store it, and make the data available for consumption both in real time and on a batch or ad-hoc basis. When you choose to store data in the video stream, Kinesis Video Streams will encrypt the data, and generate a time-based index on the stored data. In a typical configuration, a Kinesis video stream has only one producer publishing data into it. The Kinesis video stream can have multiple consuming applications processing the contents of the video stream.

Q: What is a fragment?

A fragment is a self-contained sequence of media frames. The frames belonging to a fragment should have no dependency on any frames from other fragments. As fragments arrive, Kinesis Video Streams assigns a unique fragment number, in increasing order. It also stores producer-side and server-side time stamps for each fragment, as Kinesis Video Streams-specific metadata.

Q: What is a producer?

A producer is a general term used to refer to a device or source that puts data into a Kinesis video stream. A producer can be any video-generating device, such as a security camera, a body-worn camera, a smartphone camera, or a dashboard camera. A producer can also send non-video time-encoded data, such as audio feeds, images, or RADAR data. One producer can generate one or more video streams. For example, a video camera can push video data to one Kinesis video stream and audio data to another.

Q: What is a consumer?

Consumers are your custom applications that consume and process data in Kinesis video streams in real time, or after the data is durably stored and time-indexed when low latency processing is not required. You can create these consumer applications to run on Amazon EC2 instances. You can also use other Amazon AI services such as Amazon Rekognition, or third party video analytics providers to process your video streams.

Q: What is a chunk?

Upon receiving the data from a producer, Kinesis Video Streams stores incoming media data as chunks. Each chunk consists of the actual media fragment, a copy of media metadata sent by the producer, and the Kinesis Video Streams-specific metadata such as the fragment number, and server-side and producer-side timestamps. When a consumer requests media data through the GetMedia API operation, Kinesis Video Streams returns a stream of chunks, starting with the fragment number that you specify in the request.

Q: How do I think about latency in Amazon Kinesis Video Streams?

There are four key contributors to latency in an end-to-end media data flow.

- Time spent in the device's hardware media pipeline: This pipeline can comprise of the image sensor and any hardware encoders as appropriate. In theory, this can be as little as a single frame duration. In practice it rarely is. All encoders in order to work effectively for media encoding (compression) will accumulate several frames to construct a fragment. This process and any corresponding motion compensation algorithms will add anywhere from one second to several seconds of latency on the device before the data is packaged for transmission.
- Latency incurred on actual data transmission on the internet: The quality of the network throughput and latency can vary significantly based on where the producing device is located.

- Latency added by the Kinesis Video Streams as it receives data from the producer device: The incoming data is made available immediately on the GetMedia API operation for any consuming application. If you choose to retain data, then Kinesis Video Streams will ensure that the data is encrypted using AWS Key Management Service (AWS KMS) and generate a time-based index on the individual fragments in the video stream. When you access this retained data using the GetMediaforFragmentList API, Kinesis Video Streams fetches the fragments from durable storage, decrypt the data, and make it available for the consuming application.
- Time latency on data transmission back to the consumer: There can be consuming devices on the internet or other AWS regions that request the media data. The quality of the network throughput and latency can vary significantly based on where the consuming device is located.

Publishing data to streams

Q: How do I publish data to my Kinesis video stream?

You can publish media data to a Kinesis video stream via the PutMedia operation, or use the Kinesis Video Streams Producer SDKs in Java, C++, or Android. If you choose to use the PutMedia operation directly, you will be responsible for packaging the media stream according to the Kinesis Video Streams data specification, handle the stream creation, token rotation, and other actions necessary for reliable streaming of media data to the AWS cloud. We recommend using the Producer SDKs to make these tasks simpler and get started faster.

Q: What is the Kinesis Video Streams PutMedia operation?

Kinesis Video Streams provides a PutMedia API to write media data to a Kinesis video stream. In a PutMedia request, the producer sends a stream of media fragments. As fragments arrive, Kinesis Video Streams assigns a unique fragment number, in increasing order. It also stores producer-side and server-side time stamps for each fragment, as Kinesis Video Streams-specific metadata.

Q: What is the Kinesis Video Streams Producer SDK?

The Amazon Kinesis Video Streams Producer SDK are a set of easy-to-use and highly configurable libraries that you can install and customize for your specific producers. The SDK makes it easy to build an on-device application that securely connects to a video stream, and reliably publishes video and other media data to Kinesis Video Streams. It takes care of all the underlying tasks required to package the frames and fragments generated by the device's media pipeline. The SDK also handles stream creation, token rotation for secure and uninterrupted streaming, processing acknowledgements returned by Kinesis Video Streams, and other tasks.

Q: In which programming platforms is the Kinesis Video Streams Producer SDK available?

Kinesis Video Streams Producer SDK's core is built in C, so it is efficient and portable to a variety of hardware platforms. Most developers will prefer to use the C, C++ or Java versions of the Kinesis Video Streams producer SDK. There is also an Android version of the producer SDK for mobile app developers who want to stream video data from Android devices.

Q: What should I be aware of before getting started with the Kinesis Video Streams producer SDK?

The Kinesis Video Streams producer SDK does all the heavy lifting of packaging frames and fragments, establishes a secure connection, and reliably streams video to AWS. However there are many different varieties of hardware devices and media pipelines running on them. To make

the process of integration with the media pipeline easier, we recommend having some knowledge of: 1) the frame boundaries, 2) the type of a frame used for the boundaries, I-frame or non I-frame, and 3) the frame encoding time stamp.

Reading data from streams

Q: What is the GetMedia API?

You can use the GetMedia API to retrieve media content from a Kinesis video stream. In the request, you identify stream name or stream Amazon Resource Name (ARN), and the starting chunk. Kinesis Video Streams then returns a stream of chunks in order by fragment number. When you put media data (fragments) on a stream, Kinesis Video Streams stores each incoming fragment and related metadata in what is called a "chunk." The GetMedia API returns a stream of these chunks starting from the chunk that you specify in the request.

Q: What is the GetMediaForFragmentList API?

You can use the GetMediaForFragmentList API to retrieve media data for a list of fragments (specified by fragment number) from the archived data in a Kinesis video stream. Typically a call to this API operation is preceded by a call to the ListFragments API.

Q: What is the ListFragments API?

You can use the ListFragments API to return a list of Fragments from the specified video stream and start location - using the fragment number or timestamps - within the retained data.

Q: How long can I store data in Kinesis Video Streams?

You can store data in their streams for as long as you like. Kinesis Video Streams allows you to configure the data retention period to suit your archival and storage requirements.

Q: What is the Kinesis Video Streams parser library?

The Kinesis Video Streams parser library makes it easy for developers to consume and process the output of Kinesis Video Streams GetMedia operation. Application developers will include the library in their video analytics and processing applications that operate on video streams. The applications themselves will run on your EC2 instances, although they can be run elsewhere. The library has features that make it easy to get a frame-level object and its associated metadata, extract and collect Kinesis Video Streams-specific metadata attached to fragments, and consecutive fragments. You can then build custom applications that can more easily use the raw video data for your use cases.

Q: If I have a custom processing application that needs to use the frames (and fragments) carried by the Kinesis video stream, how do I do that?

In general, if you want to consume video streams and then manipulate them to fit your custom application's needs, then there are two key steps to consider. First, get the bytes in a frame from the formatted stream vended by the GetMedia API. You can use the stream parser library to get the frame objects. Next, get the metadata necessary to decode a frame such as the pixel height, width, codec id, and codec private data. Such metadata is embedded in the track elements. The parser library makes extracting this information easier by providing helper classes to collect the track information for a fragment.

The steps after this are highly application dependent. You may wish to decode frames, format them for a playback engine, transcode them for content distribution, or feed them into a custom

deep learning application format. The Kinesis Video Streams stream parser library is open-sourced so that you can extend it for your specific use cases.

Playing back video from streams

Q: How do I playback the video captured in my own application?

You can use Amazon Kinesis Video Streams' HTTP Live Streams (HLS) and Dynamic Adaptive Streaming over HTTP (DASH) capabilities to playback the ingested video in fragmented MP4 or MPEG_TS packaged format. HLS and DASH are industry-standard, HTTP-based media streaming protocols. As you capture video from devices using Amazon Kinesis Video Streams, you can use the HLS or DASH APIs to playback live or recorded video. This capability is fully managed, so you do not have to build any cloud-based infrastructure to support video playback. For low-latency playback and two-way media streaming, see the FAQs on WebRTC–based streaming.

Q: How do I get started with Kinesis Video Streams HLS or DASH APIs?

To view a Kinesis video stream using HLS or DASH, you first create a streaming session using GetHLSStreamingSessionURL or GetDASHStreamingSessionURL APIs. This action returns a URL (containing a session token) for accessing the HLS or DASH session, which you can then use in a media player or a standalone application to playback the stream. You can use a third-party player (such as Video.js or Google Shaka Player) to display the video stream, by providing the HLS or DASH streaming session URL, either programmatically or manually. You can also play back video by entering the HLS or DASH streaming session URL in the Location bar of the Apple Safari or Microsoft Edge browsers. Additionally, you can use the video players for Android (Exoplayer) and iOS (AVMediaPlayer) for mobile apps.

Q: What are the basic requirements to use the Kinesis Video Streams HLS APIs?

An Amazon Kinesis video stream has the following requirements for providing data through HLS:

- The media must contain h.264 or h.265 encoded video and, optionally, AAC encoded audio. Specifically, the codec ID of track 1 should be V_MPEG/ISO/AVC for h.264 or V_MPEG/ISO/HEVC for h.265. Optionally, the codec ID of track 2 should be A_AAC.
- The video track of each fragment must contain codec private data in the Advanced Video Coding (AVC) for h.264 format or HEVC for h.265 format (MPEG-4 specification ISO/IEC 14496-15). For information about adapting stream data to a given format, see NAL Adaptation Flags.
- Data retention must be greater than 0.
- The audio track (if present) of each fragment must contain codec private data in the AAC format (AAC specification ISO/IEC 13818-7)

Q: What are the basic requirements to use the Kinesis Video Streams DASH APIs?

An Amazon Kinesis video stream has the following requirements for providing data through DASH:

- The media must contain h.264 or h.265 encoded video and, optionally, AAC or G.711 encoded audio. Specifically, the codec ID of track 1 should be V_MPEG/ISO/AVC (for h.264) or V_MPEGH/ISO/HEVC (for H.265). Optionally, the codec ID of track 2 should be A_AAC (for AAC) or A_MS/ACM (for G.711).

- The video track of each fragment must contain codec private data in the Advanced Video Coding (AVC) for H.264 format and HEVC for H.265 format. For more information, see MPEG-4 specification ISO/IEC 14496-15. For information about adapting stream data to a given format, see NAL Adaptation Flags.
- Data retention must be greater than 0.
- The audio track (if present) of each fragment must contain codec private data in the AAC format (AAC specification ISO/IEC 13818-7) or the MS Wave format.

Q: What are the available playback modes for HLS or DASH streaming in Kinesis Video Streams?

There are two different playback modes supported by both HLS and DASH: Live and On Demand.

LIVE: For live sessions, the HLS media playlist is continually updated with the latest fragments as they become available. When this type of session is played in a media player, the user interface typically displays a "live" notification, with no scrubber control for choosing the position in the playback window to display.

ON DEMAND: For on-demand, the HLS media playlist contains all the fragments for the session, up to the number that is specified in MaxMediaPlaylistFragmentResults. The playlist can only be retrieved once for each session.
Additionally, HLS also supports playback in LIVE_REPLAY mode. In this mode, the HLS media playlist is updated similarly to how it is updated for LIVE mode except that it starts by including fragments from a given start time. This mode is useful for cases when you want to start playback from a point in the past from stored media and continue into live streaming.

Q: What is the delay in the playback of video using the API?

The latency for live playback is typically between 3 and 5 seconds, but this could vary. We strongly recommend running your own tests and proof-of-concepts to determine the target latencies. There are a variety of factors that impact latencies, including the use case, how the producer generates the video fragments, the size of the video fragment, the player tuning, and network conditions both streaming into AWS and out of AWS for playback. For low-latency playback, see the FAQs on WebRTC–based streaming.

Q: What are the relevant limits to using HLS or DASH?

A Kinesis video stream supports a maximum of ten active HLS or DASH streaming sessions. If a new session is created when the maximum number of sessions is already active, the oldest (earliest created) session is closed. The number of active GetMedia connections on a Kinesis video stream does not count against this limit, and the number of active HLS sessions does not count against the active GetMedia connection limit. See Kinesis Video Streams Limits for more details.

Q: What’s the difference between Kinesis Video Streams and AWS Elemental MediaLive?

AWS Elemental MediaLive is a broadcast-grade live video encoding service. It lets you create high-quality video streams for delivery to broadcast televisions and internet-connected multiscreen devices, like connected TVs, tablets, smart phones, and set-top boxes. The service functions independently or as part of AWS Media Services.

Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for real-time and batch-driven machine learning (ML), video playback, analytics, and

other processing. It enables customers to build machine-vision based applications that power smart homes, smart cities, industrial automation, security monitoring, and more.

Q: Am I charged to use this capability?

Kinesis Video Streams uses a simple pay as you go pricing. There are no upfront costs and you only pay for the resources you use. Kinesis Video Streams pricing is based on the data volume (GB) ingested, volume of data consumed (GB) including through the HLS or DASH APIs, and the data stored (GB-Month) across all the video streams in your account.

Low-latency two-way media streaming with WebRTC

Q: What is WebRTC and how does Kinesis Video Streams support this capability?

WebRTC is an open technology specification for enabling real-time communication (RTC) across browsers and mobile applications via simple APIs. It leverages peering techniques for real-time data exchange between connected peers and provides low media streaming latency required for human-to-human interaction. WebRTC specification includes a set of IETF protocols including Interactive Connectivity Establishment (ICE RFC5245), Traversal Using Relay around NAT (TURN RFC5766), and Session Traversal Utilities for NAT (STUN RFC5389) for establishing peer-to-peer connectivity, in addition to protocol specifications for real-time media and data streaming. Kinesis Video Streams provides a standards compliant WebRTC implementation, as a fully-managed capability. You can use this capability to securely live stream media or perform two-way audio or video interaction between any camera IoT device and WebRTC compliant mobile or web players. As a fully-managed capability, you do not have to build, operate, or scale any WebRTC related cloud infrastructure such as signaling or media relay servers to securely stream media across applications and devices.

Q: What does Amazon Kinesis Video Streams manage on my behalf to enable live media streaming with WebRTC?

Kinesis Video Streams provides managed end-points for WebRTC signaling that allows applications to securely connect with each other for peer-to-peer live media streaming. Next, it includes managed end-points for TURN that enables media relay via the cloud when applications cannot stream peer-to-peer media. It also includes managed end-points for STUN that enables applications to discover their public IP address when they are located behind a NAT or a firewall. Additionally, it provides easy to use SDKs to enable camera IoT devices with WebRTC capabilities. Finally, it provides client SDKs for Android, iOS, and for Web applications to integrate Kinesis Video Streams WebRTC signaling, TURN, and STUN capabilities with any WebRTC compliant mobile or web player.

Q: What can I build using Kinesis Video Streams WebRTC capability?

With Kinesis Video Streams WebRTC, you can easily build applications for live media streaming or real-time audio or video interactivity between camera IoT devices, web browsers, and mobile devices for usecases such as helping parents keep an eye on their baby's room, enable home-owners use a video doorbell to check who's at the door, allow owners of camera-enabled robot vacuums to remotely control the robot by viewing the live camera stream on a mobile phone, and much more.

Q: How do I get started with Kinesis Video Streams WebRTC capability?

You can get started by building and running the sample applications in the Kinesis Video Streams SDKs for WebRTC available for Web browsers, Android or iOS based mobile devices, and for Linux, Raspbian, and MacOS based IoT devices. You can also run a quick demo of this capability in the Kinesis Video Streams management console by creating a signaling channel,

and running the demo application to live stream audio and video from your laptop's built-in camera and microphone.

Q: What is a Signaling Channel?

A signaling channel is a resource that enables applications to discover, set up, control, and terminate a peer-to-peer connection by exchanging signaling messages. Signaling messages are metadata that two applications exchange with each other to establish peer-to-peer connectivity. This metadata includes local media information such as media codecs and codec parameters, and possible network candidate paths for the two applications to connect with each other for live streaming.

Q: How do applications use a signaling channel to enable peer-to-peer connectivity?

Streaming applications can maintain persistent connectivity with a signaling channel and wait for other applications to connect to them or they can connect to a signaling channel only when they need to live stream media. The signaling channel enables applications to connect with each other in a one to few model using the concept of one master connecting to multiple viewers. The application that initiates the connection assumes the responsibility of a master via the ConnectAsMaster API and wait for viewers. Upto 10 applications can then connect to that signaling channel by assuming the viewer responsibility via the ConnectAsViewer API. Once connected to the signaling channel, the master and viewer applications can send each other signaling messages to establish peer-t0-peer connectivity for live media streaming.

Q: How do applications live stream peer-to-peer media when they are located behind a NAT or a firewall?

Applications use Kinesis Video Streams STUN end point to discover their public IP address when they are located behind a NAT or a firewall. An application provides its public IP address as a possible location where it can receive connection requests from other applications for live streaming. The default option for all WebRTC communication is direct peer-to-peer connectivity but if the NAT or firewall does now allow direct connectivity (e.g. in case of symmetric NATs), applications can connect to the Kinesis Video Streams TURN end points for relaying media via the cloud. The GetIceServerConfig API provides the necessary TURN end point information that applications can use in their WebRTC configuration. This configuration allows applications to use TURN relay as a fallback when they are unable to establish a direct peer-to-peer connection for live streaming.

Q: How does Kinesis Video Streams secure the live media streaming with WebRTC?

End to end encryption is a mandatory feature of WebRTC, and Kinesis Video Streams enforces it on all the components, including signaling and media or data streaming. Regardless of whether the communication is peer-to-peer or relayed via Kinesis Video Streams TURN end points, all WebRTC communications are securely encrypted through standardized encryption protocols. The signaling messages are exchanged using secure Websockets (WSS), data streams are encrypted using Datagram Transport Layer Security (DTLS), and media streams are encrypted using Secure Real-time Transport Protocol (SRTP).

Console

Q: What is the Kinesis Video Streams management console?

The Kinesis Video Streams management console enables you to create, update, manage, and monitor your video streams. It console can also playback your media streams live or on an on-demand basis, as long as the content in the streams is in the supported media type. Using the player controls, you can view the live stream, skip forwards or backwards 10 seconds, use the

date and time picker to rewind to a point in the past when you have set the corresponding retention period for the video stream. The Kinesis Video Streams management console's video playback capabilities are offered as a quick diagnostic tool for development and test scenarios for developers as they build solutions using Kinesis Video Streams.

Q: What media type does the console support?

The only supported video media type for playback in the Kinesis Video Streams management console is the popular H.264 format. This media format has wide support on devices, hardware and software encoders and playback engines. While, you can ingest any variety of video, audio, or other custom time-encoded data types for your own consumer applications and use cases, the management console will not perform playback of those other data types.

Q: What is the delay in the playback of video on the Kinesis Video Streams management console?

For a producer that is transmitting video data into the video stream, you will experience a 2 - 10 second lag in the live playback experience in the Kinesis Video Streams management console. The majority of the latency is added by the producer device as it accumulates frames into fragments before it transmits data over the internet. Once the data enters into the Kinesis Video Streams endpoint and you request playback, the console will get H.264 media type fragments from the durable storage, trans-package the fragments into a media format suitable for playback across different internet browsers. The trans-packaged media content will then be transferred to your location where you requested the playback from over the internet.

Encryption

Q: What Is Server-Side Encryption for Kinesis Video Streams?

Server-side encryption is a feature in Kinesis Video Streams that automatically encrypts data before it's at rest by using an AWS KMS key that you specify. Data is encrypted before it is written to the Kinesis Video Streams storage layer, and it is decrypted after it is retrieved from storage. As a result, your data is always encrypted at rest within the Kinesis Video Streams service.

Q: How do I get started with server-side encryption?

Server-side encryption is always enabled on Kinesis video streams. If a user-provided key is not specified when the stream is created, the default key (provided by Kinesis Video Streams) is used.

A user-provided AWS KMS key must be assigned to a Kinesis Video Streams stream when it is created. You can't later assign a different key to a stream using the UpdateStream API.

You can assign a user-provided AWS KMS key to a Kinesis video stream in two ways: When creating a Kinesis video stream in the console, specify the AWS KMS key in the Encryption section on the Create new Kinesis Video stream page. Or when creating a Kinesis Video Streams stream using the CreateStream API, specify the key ID in the KmsKeyId parameter.

Q: How much does it cost to use server-side encryption?

When you apply server-side encryption, you are subject to AWS KMS API usage and key costs. Unlike custom AWS KMS keys, the (Default) aws/kinesis-video KMS key is offered free of charge. However, you still pay for the API usage costs that Kinesis Video Streams incurs on your behalf. API usage costs apply for every KMS key, including custom ones. Kinesis Video Streams calls AWS KMS approximately every 45 minutes when it is rotating the data key. In a

30-day month, the total cost of AWS KMS API calls that are initiated by a Kinesis Video Streams stream should be less than a few dollars. This cost scales with the number of user credentials that you use on your data producers and consumers because each user credential requires a unique API call to AWS KMS.

Amazon Kinesis Data Streams FAQs

General

Q: What is Amazon Kinesis Data Streams?

With Amazon Kinesis Data Streams, you can build custom applications that process or analyze streaming data for specialized needs. You can add various types of data such as clickstreams, application logs, and social media to a Kinesis data stream from hundreds of thousands of sources. Within seconds, the data will be available for your applications to read and process from the stream.

Q: What does Amazon Kinesis Data Streams manage on my behalf?

Amazon Kinesis Data Streams manages the infrastructure, storage, networking, and configuration needed to stream your data at the level of your data throughput. You don't have to worry about provisioning, deployment, or ongoing maintenance of hardware, software, or other services for your data streams. In addition, Kinesis Data Streams synchronously replicates data across three Availability Zones, providing high availability and data durability. By default, Kinesis Data Streams scales capacity automatically, freeing you from provisioning and managing capacity. You can choose provisioned mode if you want to provision and manage throughput on your own.

Q: What can I do with Amazon Kinesis Data Streams?

Kinesis Data Streams is useful for rapidly moving data off data producers and then continuously processing the data, whether that means transforming it before emitting to a data store, running real-time metrics and analytics, or deriving more complex data streams for further processing.

The following are typical scenarios for using Kinesis Data Streams:

Accelerated log and data feed intake: Instead of waiting to batch the data, you can have your data producers push data to a Kinesis data stream as soon as the data is produced, preventing data loss in case of producer failure. For example, system and application logs can be continuously added to a data stream and be available for processing within seconds.

Real-time metrics and reporting: You can extract metrics and generate reports from Kinesis data stream data in real time. For example, your Amazon Kinesis application can work on metrics and reporting for system and application logs as the data is streaming in, rather than waiting to receive data batches.

Real-time data analytics: With Kinesis Data Streams, you can run real-time streaming data analytics. For example, you can add clickstreams to your Kinesis data stream and have your Kinesis application run analytics in real time, allowing you to gain insights from your data in minutes instead of hours or days.

Log and event data collection: Collect log and event data from sources such as servers, desktops, and mobile devices. You can then build applications using Amazon Lambda or Kinesis Data Analytics to continuously process the data, generate metrics, power live

dashboards, and emit aggregated data into stores such as Amazon Simple Storage Service (S3).

Power event-driven applications: Quickly pair with AWS Lambda to respond or adjust to immediate occurrences within the event-driven applications in your environment, at any scale.

Q: How do I use Amazon Kinesis Data Streams?

After you sign up for AWS, you can start using Kinesis Data Streams by creating a Kinesis data stream through either the AWS Management Console or the CreateStream operation. Then configure your data producers to continuously add data to your data stream. You can optionally send data from existing resources in AWS services such as Amazon DynamoDB, Amazon Aurora, Amazon CloudWatch, and AWS IoT Core. You can then use AWS Lambda, Amazon Kinesis Data Analytics, or AWS Glue Streaming to quickly process data stored in Kinesis Data Streams. You can also build custom applications that run on Amazon Elastic Compute Cloud (EC2), Amazon Elastic Container Service (ECS), and Amazon Elastic Kubernetes Service (EKS) using either Amazon Kinesis API or Amazon Kinesis Client Library (KCL).

Key concepts

Q: What is a shard, producer, and consumer in Kinesis Data Streams?

A shard has a sequence of data records in a stream. It serves as a base throughput unit of a Kinesis data stream. A shard supports 1 MB/second and 1,000 records per second for writes and 2 MB/second for reads. The shard limits ensure predictable performance, making it easy to design and operate a highly reliable data streaming workflow. A producer puts data records into shards and a consumer gets data records from shards. Consumers use shards for parallel data processing and for consuming data in the exact order in which they are stored. If writes and reads exceed the shard limits, the producer and consumer applications will receive throttles, which can be handled through retries.

Q: What is a record?

A record is the unit of data stored in an Amazon Kinesis data stream. A record is composed of a sequence number, partition key, and data blob. Data blob is the data of interest your data producer adds to a data stream. The maximum size of a data blob (the data payload before Base64-encoding) is 1 megabyte (MB).

Q: What is a partition key?

A partition key is used to segregate and route records to different shards of a data stream. A partition key is specified by your data producer while adding data to a Kinesis data stream. For example, let's say you have a data stream with two shards (shard 1 and shard 2). You can configure your data producer to use two partition keys (key A and key B) so that all records with key A are added to shard 1 and all records with key B are added to shard 2.

Q: What is a sequence number?

A sequence number is a unique identifier for each record. Sequence number is assigned by Amazon Kinesis when a data producer calls PutRecord or PutRecords operation to add data to a Amazon Kinesis data stream. Sequence numbers for the same partition key generally increase over time; the longer the time period between PutRecord or PutRecords requests, the larger the sequence numbers become.

Q: What is a capacity mode?

The capacity mode of Kinesis Data Streams determines how capacity is managed and usage is charged for a data stream. You can choose between provisioned and on-demand modes. In provisioned mode, you specify the number of shards for the data stream. The total capacity of a data stream is the sum of the capacities of its shards. You can increase or decrease the number of shards in a data stream as needed, and you pay for the number of shards at an hourly rate. In on-demand mode, AWS manages the shards to provide the necessary throughput. You pay only for the actual throughput used, and Kinesis Data Streams automatically accommodates your workload throughput needs as they ramp up or down. All Kinesis Data Streams write and read APIs, along with optional features such as Extended Retention and Enhanced Fan-Out, are supported in both capacity modes.

Q: How do I choose between on-demand and provisioned mode?

On-demand mode is best suited for workloads with unpredictable and highly variable traffic patterns. You should use this mode if you prefer AWS to manage capacity on your behalf or prefer pay-per-throughput pricing. Provisioned mode is best suited for predictable traffic, where capacity requirements are easy to forecast. You should consider using provisioned mode if you want fine-grained control over how data is distributed across shards. Provisioned mode is also suitable if you want to provision additional shards so the consuming application can have more read throughput to speed up the overall processing.

Q: Can I switch between on-demand and provisioned mode?

Yes. You can switch between on-demand and provisioned mode twice a day. The shard count of your data stream remains the same when you switch from provisioned mode to on-demand mode and vice versa. With the switch from provisioned to on-demand capacity mode, your data stream retains whatever shard count it had before the transition. But from that point on, Kinesis Data Streams monitors your data traffic and scales the shard count of this on-demand data stream up or down depending on traffic increase or decrease.

Adding data to Kinesis Data Streams

Q: How do I add data to my Amazon Kinesis data stream?

You can add data to a Kinesis data stream through PutRecord and PutRecords operations, Amazon Kinesis Producer Library (KPL), or Amazon Kinesis Agent.

Q: What is the difference between PutRecord and PutRecords?

PutRecord operation allows a single data record within an API call, and PutRecords operation allows multiple data records within an API call. For more information, see PutRecord and PutRecords.

Q: What is Amazon Kinesis Producer Library (KPL)?

Amazon Kinesis Producer Library (KPL) is an easy-to-use and highly configurable library that helps you put data into an Amazon Kinesis data stream. KPL presents a simple, asynchronous, and reliable interface that enables you to quickly achieve high producer throughput with minimal client resources.

Q: What is Amazon Kinesis Agent?

Amazon Kinesis Agent is a prebuilt Java application that offers an easy way to collect and send data to your Amazon Kinesis data stream. You can install the agent on Linux-based server environments such as web servers, log servers, and database servers. The agent monitors

certain files and continuously sends data to your data stream. For more information, see Writing with Agents.

Q: What data is counted against the data throughput of an Amazon Kinesis data stream during a PutRecord or PutRecords call?

Your data blob, partition key, and data stream name are required parameters of a PutRecord or PutRecords call. The size of your data blob (before Base64 encoding) and partition key will be counted against the data throughput of your Amazon Kinesis data stream, which is determined by the number of shards within the data stream.

Reading and processing data from Kinesis Data Streams

Q: What is a consumer, and what are different consumer types offered by Amazon Kinesis Data Streams?

A consumer is an application that processes all data from a Kinesis data stream. You can choose between shared fan-out and enhanced fan-out consumer types to read data from a Kinesis data stream. The shared fan-out consumers all share a shard's 2 MB/second of read throughput and five transactions per second limits and require the use of the GetRecords API. An enhanced fan-out consumer gets its own 2 MB/second allotment of read throughput, allowing multiple consumers to read data from the same stream in parallel, without contending for read throughput with other consumers. You need to use the SubscribeToShard API with the enhanced fan-out consumers. We recommend using enhanced fan-out consumers if you want to add more than one consumer to your data stream.

Q: How I can process data captured and stored in Amazon Kinesis Data Streams?

You can use managed services such as AWS Lambda, Amazon Kinesis Data Analytics, and AWS Glue to process data stored in Kinesis Data Streams. These managed services take care of provisioning and managing the underlying infrastructure so you can focus on writing your business logic. You can also deliver data stored in Kinesis Data Streams to Amazon S3, Amazon OpenSearch Service, Amazon Redshift, and custom HTTP endpoints using its prebuilt integration with Kinesis Data Firehose. You can also build custom applications using Amazon Kinesis Client Library, a prebuilt library, or the Amazon Kinesis Data Streams API.

Q: What is Amazon Kinesis Client Library (KCL)?

Amazon Kinesis Client Library (KCL) for Java, Python, Ruby, Node.js, and .NET is a prebuilt library that helps you easily build Amazon Kinesis applications for reading and processing data from an Amazon Kinesis data stream.

KCL handles complex issues such as adapting to changes in data stream volume, load-balancing streaming data, coordinating distributed services, and processing data with fault tolerance. KCL enables you to focus on business logic while building applications. Refer to Kinesis Data Streams documentation here for more details on KCL.

Q: What is the SubscribeToShard API?

The SubscribeToShard API is a high-performance streaming API that pushes data from shards to consumers over a persistent connection without a request cycle from the client. The SubscribeToShard API uses the HTTP/2 protocol to deliver data to registered consumers whenever new data arrives on the shard, typically within 70 milliseconds, offering approximately 65% faster delivery compared to the GetRecords API. The consumers will enjoy fast delivery even when multiple registered consumers are reading from the same shard.

Q: What is enhanced fan-out?

Enhanced fan-out is an optional feature for Kinesis Data Streams consumers that provides logical 2 MB/second throughput pipes between consumers and shards. This allows you to scale the number of consumers reading from a data stream in parallel, while maintaining high performance.

Q: When should I use enhanced fan-out?

You should use enhanced fan-out if you have, or expect to have, multiple consumers retrieving data from a stream in parallel, or if you have at least one consumer that requires the use of the SubscribeToShard API to provide sub-200 millisecond data delivery speeds between producers and consumers.

Q: How is enhanced fan-out used by a consumer?

Consumers use enhanced fan-out by retrieving data with the SubscribeToShard API. The name of the registered consumer is used within the SubscribeToShard API, which leads to utilization of the enhanced fan-out benefit provided to the registered consumer.

Q: Can I have some consumers using enhanced fan-out, and other not?

Yes. You can have multiple consumers using enhanced fan-out and others not using enhanced fan-out at the same time. The use of enhanced fan-out does not impact the limits of shards for traditional GetRecords usage.

Q: Do I need to use enhanced fan-out if I want to use SubscribeToShard?

Yes. To use SubscribeToShard, you need to register your consumers, which activates enhanced fan-out. By default, your consumer will use enhanced fan-out automatically when data is retrieved through SubscribeToShard.

On-demand mode

Q: What are the default throughput quotas to write data into data stream using on-demand mode?

A new data stream created in on-demand mode has a quota of 4 MB/second and 4,000 records per second for writes. By default, these streams automatically scale up to 200 MB/second and 200,000 records per second for writes.

Q: How do data streams scale in on-demand mode to handle increase in write throughput?

A data stream in on-demand mode accommodates up to double its previous peak write throughput observed in the last 30 days. As your data stream's write throughput hits a new peak, Kinesis Data Streams scales the stream's capacity automatically. For example, if your data stream has a write throughput that varies between 10 MB/second and 40 MB/second, Kinesis Data Streams will ensure that you can easily burst to double the peak throughput of 80 MB/second. Subsequently, if the same data stream sustains a new peak throughput of 50 MB/second, Data Streams will ensure that there is enough capacity to ingest 100 MB/second of write throughput. However, you will see "ProvisionedThroughputExceeded" exceptions if your traffic grows more than double the previous peak within a 15-minute duration. You need to retry these throttled requests.

Q: What are the throughput limits for reading data from streams in on-demand mode?

On-demand mode's aggregate read capacity increases proportionally to write throughput to ensure that consuming applications always have adequate read throughput to process incoming data in real time. You get at least twice the write throughput to read data using the GetRecords API. We recommend using one consumer with the GetRecord API so it has enough room to catch up when the application needs to recover from downtime. To add more than one consuming application, you need to use enhanced fan-out, which supports adding up to 20 consumers to a data stream using the SubscribeToShard API, with each having dedicated throughput.

Provisioned mode

Q: What are the limits of Kinesis Data Streams in provisioned mode?

The throughput of a Kinesis data stream in provisioned mode is designed to scale without limits by increasing the number of shards within a data stream.

Q: How do I scale capacity of Kinesis Data Streams in provisioned mode?

You can scale up a Kinesis Data Stream capacity in provisioned mode by splitting existing shards using the SplitShard API. You can scale down capacity by merging two shards using the MergeShard API. Alternatively, you can use UpdateShardCount API to scale up (or down) a stream capacity to a specific shard count.

Q: How do I decide the throughput of my Amazon Kinesis data stream in provisioned mode?

The throughput of a Kinesis data stream is determined by the number of shards within the data stream. Follow the steps below to estimate the initial number of shards your data stream needs in provisioned mode. Note that you can dynamically adjust the number of shards within your data stream through resharding.

Estimate the average size of the record written to the data stream in kilobytes (KB), rounded up to the nearest 1 KB. (average_data_size_in_KB)

Estimate the number of records written to the data stream per second. (number_of_records_per_second)

Decide the number of Amazon Kinesis Applications consuming data concurrently and independently from the data stream. (number_of_consumers)

Calculate the incoming write bandwidth in KB (incoming_write_bandwidth_in_KB), which is equal to the average_data_size_in_KB multiplied by the number_of_records_per_second.

Calculate the outgoing read bandwidth in KB (outgoing_read_bandwidth_in_KB), which is equal to the incoming_write_bandwidth_in_KB multiplied by the number_of_consumers.

You can then calculate the initial number of shards (number_of_shards) your data stream needs using the following formula: number_of_shards = max (incoming_write_bandwidth_in_KB/1000, outgoing_read_bandwidth_in_KB/2000)

Q: What is the maximum throughput I can request for my Amazon Kinesis data stream in provisioned mode?

The throughput of a Kinesis data stream is designed to scale without limits. The default shard quota is 500 shards per stream for the following AWS Regions: US East (N. Virginia), US West (Oregon), and Europe (Ireland). For all other Regions, the default shard quota is 200 shards per stream. You can request the increase in the shard quota using the AWS Service Quotas console.

Q: What happens if the capacity limits of an Amazon Kinesis data stream are exceeded while the data producer adds data to the data stream in provisioned mode?

In provisioned mode, the capacity limits of a Kinesis data stream are defined by the number of shards within the data stream. The limits can be exceeded either by data throughput or by the number of PUT records. While the capacity limits are exceeded, the put data call will be rejected with a ProvisionedThroughputExceeded exception. If this is due to a temporary rise of the data stream's input data rate, retry by the data producer will eventually lead to completion of the requests. If it's due to a sustained rise of the data stream's input data rate, you should increase the number of shards within your data stream to provide enough capacity for the put data calls to consistently succeed. In both cases, Amazon CloudWatch metrics allow you to learn about the change of the data stream's input data rate and the occurrence of ProvisionedThroughputExceeded exceptions.

Q: What happens if the capacity limits of an Amazon Kinesis data stream are exceeded while the Amazon Kinesis application reads data from the data stream in provisioned mode?

In provisioned mode, the capacity limits of a Kinesis data stream are defined by the number of shards within the data stream. The limits can be exceeded either by data throughput or by the number of read data calls. While the capacity limits are exceeded, the read data call will be rejected with a ProvisionedThroughputExceeded exception. If this is due to a temporary rise of the data stream's output data rate, retry by the Amazon Kinesis application will eventually lead to completion of the requests. If it's due to a sustained rise of the data stream's output data rate, you should increase the number of shards within your data stream to provide enough capacity for the read data calls to consistently succeed. In both cases, Amazon CloudWatch metrics allow you to learn about the change of the data stream's output data rate and the occurrence of ProvisionedThroughputExceeded exceptions.

Extended and Long-term data retention

Q. What is the retention period supported by Kinesis Data Streams?

The default retention period of 24 hours covers scenarios where intermittent lags in processing require catch-up with the real-time data. A seven-day retention lets you reprocess data for up to seven days to resolve potential downstream data losses. Long term data retention greater than seven days and up to 365 days lets you reprocess old data for use cases such as algorithm back testing, data store backfills, and auditing.

Q. Can I use the existing Kinesis Data Streams APIs to read data older than seven days?

Yes. You can use the same getShardIterator, GetRecords, and SubscribeToShard APIs to read data retained for more than seven days. The consumers can move the iterator to the desired location in the stream, retrieve the shard map (including both open and closed), and read the records.

Q: Are there any new APIs to further assist in reading old data?

Yes. There are API enhancements to ListShards, GetRecords, and SubscribeToShard APIs. You can use the new filtering option with the TimeStamp parameter available in the ListShards API to efficiently retrieve the shard map and improve the performance of reading old data. The TimeStamp filter lets applications discover and enumerate shards from the point in time you wish to reprocess data and eliminate the need to start at the trim horizon. GetRecords and SubscribeToShards have a new field, ChildShards, which allows you to quickly discover the children shards when an application finishes reading data from a closed shard, instead of having to traverse the shard map again. The fast discovery of shards makes efficient use of the

consuming application's compute resources for any sized stream, irrespective of the data retention period.

Q: When do I use the API enhancements?

You should consider the API enhancements if you plan to retain data longer and scale your stream's capacity regularly. Stream scaling operations close existing shards and open new child shards. The data in all the open and closed shards is retained until the end of the retention period. So the total number of shards increase linearly with a longer retention period and multiple scaling operations. This increase in the shard map requires you to use ListShards with the TimeStamp filter and ChildShards field in GetRecords, and SubscribeToShard API for efficient discovery of shards for data retrieval. You will need to upgrade your KCL to the latest version (1.x for standard consumers and 2.x for enhanced fan-out consumers) for these features.

Q: Does Amazon Kinesis Data Streams support schema registration?

Yes. Clients of Kinesis Data Streams can use the AWS Glue Schema Registry, a serverless feature of AWS Glue, either through the Kinesis Producer Library (KPL) and Kinesis Client Libraries (KCL) or through AWS Glue Schema Registry APIs in the AWS Java SDK. The Schema Registry is available at no additional charge.

Visit the Schema Registry user documentation to get started and to learn more.

Managing Kinesis Data Streams

Q: How do I change the throughput of my Amazon Kinesis data stream in provisioned mode?

There are two ways to change the throughput of your data stream. You can use the UpdateShardCount API or the AWS Management Console to scale the number of shards in a data stream, or you can change the throughput of an Amazon Kinesis data stream by adjusting the number of shards within the data stream (resharding).

Q: How long does it take to change the throughput of my Amazon Kinesis data stream running in provisioned mode using UpdateShardCount or the AWS Management Console?

Typical scaling requests should take a few minutes to complete. Larger scaling requests will take longer than smaller ones.

Q: Does Amazon Kinesis Data Streams remain available when I change the throughput of my Kinesis data stream in provisioned mode or when the scaling happens automatically in on-demand mode?

Yes. You can continue adding data to and reading data from your Kinesis data stream while you use UpdateShardCount or reshard to change the throughput of the data stream or when Kinesis Data Streams does it automatically in on-demand mode.

Q: How do I monitor the operations and performance of my Amazon Kinesis data stream?

Amazon Kinesis Data Streams Management Console displays key operational and performance metrics such as throughput of data input and output of your Kinesis data streams. Kinesis Data Streams also integrates with Amazon CloudWatch so you can collect, view, and analyze CloudWatch metrics for your data streams and shards within those data streams. For more information about Amazon Kinesis Data Streams metrics, see Monitoring Amazon Kinesis Data Streams with Amazon CloudWatch.

Note that all stream-level metrics are free of charge. All enabled shard-level metrics are charged at Amazon CloudWatch Pricing.

Q: How do I manage and control access to my Amazon Kinesis data stream?

Amazon Kinesis Data Streams integrates with AWS Identity and Access Management (IAM), a service that enables you to securely control access to your AWS services and resources for your users. For example, you can create a policy that allows only a specific user or group to add data to your Kinesis data stream. For more information about access management and control of your data stream, see Controlling Access to Amazon Kinesis Data Streams Resources using IAM.

Q: How do I log API calls made to my Amazon Kinesis data stream for security analysis and operational troubleshooting?

Kinesis Data Streams integrates with Amazon CloudTrail, a service that records AWS API calls for your account and delivers log files to you. For more information about API call logging and a list of supported Amazon Kinesis API operations, see Logging Amazon Kinesis API calls Using Amazon CloudTrail.

Q: How do I effectively manage my Amazon Kinesis data streams and the costs associated with them?

Kinesis Data Streams allows you to tag your Kinesis data streams for easier resource and cost management. A tag is a user-defined label expressed as a key-value pair that helps organize AWS resources. For example, you can tag your data streams by cost centers so you can categorize and track your Kinesis Data Streams costs based on cost centers. For more information about Amazon Kinesis Data Streams tagging, see Tagging Your Amazon Kinesis Data Streams.

Security

Q: When I use Kinesis Data Streams, how secure is my data?

Amazon Kinesis is secure by default. Only the account and data stream owners have access to the Kinesis resources they create. Kinesis supports user authentication to control access to data. You can use AWS IAM policies to selectively grant permissions to users and groups of users. You can securely put and get your data from Kinesis through SSL endpoints using the HTTPS protocol. If you need extra security, you can use server-side encryption with AWS Key Management Service (KMS) keys to encrypt data stored in your data stream. AWS KMS allows you to use AWS-generated KMS keys for encryption, or if you prefer, you can bring your own KMS key into AWS KMS. Lastly, you can use your own encryption libraries to encrypt data on the client side before putting the data into Kinesis.

Q: Can I privately access Kinesis Data Streams APIs from my Amazon Virtual Private Cloud (VPC) without using public IPs?

Yes. You can privately access Kinesis Data Streams APIs from your Amazon VPC by creating VPC Endpoints. With VPC Endpoints, the routing between the VPC and Kinesis Data Streams is handled by the AWS network without the need for an internet gateway, NAT gateway, or VPN connection. The latest generation of VPC Endpoints used by Kinesis Data Streams are powered by AWS PrivateLink, a technology that enables private connectivity between AWS services using Elastic Network Interfaces (ENI) with private IPs in your VPCs. To learn more about PrivateLink, visit the PrivateLink documentation.

Encryption

Q: Can I encrypt the data I put into a Kinesis data stream?

Yes, and there are two options for doing so. You can use server-side encryption, which is a fully managed feature that automatically encrypts and decrypts data as you put and get it from a data stream. You can also write encrypted data to a data stream by encrypting and decrypting on the client side.

Q: Why should I use server-side encryption instead of client-side encryption?

You might choose server-side encryption over client-side encryption for any of the following reason:

1. It is hard to enforce client-side encryption.
2. They want a second layer of security on top of client-side encryption.
3. It is hard to implement client-side key management schemes.

Q: What is server-side encryption?

Server-side encryption for Kinesis Data Streams automatically encrypts data using a user specified AWS KMS key before it is written to the data stream storage layer, and decrypts the data after it is retrieved from storage. Encryption makes writes impossible and the payload and the partition key unreadable unless the user writing or reading from the data stream has the permission to use the key selected for encryption on the data stream. As a result, server-side encryption can make it easier to meet internal security and compliance requirements governing your data.

With server-side encryption your client-side applications (producers and consumers) do not need to be aware of encryption, they do not need to manage KMS keys or cryptographic operations, and your data is encrypted when it is at rest and in motion within the Kinesis Data Streams service. All KMS keys used by the server-side encryption feature are provided by the AWS KMS. AWS KMS makes it easy to use an AWS-managed KMS key for Kinesis (a “one-click” encryption method), your own AWS KMS customer-managed key, or a KMS key that you imported for encryption.

Q: Is there a server-side encryption getting started guide?

Yes, there is a getting started guide in the user documentation.

Q: Does server-side encryption interfere with how my applications interact with Kinesis Data Streams?

Possibly. It depends on the key you use for encryption and the permissions governing access to the key.

- If you use the AWS-managed KMS key for Kinesis (key alias = aws/kinesis) your applications will not be impacted by enabling or disabling encryption with this key.
- If you use a different KMS key, like a custom AWS KMS key or one you imported into the AWS KMS service, and if your producers and consumers of a data stream do not have permission to use the KMS key used for encryption, then your PUT and GET requests will fail. Before you can use server-side encryption you must configure AWS KMS key policies to allow encryption and decryption of messages. For examples and more information about AWS KMS permissions, see AWS KMS API Permissions:

Actions and Resources Reference in the AWS Key Management Service Developer Guide or the permissions guidelines in the Kinesis Data Streams server-side encryption user documentation.

Q: Is there an additional cost associated with the use of server-side encryption?

Yes, however if you are using the AWS-managed KMS key for Kinesis and are not exceeding the AWS Free Tier KMS API usage costs, your use of server-side encryption is free. The following describes the costs by resource:

Keys:

The AWS-managed KMS key for Kinesis (alias = “aws/kinesis”) is free.
Customer managed KMS keys are subject to KMS key costs. Learn more.

KMS API Usage:

API usage costs apply for every KMS key, including custom ones. Kinesis Data Streams calls KMS approximately every five minutes when it’s rotating the data key. In a 30-day month, the total cost of KMS API calls initiated by a Kinesis data stream should be less than a few dollars. Note that this cost scales with the number of user credentials you use on your data producers and consumers because each user credential requires a unique API call to AWS KMS. When you use IAM role for authentication, each assume role-call will result in unique user credentials, and you might want to cache user credentials returned by the assume-role-call to save KMS costs.

Q: Which AWS regions offer server-side encryption for Kinesis Data Streams?

Kinesis Data Streams server-side encryption is available in the AWS GovCloud Region and all public Regions except the China (Beijing) region.

Q: How do I start, update, or remove server-side encryption from a data stream?

All of these operations can be completed using the AWS Management Console or the AWS SDK. To learn more, see the Kinesis Data Streams server-side encryption getting started guide.

Q: What encryption algorithm is used for server-side encryption?

Kinesis Data Streams uses an AES-GCM 256 algorithm for encryption.

Q: If I encrypt a data stream that already has data written to it, either in plain text or ciphertext, will all of the data in the data stream be encrypted or decrypted if I update encryption?

No. Only new data written into the data stream will be encrypted (or left decrypted) by the new application of encryption.

Q: What does server-side encryption for Kinesis Data Streams encrypt?

Server-side encryption encrypts the payload of the message along with the partition key, which is specified by the data stream producer applications.

Q: Is server-side encryption a shard specific feature or a stream specific feature?

Server-side encryption is a stream specific feature.

Q: Can I change the KMS key that is used to encrypt a specific data stream?

Yes, using the AWS Management Console or the AWS SDK, you can choose a new KMS key to apply to a specific data stream.

Q: Is Amazon Kinesis Data Streams available in the AWS Free Tier?

No. Amazon Kinesis Data Streams is not currently available in the AWS Free Tier. AWS Free Tier is a program that offers free trial for a group of AWS services. For more details about AWS Free Tier, see AWS Free Tier.

Comparison with other AWS services

Q: How does Amazon Kinesis Data Streams differ from Amazon SQS?

Amazon Kinesis Data Streams enables real-time processing of streaming big data. It provides ordering of records, as well as the ability to read and/or replay records in the same order to multiple Amazon Kinesis Applications. The Amazon Kinesis Client Library (KCL) delivers all records for a given partition key to the same record processor, making it easier to build multiple applications reading from the same Kinesis data stream (for example, to perform counting, aggregation, and filtering). Amazon Simple Queue Service (SQS) offers a reliable, highly scalable hosted queue for storing messages as they travel between computers. Amazon SQS lets you easily move data between distributed application components and helps you build applications in which messages are processed independently (with message-level ack/fail semantics), such as automated workflows.

Q: When should I use Amazon Kinesis Data Streams, and when should I use Amazon SQS?

We recommend Kinesis Data Streams for use cases with requirements that are similar to the following:

Routing related records to the same record processor (as in streaming MapReduce). For example, counting and aggregation are simpler when all records for a given key are routed to the same record processor.

Ordering of records. For example, you want to transfer log data from the application host to the processing/archival host while maintaining the order of log statements.

Ability for multiple applications to consume the same stream concurrently. For example, you have one application that updates a real-time dashboard and another that archives data to Amazon Redshift. You want both applications to consume data from the same stream concurrently and independently.

Ability to consume records in the same order a few hours later. For example, you have a billing application and an audit application that runs a few hours behind the billing application. Because Kinesis Data Streams stores data for up to 365 days, you can run the audit application up to 365 days behind the billing application.

We recommend Amazon SQS for use cases with requirements that are similar to the following:

Messaging semantics (such as message-level ack/fail) and visibility timeout. For example, you have a queue of work items and want to track the successful completion of each item independently. Amazon SQS tracks the ack/fail so the application doesn’t have to maintain a persistent checkpoint/cursor. Amazon SQS will delete acked messages and redeliver failed messages after a configured visibility timeout.

Individual message delay. For example, you have a job queue and need to schedule individual jobs with a delay. With Amazon SQS, you can configure individual messages to have a delay of up to 15 minutes.

Dynamically increasing concurrency/throughput at read time. For example, you have a work queue and want to add more readers until the backlog is cleared. With Kinesis Data Streams, you can scale up to a sufficient number of shards (note, however, that you'll need to provision enough shards ahead of time).

Using the ability of Amazon SQS to scale transparently. For example, you buffer requests and the load changes as a result of occasional load spikes or the natural growth of your business. Because each buffered request can be processed independently, Amazon SQS can scale transparently to handle the load without any provisioning instructions from you.

Amazon Kinesis Data Firehose FAQs

General and Streaming ETL Concepts

Q: What is Streaming ETL?

Streaming ETL is the processing and movement of real-time data from one place to another. ETL is short for the database functions extract, transform, and load. Extract refers to collecting data from some source. Transform refers to any processes performed on that data. Load refers to sending the processed data to a destination, such as a warehouse, a datalake, or an analytical tool.

Q: What is Amazon Kinesis Data Firehose?

Kinesis Data Firehose is a streaming ETL solution. It is the easiest way to load streaming data into data stores and analytics tools. It can capture, transform, and load streaming data into Amazon S3, Amazon Redshift, Amazon OpenSearch Service, and Splunk, enabling near real-time analytics with existing business intelligence tools and dashboards you're already using today. It is a fully managed service that automatically scales to match the throughput of your data and requires no ongoing administration. It can also batch, compress, and encrypt the data before loading it, minimizing the amount of storage used at the destination and increasing security.

Q: What is a source in Kinesis Data Firehose?

A source is where your streaming data is continuously generated and captured. For example, a source can be a logging server running on Amazon EC2 instances, an application running on mobile devices, or a sensor on an IoT device. You can connect your sources to Kinesis Data Firehose using 1) Amazon Kinesis Data Firehose API, which uses the AWS SDK for Java, .NET, Node.js, Python, or Ruby. 2) Kinesis Data Stream, where Kinesis Data Firehose reads data easily from an existing Kinesis data stream and load it into Kinesis Data Firehose destinations. 3) AWS natively supported Service like AWS Cloudwatch, AWS EventBridge, AWS IOT, or AWS Pinpoint. For complete list, see the Amazon Kinesis Data Firehose developer guide. 4) Kinesis Agents, which is a stand-alone Java software application that continuously monitors a set of files and sends new data to your stream. 5) Fluentbit, which an open source Log Processor and Forwarder. 6) AWS Lambda, which is a serverless compute service that lets you run code without provisioning or managing servers. You can use write your Lambda function to send traffic from S3 or DynamoDB to Kinesis Data Firehose based on a triggered event.

Q: What is a destination in Kinesis Data Firehose?

A destination is the data store where your data will be delivered. Kinesis Data Firehose currently supports Amazon S3, Amazon Redshift, Amazon OpenSearch Service, Splunk, Datadog, NewRelic, Dynatrace, Sumologic, LogicMonitor, MongoDB, and HTTP End Point as destinations.

Q: What does Kinesis Data Firehose manage on my behalf?

Kinesis Data Firehose manages all underlying infrastructure, storage, networking, and configuration needed to capture and load your data into Amazon S3, Amazon Redshift, Amazon OpenSearch Service, or Splunk. You do not have to worry about provisioning, deployment, ongoing maintenance of the hardware, software, or write any other application to manage this process. Kinesis Data Firehose also scales elastically without requiring any intervention or associated developer overhead. Moreover, Kinesis Data Firehose synchronously replicates data across three facilities in an AWS Region, providing high availability and durability for the data as it is transported to the destinations.

Q: How do I use Kinesis Data Firehose?

After you sign up for Amazon Web Services, you can start using Kinesis Data Firehose with the following steps:

- Create an Kinesis Data Firehose delivery stream through the Firehose Console or the CreateDeliveryStream operation. You can optionally configure an AWS Lambda function in your delivery stream to prepare and transform the raw data before loading the data.
- Configure your data producers to continuously send data to your delivery stream using the Amazon Kinesis Agent or the Firehose API.
- Firehose automatically and continuously loads your data to the destinations you specify.

Q: What is a delivery stream in Kinesis Data Firehose?

A delivery stream is the underlying entity of Kinesis Data Firehose. You use Firehose by creating a delivery stream and then sending data to it. You can create an Kinesis Data Firehose delivery stream through the Firehose Console or the CreateDeliveryStream operation. For more information, see Creating a Delivery Stream.

Q: What is a record in Kinesis Data Firehose?

A record is the data of interest your data producer sends to a delivery stream. The maximum size of a record (before Base64-encoding) is 1024 KB.

Q: What are the limits of Kinesis Data Firehose?

For information about limits, see Amazon Kinesis Data Firehose Limits in the developer guide.

Data Sources

Q: What programming languages or platforms can I use to access Kinesis Data Firehose API?

Kinesis Data Firehose API is available in Amazon Web Services SDKs. For a list of programming languages or platforms for Amazon Web Services SDKs, see Tools for Amazon Web Services.

Q: What is Amazon Kinesis Agent?

Kinesis Agent is a pre-built Java application that offers an easy way to collect and send data to your delivery stream. You can install the agent on Linux-based server environments such as web servers, log servers, and database servers. The agent monitors certain files and continuously sends data to your delivery stream. Amazon Kinesis Agent currently supports Amazon Linux, Red Hat Enterprise Linux, and Microsoft Windows. For more information, see Writing with Agents.

Q: Where do I get Amazon Kinesis Agent?

You can download and install Kinesis Agent using the following command and link:

- On Amazon Linux: sudo yum install –y aws-kinesis-agent
- On Red Hat Enterprise Linux: sudo yum install –y https://s3.amazonaws.com/streaming-data-agent/aws-kinesis-agent-latest.amzn1.noarch.rpm
- From GitHub: awlabs/amazon-kinesis-agent
- On Windows: https://docs.aws.amazon.com/kinesis-agent-windows/latest/userguide/getting-started.html#getting-started-installation

Q: What is the difference between PutRecord and PutRecordBatch operations?

You can add data to an Kinesis Data Firehose delivery stream through Kinesis Agent or Firehose's PutRecord and PutRecordBatch operations. PutRecord operation allows a single data record within an API call and PutRecordBatch operation allows multiple data records within an API call. For more information, see PutRecord and PutRecordBatch.

Q: How do I add data to my Kinesis Data Firehose delivery stream from my Kinesis Data Stream?

When you create or update your delivery stream through AWS console or Firehose APIs, you can configure a Kinesis Data Stream as the source of your delivery stream. Once configured, Firehose will automatically read data from your Kinesis Data Stream and load the data to specified destinations.

Q: How often does Kinesis Data Firehose read data from my Kinesis stream?
Kinesis Data Firehose calls Kinesis Data Streams GetRecords() once every second for each Kinesis shard.

Q: From where does Kinesis Data Firehose read data when my Kinesis Data Stream is configured as the source of my delivery stream?

Kinesis Data Firehose starts reading data from the LATEST position of your Kinesis Data Stream when it's configured as the source of a delivery stream. For more information about Kinesis Data Stream position, see GetShardIterator in the Kinesis Data Streams Service API Reference.

Q: Can I configure my Kinesis Data Stream to be the source of multiple Kinesis Data Firehose delivery streams?

Yes, you can. However, note that the GetRecords() call from Kinesis Data Firehose is counted against the overall throttling limit of your Kinesis shard so that you need to plan your delivery stream along with your other Kinesis applications to make sure you won't get throttled. For more information, see Kinesis Data Streams Limits in the Kinesis Data Streams developer guide.

Q: Can I still add data to delivery stream through Kinesis Agent or Firehose's PutRecord and PutRecordBatch operations when my Kinesis Data Stream is configured as source?

No, you cannot. When a Kinesis Data Stream is configured as the source of a Kinesis Data Firehose delivery stream, Firehose's PutRecord and PutRecordBatch operations will be disabled. You should add data to your Kinesis Data Stream through the Kinesis Data Streams PutRecord and PutRecords operations instead.

Q: How do I add data to my delivery stream from AWS IoT?

You add data to your delivery stream from AWS IoT by creating an AWS IoT action that sends events to your delivery stream. For more information. See Writing to Amazon Kinesis Data Firehose Using AWS IoT in the Kinesis Data Firehose developer guide.

Q: How can I stream my VPC flow logs to Firehose?

When you create or update your delivery stream through AWS console or Firehose APIs, you can configure Direct PUT as the source of your delivery stream. Once the stream is created, you can configure the created Firehose stream as your delivery stream in the Vended Logs section of the VPC Flow logs console.

Q: How do I add data to my delivery stream from CloudWatch Logs?

You add data to your Kinesis Data Firehose delivery stream from CloudWatch Logs by creating a CloudWatch Logs subscription filter that sends events to your delivery stream. For more information, see Using CloudWatch Logs Subscription Filters in Amazon CloudWatch user guide.

Q: How do I add data to my Kinesis Data Firehose delivery stream from CloudWatch Events?

You add data to your Kinesis Data Firehose delivery stream from CloudWatch Events by creating a CloudWatch Events rule with your delivery stream as target. For more information, see Writing to Amazon Kinesis Data Firehose Using CloudWatch Events in the Kinesis Data Firehose developer guide.

Q: How do I add data to my Amazon Kinesis Data Firehose delivery stream from AWS Eventbridge?

You add data to your Kinesis Data Firehose delivery stream from AWS EventBridge console. For more information, see AWS EventBridge documentation.

Q: What kind of encryption can I use?

Kinesis Data Firehose allows you to encrypt your data after it's delivered to your Amazon S3 bucket. While creating your delivery stream, you can choose to encrypt your data with an AWS Key Management Service (KMS) key that you own. For more information about KMS, see AWS Key Management Service.

Q: What is the IAM role that I need to specify while creating a delivery stream?

Kinesis Data Firehose assumes the IAM role you specify to access resources such as your Amazon S3 bucket and Amazon OpenSearch domain. For more information, see Controlling Access with Kinesis Data Firehose in the Kinesis Data Firehose developer guide.

Data Transformation and Format Conversion

Q: How do I prepare and transform raw data in Kinesis Data Firehose?

Kinesis Data Firehose supports built-in data format conversion from data raw or Json into formats like Apache Parquet and Apache ORC required by your destination data stores, without

having to build your own data processing pipelines. Kinesis Data Firehose also allows you to dynamically partition your streaming data before delivery to S3 using static or dynamically defined keys like “customer_id” or “transaction_id”. Kinesis Data Firehose groups data by these keys and delivers into key-unique S3 prefixes, making it easier for you to perform high performance, cost efficient analytics in S3 using Athena, EMR, and Redshift Spectrum.

In addition to the built-in format conversion option in Amazon Kinesis Data Firehose, you can also use an AWS Lambda function to prepare and transform incoming raw data in your delivery stream before loading it to destinations. You can configure an AWS Lambda function for data transformation when you create a new delivery stream or when you edit an existing delivery stream. Amazon has created multiple Lambda Blue prints that you can choose from for quick start. For complete list, see the Amazon Kinesis Data Firehose developer guide.

Q: What compression format can I use?

Amazon Kinesis Data Firehose allows you to compress your data before delivering it to Amazon S3. The service currently supports GZIP, ZIP, and SNAPPY compression formats. Only GZIP is supported if the data is further loaded to Amazon Redshift.

Q: How does compression work when I use the CloudWatch Logs subscription feature?

You can use CloudWatch Logs subscription feature to stream data from CloudWatch Logs to Kinesis Data Firehose. All log events from CloudWatch Logs are already compressed in gzip format, so you should keep Firehose’s compression configuration as uncompressed to avoid double-compression. For more information about CloudWatch Logs subscription feature, see Subscription Filters with Amazon Kinesis Data Firehose in the Amazon CloudWatch Logs user guide.

Q: How do I return prepared and transformed data from my AWS Lambda function back to Amazon Kinesis Data Firehose?

All transformed records from Lambda must be returned to Firehose with the following three parameters; otherwise, Firehose will reject the records and treat them as data transformation failure.

- recordId: Firehose passes a recordId along with each record to Lambda during the invocation. Each transformed record should be returned with the exact same recordId. Any mismatch between the original recordId and returned recordId will be treated as data transformation failure.
- result: The status of transformation result of each record. The following values are allowed for this parameter: “Ok” if the record is transformed successfully as expected. “Dropped” if your processing logic intentionally drops the record as expected. “ProcessingFailed” if the record is not able to be transformed as expected. Firehose treats returned records with “Ok” and “Dropped” statuses as successfully processed records, and the ones with “ProcessingFailed” status as unsuccessfully processed records when it generates SucceedProcessing.Records and SucceedProcessing.Bytes metrics.
- data: The transformed data payload after based64 encoding.

Q: What is error logging?

If you enable data transformation with Lambda, Firehose can log any Lambda invocation and data delivery errors to Amazon CloudWatch Logs so that you can view the specific error logs if Lambda invocation or data delivery fails. For more information, see Monitoring with Amazon CloudWatch Logs.

Q: What is source record backup?

If you use data transformation with Lambda, you can enable source record backup, and Amazon Kinesis Data Firehose will deliver the un-transformed incoming data to a separate S3 bucket. You can specify an extra prefix to be added in front of the “YYYY/MM/DD/HH” UTC time prefix generated by Firehose.

Built-in Data Transformation for Amazon S3

Q: When should I use Kinesis Data Firehose dynamic partitioning?

Kinesis Data Firehose dynamic partitioning eliminates the complexities and delays of manual partitioning at the source or after storing the data, and enables faster analytics for querying optimized data sets. This makes the data sets immediately available for analytics tools to run their queries efficiently and enhances fine-grained access control for data. For example, marketing automation customers can partition data on-the-fly by customer id, allowing customer-specific queries to query optimized data sets and deliver results faster. IT operations or security monitoring customers can create groupings based on event timestamp embedded in logs so they can query optimized data sets and get results faster. This feature combined with Amazon Kinesis Data Firehose's existing JSON-to-parquet format conversion feature makes Amazon Kinesis Data Firehose an ideal streaming ETL option for S3.

Q: How do I setup dynamic partitioning with Kinesis Data Firehose?

You can setup Kinesis Data Firehose data partitioning capability through the AWS Management Console, CLIs or SDKs. When you create or update a Kinesis Data Firehose delivery stream, select Amazon S3 as the delivery destination for the delivery stream and enable dynamic partitioning. You can specify keys or create an expression that will be evaluated at runtime to define keys used for partitioning. For example, you can select a data field in the incoming stream such as customer id and define an S3 prefix expression such as customer_id=!{partitionKey:customer_id}/, that will be evaluated in runtime based on the ingested records to define to which S3 prefix deliver the records.

Q: What kind of transformations and data processing can I do with dynamic partitioning and with partitioning keys?

Kinesis Data Firehose supports parquet/orc conversion out of the box when you write your data to Amazon S3. Kinesis Data Firehose also integrates with Lambda function, so you can write your own transformation code. Kinesis Data Firehose also has built-in support for extracting the key data fields from records that are in JSON format. Kinesis Data Firehose also supports the JQ parsing language to enable transformations on those partition keys. To learn more, see the Kinesis Data Firehose developer guide.

Data Delivery and Destinations

Q: Can I keep a copy of all the raw data in my S3 bucket?

Yes, Kinesis Data Firehose can back up all un-transformed records to your S3 bucket concurrently while delivering transformed records to destination. Source record backup can be enabled when you create or update your delivery stream.

Q: How often does Kinesis Data Firehose deliver data to my Amazon S3 bucket?

The frequency of data delivery to Amazon S3 is determined by the S3 buffer size and buffer interval value you configured for your delivery stream. Kinesis Data Firehose buffers incoming

data before delivering it to Amazon S3. You can configure the values for S3 buffer size (1 MB to 128 MB) or buffer interval (60 to 900 seconds), and the condition satisfied first triggers data delivery to Amazon S3. If you have Apache parquet or dynamic partitioning enabled, then your buffer size is in MBs and ranges from 64MB to 128MB for Amazon S3 destination, with is 128MB being the default value. Note that in circumstances where data delivery to the destination is falling behind data ingestion into the delivery stream, Kinesis Data Firehose raises the buffer size automatically to catch up and make sure that all data is delivered to the destination.

Q: How is buffer size applied if I choose to compress my data?

Buffer size is applied before compression. As a result, if you choose to compress your data, the size of the objects within your Amazon S3 bucket can be smaller than the buffer size you specify.

Q: What privilege is required for the Amazon Redshift user that I need to specify while creating a delivery stream?

The Amazon Redshift user needs to have Redshift INSERT privilege for copying data from your Amazon S3 bucket to your Redshift cluster.

Q: What do I need to do if my Amazon Redshift cluster is within a VPC?

If your Amazon Redshift cluster is within a VPC, you need to grant Amazon Kinesis Data Firehose access to your Redshift cluster by unblocking Firehose IP addresses from your VPC. For information about how to unblock IPs to your VPC, see Grant Firehose Access to an Amazon Redshift Destination in the Amazon Kinesis Data Firehose developer guide.

Q: Why do I need to provide an Amazon S3 bucket while choosing Amazon Redshift as destination?

For Amazon Redshift destination, Amazon Kinesis Data Firehose delivers data to your Amazon S3 bucket first and then issues Redshift COPY command to load data from your S3 bucket to your Redshift cluster.

Q: What is Amazon OpenSearch Service (successor to Amazon Elasticsearch Service)?

Amazon OpenSearch Service makes it easy for you to perform interactive log analytics, real-time application monitoring, website search, and more. OpenSearch is an open source, distributed search and analytics suite derived from Elasticsearch. Amazon OpenSearch Service offers the latest versions of OpenSearch, support for 19 versions of Elasticsearch (1.5 to 7.10 versions), and visualization capabilities powered by OpenSearch Dashboards and Kibana (1.5 to 7.10 versions). Click here for more information on Amazon OpenSearch.

Q: What is index rotation for Amazon OpenSearch Service destination?

Kinesis Data Firehose can rotate your Amazon OpenSearch Service index based on a time duration. You can configure this time duration while creating your delivery stream. For more information, see Index Rotation for the Amazon OpenSearch Destination in the Amazon Kinesis Data Firehose developer guide.

Q: Why do I need to provide an Amazon S3 bucket when choosing Amazon OpenSearch Service as destination?

When loading data into Amazon OpenSearch Service, Kinesis Data Firehose can back up all of the data or only the data that failed to deliver. To take advantage of this feature and prevent any data loss, you need to provide a backup Amazon S3 bucket.

Q: Can I change the configurations of my delivery stream after it's created?

You can change the configuration of your delivery stream at any time after it's created. You can do so by using the Firehose Console or the UpdateDestination operation. Your delivery stream remains in ACTIVE state while your configurations are updated and you can continue to send data to your delivery stream. The updated configurations normally take effect within a few minutes.

When delivering to a VPC destination, you can change the destination endpoint URL, as long as new destination is accessible within the same VPC, subnets and security groups. For changes of VPC, subnets and security groups, you need to re-create the Firehose delivery stream.

Q: Can I use a Kinesis Data Firehose delivery stream in one account to deliver my data into an Amazon OpenSearch Service domain VPC destination in a different account?

No, your Kinesis Data Firehose delivery stream and destination Amazon OpenSearch Service domain need to be in the same account.

Q: Can I use a Kinesis Data Firehose delivery stream in one region to deliver my data into an Amazon OpenSearch Service domain VPC destination in a different region?

No, your Kinesis Data Firehose delivery stream and destination Amazon OpenSearch Service domain need to be in the same region.

Q: How often does Kinesis Data Firehose deliver data to my Amazon OpenSearch domain?

The frequency of data delivery to Amazon OpenSearch Service is determined by the OpenSearch buffer size and buffer interval values that you configured for your delivery stream. Firehose buffers incoming data before delivering it to Amazon OpenSearch Service. You can configure the values for OpenSearch buffer size (1 MB to 100 MB) or buffer interval (60 to 900 seconds), and the condition satisfied first triggers data delivery to Amazon OpenSearch Service. Note that in circumstances where data delivery to the destination is falling behind data ingestion into the delivery stream, Amazon Kinesis Data Firehose raises the buffer size automatically to catch up and make sure that all data is delivered to the destination.

Q: What is the manifests folder in my Amazon S3 bucket?
For Amazon Redshift destination, Amazon Kinesis Data Firehose generates manifest files to load Amazon S3 objects to Redshift cluster in batch. The manifests folder stores the manifest files generated by Firehose.

Q: How do backed up OpenSearch documents look like in my Amazon S3 bucket?
If "all documents" mode is used, Amazon Kinesis Data Firehose concatenates multiple incoming records based on buffering configuration of your delivery stream, and then delivers them to your S3 bucket as an S3 object. Regardless of which backup mode is configured, the failed documents are delivered to your S3 bucket using a certain JSON format that provides additional information such as error code and time of delivery attempt. For more information, see Amazon S3 Backup for the Amazon ES Destination in the Amazon Kinesis Data Firehose developer guide.

Q: Can a single delivery stream deliver data to multiple Amazon S3 buckets?

A single delivery stream can only deliver data to one Amazon S3 bucket currently. If you want to have data delivered to multiple S3 buckets, you can create multiple delivery streams.

Q: Can a single delivery stream deliver data to multiple Amazon Redshift clusters or tables?

A single delivery stream can only deliver data to one Amazon Redshift cluster and one table currently. If you want to have data delivered to multiple Redshift clusters or tables, you can create multiple delivery streams.

Q: Can a single delivery stream deliver data to multiple Amazon OpenSearch Service domains or indexes?

A single delivery stream can only deliver data to one Amazon OpenSearch Service domain and one index currently. If you want to have data delivered to multiple Amazon OpenSearch domains or indexes, you can create multiple delivery streams.

Q: How does Amazon Kinesis Data Firehose deliver data to my Amazon OpenSearch Service domain into a VPC?

When you enable Kinesis Data Firehose to deliver data to an Amazon OpenSearch Service destination in a VPC, Amazon Kinesis Data Firehose creates one or more cross account elastic network interfaces (ENI) in your VPC for each subnet(s) that you choose. Amazon Kinesis Data Firehose uses these ENIs to deliver the data into your VPC. The number of ENIs scales automatically to meet the service requirements.

Troubleshooting and managing delivery streams

Q: Why do I get throttled when sending data to my Amazon Kinesis Data Firehose delivery stream?

By default, each delivery stream can intake up to 2,000 transactions/second, 5,000 records/second, and 5 MB/second. You can have this limit increased easily by submitting a service limit increase form.

Q: Why do I see duplicated records in my Amazon S3 bucket, Amazon Redshift table, Amazon OpenSearch index, or Splunk clusters?

Amazon Kinesis Data Firehose uses at least once semantics for data delivery. In rare circumstances such as request timeout upon data delivery attempt, delivery retry by Firehose could introduce duplicates if the previous request eventually goes through.

Q: What happens if data delivery to my Amazon S3 bucket fails?

If your data source is Direct put and the data delivery to your Amazon S3 bucket fails, then Amazon Kinesis Data Firehose will retry to deliver data every 5 seconds for up to a maximum period of 24 hours. If the issue continues beyond the 24-hour maximum retention period, then Amazon Kinesis Data Firehose discards the data.

If your data source is Kinesis Data Streams and the data delivery to your Amazon S3 bucket fails, then Amazon Kinesis Data Firehose will retry to deliver data every 5 seconds for up to a maximum period of what is configured on Kinesis Data Streams.

Q: What happens if data delivery to my Amazon Redshift cluster fails?

If data delivery to your Amazon Redshift cluster fails, Amazon Kinesis Data Firehose retries data delivery every 5 minutes for up to a maximum period of 120 minutes. After 120 minutes, Amazon Kinesis Data Firehose skips the current batch of S3 objects that are ready for COPY and moves on to the next batch. The information about the skipped objects is delivered to your S3 bucket as a manifest file in the errors folder, which you can use for manual backfill. For information about how to COPY data manually with manifest files, see Using a Manifest to Specify Data Files.

Q: What happens if data delivery to my Amazon OpenSearch domain fails?

For Amazon OpenSearch Service destination, you can specify a retry duration between 0 and 7200 seconds when creating the delivery stream. If data delivery to your Amazon ES domain fails, Amazon Kinesis Data Firehose retries data delivery for the specified time duration. After the retrial period, Amazon Kinesis Data Firehose skips the current batch of data and moves on to the next batch. Details on skipped documents are delivered to your S3 bucket in the opensearch_failed folder, which you can use for manual backfill.

Q: What happens if there is a data transformation failure?

There are two types of failure scenarios when Firehose attempts to invoke your Lambda function for data transformation:

- The first type is when the function invocation fails for reasons such as reaching network timeout, and hitting Lambda invocation limits. Under these failure scenarios, Firehose retries the invocation for three times by default and then skips that particular batch of records. The skipped records are treated as unsuccessfully processed records. You can configure the number of invocation re-trials between 0 and 300 using the CreateDeliveryStream and UpdateDeliveryStream APIs. For this type of failure, you can also use Firehose's error logging feature to emit invocation errors to CloudWatch Logs. For more information, see Monitoring with Amazon CloudWatch Logs.
- The second type of failure scenario occurs when a record's transformation result is set to "ProcessingFailed" when it is returned from your Lambda function. Firehose treats these records as unsuccessfully processed records. For this type of failure, you can use Lambda's logging feature to emit error logs to CloudWatch Logs. For more information, see Accessing Amazon CloudWatch Logs for AWS Lambda.

For both types of failure scenarios, the unsuccessfully processed records are delivered to your S3 bucket in the processing_failed folder.

Q: Why is the size of delivered S3 objects larger than the buffer size I specified in my delivery stream configuration?

The size of delivered S3 objects should reflect the specified buffer size most of the time if buffer size condition is satisfied before buffer interval condition. However, when data delivery to destination is falling behind data writing to delivery stream, Firehose raises buffer size dynamically to catch up and make sure that all data is delivered to the destination. In these circumstances, the size of delivered S3 objects might be larger than the specified buffer size.

Q: What is the errors folder in my Amazon S3 bucket?

The errors folder stores manifest files that contain information of S3 objects that failed to load to your Amazon Redshift cluster. You can reload these objects manually through Redshift COPY command. For information about how to COPY data manually with manifest files, see Using a Manifest to Specify Data Files.

Q: What is the opensearch_failed folder in my Amazon S3 bucket?

The opensearch_failed folder stores the documents that failed to load to your Amazon OpenSearch What happens if data delivery to my Amazon OpenSearch domain fails?domain. You can re-index these documents manually for backfill.

Q: What is the processing_failed folder in my Amazon S3 bucket?

The processing_failed folder stores the records that failed to transform in your AWS Lambda function. You can re-process these records manually.

Q: How do I monitor the operations and performance of my Amazon Kinesis Data Firehose delivery stream?

Firehose Console displays key operational and performance metrics such as incoming data volume and delivered data volume. Amazon Kinesis Data Firehose also integrates with Amazon CloudWatch Metrics so that you can collect, view, and analyze metrics for your delivery streams. For more information about Amazon Kinesis Data Firehose metrics, see Monitoring with Amazon CloudWatch Metrics in the Amazon Kinesis Data Firehose developer guide.

Q: How do I monitor data transformation and delivery failures of my Amazon Kinesis Data Firehose delivery stream?

Amazon Kinesis Data Firehose integrates with Amazon CloudWatch Logs so that you can view the specific error logs if data transformation or delivery fails. You can enable error logging when creating your delivery stream. For more information, see Monitoring with Amazon CloudWatch Logs in the Amazon Kinesis Data Firehose developer guide.

Q: How do I manage and control access to my Amazon Kinesis Data Firehose delivery stream?

Amazon Kinesis Data Firehose integrates with AWS Identity and Access Management, a service that enables you to securely control access to your AWS services and resources for your users. For example, you can create a policy that only allows a specific user or group to add data to your Firehose delivery stream. For more information about access management and control of your stream, see Controlling Access with Amazon Kinesis Data Firehose.

Q: How do I log API calls made to my Amazon Kinesis Data Firehose delivery stream for security analysis and operational troubleshooting?

Amazon Kinesis Data Firehose integrates with AWS CloudTrail, a service that records AWS API calls for your account and delivers log files to you. For more information about API call logging and a list of supported Amazon Kinesis Data Firehose API operations, see Logging Amazon Kinesis Data Firehose API calls Using AWS CloudTrail.

Amazon Kinesis Data Analytics FAQs

General

Q: What is Amazon Kinesis Data Analytics?
Amazon Kinesis Data Analytics is the easiest way to transform and analyze streaming data in real time with Apache Flink. Apache Flink is an open-source framework and engine for processing data streams. Kinesis Data Analytics reduces the complexity of building, managing, and integrating Apache Flink applications with other AWS services.

Kinesis Data Analytics takes care of everything required to run streaming applications continuously, and scales automatically to match the volume and throughput of your incoming data. With Kinesis Data Analytics, there are no servers to manage, no minimum fee or setup cost, and you only pay for the resources your streaming applications consume.

Q: What is real-time stream processing and why do I need it?
Companies are ingesting data faster than ever due to an explosive growth of real-time data sources. Whether you are handling log data from mobile and web applications, purchase data

from e-commerce platforms, or sensor data from IoT devices, ingesting data in real time helps you learn what your customers, organization, and business are doing right now.

Q: What can I do with Kinesis Data Analytics?
You can use Kinesis Data Analytics for many use cases to process data continuously, getting insights in seconds or minutes rather than waiting days or even weeks. Kinesis Data Analytics enables you to quickly build end-to-end stream processing applications for log analytics, clickstream analytics, Internet of Things (IoT), ad tech, gaming, and more. The four most common use cases are streaming extract-transform-load (ETL), continuous metric generation, responsive real-time analytics, and interactive querying of data streams.

Streaming ETL
Streaming ETL applications enable you to clean, enrich, organize, and transform raw data prior to loading your data lake or data warehouse in real-time, reducing or eliminating batch ETL steps. These applications can buffer small records into larger files prior to delivery, and perform sophisticated joins across streams and tables. For example, you can build an application that continuously reads IoT sensor data stored in Amazon Managed Streaming for Apache Kafka (Amazon MSK), organize the data by sensor type, remove duplicate data, normalizes data per a specified schema, and then deliver the data to Amazon Simple Storage Service (Amazon S3).

Continuous metric generation
Continuous metric generation applications let you monitor and understand how your data is trending over time. Your applications can aggregate streaming data into critical information and seamlessly integrate it with reporting databases and monitoring services to serve your applications and users in real-time. With Kinesis Data Analytics, you can use Apache Flink code or SQL to continuously generate time-series analytics over time windows. For example, you can build a live leaderboard for a mobile game by computing the top players every minute and then sending it to Amazon DynamoDB. Or, you can track the traffic to your website by calculating the number of unique website visitors every five minutes and then sending the processed results to Amazon Redshift.

Responsive real-time analytics
Responsive real-time analytics applications send real-time alarms or notifications when certain metrics reach predefined thresholds, or in more advanced cases, when your application detects anomalies using machine learning (ML) algorithms. These applications enable you to respond immediately to changes in your business in real-time like predicting user abandonment in mobile apps and identifying degraded systems. For example, an application can compute the availability or success rate of a customer-facing API over time, and then send results to Amazon CloudWatch. You can build another application to look for events that meet certain criteria, and then automatically notify the right customers using Amazon Kinesis Data Streams and Amazon Simple Notification Service (SNS).

Interactive analysis of data streams
Interactive analysis enables streaming data exploration in real time. With ad hoc queries or programs, you can inspect streams from Amazon MSK or Amazon Kinesis Data Streams and visualize how data looks like within those streams. For example, you can view how a real- time metric that computes the average over a time window behaves and send the aggregated data to a destination of your choice. Interactive analysis also helps with iterative development of stream processing applications. The queries you build will continuously update as new data arrives. With Kinesis Data Analytics Studio you can deploy these queries to run continuously with autoscaling and durable state backups enabled.

Q: How do I get started with Apache Flink applications for Kinesis Data Analytics?

Sign into the Amazon Kinesis Data Analytics console and create a new stream processing application. You can also use the AWS CLI and AWS SDKs. Once you create an application, go to your favorite Integrated Development Environment, connect to AWS, and install the open-source Apache Flink libraries and AWS SDKs in your language of choice. Apache Flink is an open- source framework and engine for processing data streams and AWS SDKs. The extensible libraries include more than 25 pre-built stream processing operators like window and aggregate, and AWS service integrations like Amazon MSK, Amazon Kinesis Data Streams, and Amazon Kinesis Data Firehose. Once built, you upload your code to Kinesis Data Analytics and the service takes care of everything required to run your real-time applications, continuously including scaling automatically to match the volume and throughput of your incoming data.

Q: How do I get started with Apache Beam applications for Kinesis Data Analytics?
Using Apache Beam to create your Kinesis Data Analytics application is very similar to getting started with Apache Flink. Please follow the instructions in the question above and be sure to install any components necessary for applications to run on Apache Beam, per the instructions in the Developer Guide. Note that Kinesis Data Analytics supports Java SDK's only when running on Apache Beam.

Q: How do I get started with Amazon Kinesis Data Analytics Studio?
You can get started from the Amazon Kinesis Data Analytics console and create a new Studio notebook. Once you start the notebook, you can open it in Apache Zeppelin to immediately write code in SQL, Python, or Scala. You can interactively develop applications using the notebook interface for Amazon Kinesis Data Streams, Amazon MSK, and Amazon S3 using built-in integrations, and various other sources with custom connectors. You can use all the operators that Apache Flink supports in Flink SQL and the Table API to perform ad hoc data stream querying and develop your stream processing application. Once you are ready, with a few clicks, you can easily promote your code to a continuously running stream processing application with autoscaling and durable state.

Q: How do I get started with Kinesis Data Analytics for SQL?
Sign into the Amazon Kinesis Data Analytics console and create a new stream processing application. You can also use the AWS CLI and AWS SDKs. You can build an end-to-end application in three simple steps: 1) configure incoming streaming data, 2) write your SQL queries, and 3) point to where you want the results loaded. Kinesis Data Analytics recognizes standard data formats such as JSON, CSV, and TSV, and automatically creates a baseline schema. You can refine this schema, or if your data is unstructured, you can define a new one using our intuitive schema editor. Then, the service applies the schema to the input stream and makes it look like a SQL table that is continually updated so that you can write standard SQL queries against it. You use our SQL editor to build your queries.

The SQL editor comes with all the bells and whistles including syntax checking and testing against live data. We also give you templates that provide the SQL code for anything from a simple stream filter to advanced anomaly detection and top-K analysis. Kinesis Data Analytics takes care of provisioning and elastically scaling all of the infrastructure to handle any data throughput. You don't need to plan, provision, or manage infrastructure.

Q: What are the limits of Kinesis Data Analytics?
Kinesis Data Analytics elastically scales your application to accommodate for the data throughput of your source stream and your query complexity for most scenarios. For detailed information on service limits, see Limits in the Amazon Kinesis Data Analytics for SQL Developer Guide. For detailed information on service Limits for Apache Flink applications, visit the Limits section in the Amazon Kinesis Data Analytics for Apache Flink Developer Guide.

Q: Does Amazon Kinesis Data Analytics support schema registration?
Yes, using Apache Flink DataStream Connectors, Amazon Kinesis Data Analytics for Apache

Flink applications can use AWS Glue Schema Registry, a serverless feature of AWS Glue. You can integrate Apache Kafka/Amazon MSK and Amazon Kinesis Data Streams, as a sink or a source, with your Amazon Kinesis Data Analytics for Apache Flink workloads. Visit the Schema Registry user documentation to get started and to learn more.

Key concepts

Q: What is a Kinesis Data Analytics application?
An application is the Kinesis Data Analytics entity that you work with. Kinesis Data Analytics applications continuously read and process streaming data in real time. You write application code in a language supported by Apache Flink to process the incoming streaming data and produce output. Then, Kinesis Data Analytics writes the output to a configured destination.

Each application consists of three primary components:

Input – The streaming source for your application. In the input configuration, you map the streaming source to an in-application data stream(s). Data flows from in your data source(s) into your in-application data streams. You process data from these in-application data streams using your application code, sending processed data to subsequent in-application data streams or destinations. You add inputs inside application code for Apache Flink applications and Studio notebooks, and via the API for Kinesis Data Analytics for SQL applications.

Application code – A series of Apache Flink operators or SQL statements that process input and produce output. In its simplest form, application code can be a single Apache Flink operator or SQL statement that reads from an in-application data stream associated with a streaming source and writes to an in-application data stream associated with an output. For a Studio notebook, this could be a simple Flink SQL select query, with the results shown in context within the notebook. You can write Apache Flink code in its supported languages for Kinesis Data Analytics for Apache Flink applications or Studio notebooks, or SQL code that splits the initial in-application data stream into multiple streams and applies additional logic to these separate streams for Kinesis Data Analytics for SQL applications.

Output – You can create one or more in-application streams to store intermediate results. You can then optionally configure an application output to persist data from specific in- application streams to an external destination. You add these outputs inside application code for Apache Flink applications and Studio notebooks, and to Kinesis Data Analytics for SQL applications.

Q: What is an in-application data stream?
An in-application data stream is an entity that continuously stores data in your application for processing. Your applications continuously write to and read from in-application data streams. For Apache Flink and Studio applications, you interact with in-application stream by processing data via stream operators. Operators transform one or more data streams into a new data stream. For SQL applications, you interact with an in-application stream in the same way you would a SQL table by using SQL statements. You apply SQL statements to one or more data streams and insert the results into a new data stream.

Q: What application code is supported?
Kinesis Data Analytics supports applications built using Java, Scala, and Python with the open-source Apache Flink libraries and your own custom code. Kinesis Data Analytics also supports applications built using Java with the open-source Apache Beam libraries and your own customer code. Kinesis Data Analytics Studio supports code built using Apache Flink-compatible SQL, Python, and Scala.

Managing applications

Q: How can I monitor the operations and performance of my Kinesis Data Analytics applications?
AWS provides various tools that you can use to monitor your Kinesis Data Analytics applications including access to the Flink Dashboard for Apache Flink applications. You can configure some of these tools to do the monitoring for you. For more information about how to monitor your application, see:

- Monitoring Kinesis Data Analytics in the Amazon Kinesis Data Analytics for Apache Flink Developer Guide.
- Monitoring Kinesis Data Analytics in the Amazon Kinesis Data Analytics for Studio Developer Guide.
- Monitoring Kinesis Data Analytics in the Amazon Kinesis Data Analytics for SQL Developer Guide.

Q: How do I manage and control access to my Kinesis Data Analytics applications?
Kinesis Data Analytics needs permissions to read records from the streaming data sources you specify in your application. Kinesis Data Analytics also needs permissions to write your application output to specified destinations in your application output configuration. You can grant these permissions by creating AWS Identity and Access Management (IAM) roles that Kinesis Data Analytics can assume. The permissions you grant to this role determine what Kinesis Data Analytics can do when the service assumes the role. For more information, see:

- Granting Permissions in the Amazon Kinesis Data Analytics for Apache Flink Developer Guide.
- Granting Permissions in the Amazon Kinesis Data Analytics Studio Developer Guide.
- Granting Permissions in the Amazon Kinesis Data Analytics for SQL Developer Guide.

Q: How does Kinesis Data Analytics scale my application?
Kinesis Data Analytics elastically scales your application to accommodate the data throughput of your source stream and your query complexity for most scenarios. Kinesis Data Analytics provisions capacity in the form of Amazon Kinesis Processing Units (KPU). One KPU provides you with 1 vCPU and 4 GB memory.

For Apache Flink applications and Studio notebooks, Kinesis Data Analytics assigns 50 GB of running application storage per KPU that your application uses for checkpoints and is available for you to use via temporary disk. A checkpoint is an up-to-date backup of a running application used to recover immediately from an application disruption. You can also control the parallel execution for your Kinesis Data Analytics for Apache Flink application tasks (such as reading from a source or executing an operator) using the Parallelism and ParallelismPerKPU parameters in the API. Parallelism defines the number of concurrent instances of a task. All operators, sources, and sinks execute with a defined parallelism, by default one. Parallelism per KPU defines the amount of the number of parallel tasks that can be scheduled per Kinesis Processing Unit (KPU) of your application, by default one. For more information, see Scaling in the Amazon Kinesis Data Analytics for Apache Flink Developer Guide.

For SQL applications, each streaming source is mapped to a corresponding in-application stream. While this is not required for many customers, you can more efficiently use KPUs by increasing the number of in-application streams your source is mapped to by specifying the input parallelism parameter. Kinesis Data Analytics evenly assigns the streaming data source's partitions, such as Amazon Kinesis data stream's shards, to the number of in- application data streams that you specified. For example, if you have a 10-shard Amazon Kinesis data stream as

a streaming data source and you specify an input parallelism of two, Kinesis Data Analytics assigns five Amazon Kinesis shards to two in-application streams named "SOURCE_SQL_STREAM_001" and "SOURCE_SQL_STREAM_002". For more information, see Configuring Application Input in the Amazon Kinesis Data Analytics for SQL Developer Guide.

Q: What are the best practices associated for building and managing my Kinesis Data Analytics applications?
For information about best practices for Apache Flink, see the Best Practices section of the Amazon Kinesis Data Analytics for Apache Flink Developer Guide. The section covers best practices for fault tolerance, performance, logging, coding, and more.

For information about best practices for Amazon Kinesis Data Analytics Studio, see the Best Practices section of the Amazon Kinesis Data Analytics Studio Developer Guide. The section covers best practices and samples for SQL, Python, and Scala applications, requirements for deploying your code as a continuously running stream processing application, performance, logging, and more.

For information about best practices for SQL, see the Best Practices section of the Amazon Kinesis Data Analytics for SQL Developer Guide. The section covers managing applications, defining input schema, connecting to outputs, and authoring application code.

Q: Can I access resources behind an Amazon VPC with a Kinesis Data Analytics for Apache Flink application?
Yes. You can access resources behind an Amazon VPC. You can learn how to configure your application for VPC access in the Using an Amazon VPC section of the Amazon Kinesis Data Analytics Developer Guide.

Q: Can a single Kinesis Data Analytics for Apache Flink application have access to multiple VPCs?
No. If multiple subnets are specified, they must all be in the same VPC. You can connect to other VPCs by peering your VPCs.

Q: Can a Kinesis Data Analytics for Apache Flink application connected to a VPC also be able to access the internet and AWS Service endpoints?
Kinesis Data Analytics for Apache Flink applications and Kinesis Data Analytics Studio notebooks that are configured to access resources in a particular VPC will not have access to the internet as a default configuration. You can learn how to configure access to the internet for your application in the Internet and Service Access section of the Amazon Kinesis Data Analytics Developer Guide.

Building Apache Flink Applications

Authoring application code for applications using Apache Flink

Q: What is Apache Flink?
Apache Flink is an open-source framework and engine for stream and batch data processing. It makes streaming applications easy to build, because it provides powerful operators and solves the core streaming problems like duplicate processing very well. Apache Flink provides data distribution, communication, and fault tolerance for distributed computations over data streams.

Q: How do I develop applications?
You can start by downloading the open-source libraries including the AWS SDK, Apache Flink, and connectors for AWS services. Get instructions on how to download the libraries and create your first application in the Amazon Kinesis Data Analytics for Apache Flink Developer Guide.

Q: What does my application code look like?
You write your Apache Flink code using data streams and stream operators. Application data streams are the data structure you perform processing against using your code. Data continuously flows from the sources into application data streams. One or more stream operators are used to define your processing on the application data streams, including transform, partition, aggregate, join, and window. Data streams and operators can be connected in serial and parallel chains. A short example using pseudo code is shown below.

```
DataStream <GameEvent> rawEvents = env.addSource(
            New KinesisStreamSource("input_events"));
DataStream <UserPerLevel> gameStream =
            rawEvents.map(event - > new UserPerLevel(event.gameMetadata.gameId,
                                    event.gameMetadata.levelId,event.userId));
gameStream.keyBy(event -> event.gameId)
        .keyBy(1)
        .window(TumblingProcessingTimeWindows.of(Time.minutes(1)))
        .apply(...) - > {...};
gameStream.addSink(new KinesisStreamSink("myGameStateStream"));
```

Q: How do I use the operators?
Operators take an application data stream as input and send processed data to an application data stream as output. Operators can be connected to build applications with multiple steps and don't require advanced knowledge of distributed systems to implement and operate.

Q: What operators are supported?
Kinesis Data Analytics for Apache Flink includes over 25 operators from Apache Flink that can be used to solve a wide variety of use cases including Map, KeyBy, aggregations, Window Join, and Window. Map allows you to perform arbitrary processing, taking one element from an incoming data stream and producing another element. KeyBy logically organizes data using a specified key enabling you to process similar data points together. Aggregations performs processing across multiple keys like sum, min, and max. Window Join joins two data streams together on a given key and window. Window group date using a key and typically time-based operation, like counting the number of unique items over a five-minute time-period.

You can build custom operators if these do not meet your needs. You can find more examples in the Operators section of the Amazon Kinesis Data Analytics for Apache Flink Developer Guide. You can find a full list of Apache Flink operators in the Operators section of the Apache Flink documentation.

Q: What integrations are supported in a Kinesis Data Analytics for Apache Flink application?
You can setup pre-built integrations with minimal code, or build your own integration to connect to virtually any data source. The open-source libraries based on Apache Flink support streaming sources and destinations, or sinks, to process data delivery. This also includes data enrichment support via asynchronous input/output (I/O) connectors. A list of specific connectors included in the open-source libraries is shown below.

- Streaming data sources: Amazon Managed Streaming for Apache Kafka (Amazon MSK), Amazon Kinesis Data Streams • Destinations, or sinks: Amazon Kinesis Data Streams
- Amazon Kinesis Data Firehose, Amazon DynamoDB, Amazon Elasticsearch Service, and Amazon S3 (through file sink integrations)

Apache Flink also includes other connectors including Apache Kafka, Apache Casssandra, Elasticsearch, and more.

Q: Can Kinesis Data Analytics for Apache Flink applications replicate data across streams and/or topics?
Yes. You can use Kinesis Data Analytics Apache Flink applications to replicate data between Amazon Kinesis Data Streams, Amazon MSK, and other systems. An example provided in our documentation demonstrates how to read from one Amazon MSK topic and write to another.

Q: Are custom integrations supported?
You can add a source or destination to your application by building upon a set of primitives enabling you to read and write from files, directories, sockets, or anything that you can access over the internet. Apache Flink provides these primitives for data sources and data sinks. The primitives come with configurations like the ability to read and write data continuously or once, asynchronously or synchronously, and much more. For example, you can setup an application to read continuously from Amazon S3 by extending the existing file-based source integration.

Q: What delivery model does Kinesis Data Analytics for Apache Flink applications provide?
Apache Flink applications in Kinesis Data Analytics use an "exactly once" delivery model if an application is built using idempotent operators, including sources and sinks. This means the processed data will impact downstream results once and only once. Checkpoints save the current application state and enable Kinesis Data Analytics for Apache Flink applications to recover the position of the application to provide the same semantics as a failure-free execution. Checkpoints for Apache Flink applications are provided via Apache Flink's checkpointing functionality. By default, Kinesis Data Analytics for Apache Flink applications use the Apache Flink exactly-once semantics. Your application will support exactly once processing semantics if you design your applications using sources, operators, and sinks that utilize Apache Flink's exactly once semantics.

Q: Do I have access to local storage from my application storage?
Yes. Kinesis Data Analytics for Apache Flink applications provides your application 50 GB of running application storage per Kinesis Processing Unit (KPU). Kinesis Data Analytics scales storage with your application. Running application storage is used for saving application state using checkpoints. It is also accessible to your application code to use as temporary disk for caching data or any other purpose. Kinesis Data Analytics can remove data from running application storage not saved via checkpoints (e.g operators, sources, sinks) at any time. All data stored in running application storage is encrypted at rest.

Q: How does Kinesis Data Analytics for Apache Flink automatically backup my application?
Kinesis Data Analytics automatically backs up your running application's state using checkpoints and snapshots. Checkpoints save the current application state and enable Kinesis Data Analytics for Apache Flink applications to recover the application position to provide the same semantics as a failure-free execution. Checkpoints utilize running application storage. Snapshots save a point in time recovery point for applications. Snapshots utilize durable application backups.

Q: What are application snapshots?
Snapshots enable you to create and restore your application to a previous point in time. This enables you to maintain previous application state and rollback your application at any time. You control how snapshots you have at any given from zero to thousands of snapshots. Snapshots use durable application backups and Kinesis Data Analytics charges you based on their size. Kinesis Data Analytics encrypts data saved in snapshots by default. You can delete individual snapshots through the API or all snapshots by deleting your application.

Q: What versions of Apache Flink are supported?
To learn more about supported Apache Flink versions please see the Amazon Kinesis Data Analytics Release Notes Page. This page also includes the versions of Apache Beam, Java, Python, and AWS SDKs that Kinesis Data Analytics supports.

Q: Can Kinesis Data Analytics for Apache Flink applications run on Apache Beam?
Yes, Kinesis Data Analytics for Apache Flink supports streaming applications built using Apache Beam. You can build Apache Beam streaming applications in Java and run them in different engines and services including using Apache Flink on Amazon Kinesis Data Analytics. You can find information regarding supported Apache Flink and Apache Beam versions here in the Amazon Kinesis Data Analytics documentation.

Building Amazon Kinesis Data Analytics Studio Applications

Q: How do I develop a Studio application?
You can start from the Amazon Kinesis Data Analytics Studio, Amazon Kinesis Data Streams, or Amazon MSK consoles with a few clicks to launch a serverless notebook to immediately query data streams and perform interactive data analytics.

Interactive data analytics: You can write code in the notebook in SQL, Python, or Scala to interact with your streaming data, with query response times in seconds. You can use built-in visualizations to explore the data and view real-time insights on your streaming data from within your notebook, and easily develop stream processing applications powered by Apache Flink.

Once your code is ready to run as a production application, you can transition with a single click to a stream processing application that processes GBs of data per second, without servers.

Stream processing application: Once you are ready to promote your code to production, you can build your code by clicking. You can click on 'Deploy as stream processing application' in the notebook interface or issue a single command in the CLI, and Studio takes care of all the infrastructure management necessary for you to run your stream processing application at scale, with autoscaling and durable state enabled, just as in an Amazon Kinesis Data Analytics for Apache Flink application.

Q: What does my application code look like?
You can write code in the notebook in your preferred language of SQL, Python, or Scala using Apache Flink's Table API. The Table API is a high-level abstraction and relational API that supports a superset of SQL's capabilities. It offers familiar operations such as select, filter, join, group by, aggregate, etc., along with stream specific concepts like windowing. You use %<interpreter> to specify the language to be used in a section of the notebook, and can easily switch between languages. Interpreters are Apache Zeppelin plug-ins enabling developers to specify a language or data processing engine for each section of the notebook. You can also build user defined functions and reference them to improve code functionality.

Q: What SQL operations are supported?
You can perform SQL operations such as Scan and Filter (SELECT, WHERE), Aggregations (GROUP BY, GROUP BY WINDOW,HAVING), Set (UNION, UNIONALL, INTERSECT, IN, EXISTS), Order (ORDER BY, LIMIT), Joins (INNER, OUTER, Timed Window –BETWEEN, AND, joining with temporal tables – tables that track changes over time), Top N, deduplication, and pattern recognition. Some of these queries such as GROUP BY, OUTER JOIN, and Top N are "results updating" for streaming data, which means that the results are continuously updating as the streaming data is processed. Other DDL statements such as CREATE, ALTER, and DROP are also supported. For a complete list of queries and samples, see https://ci.apache.org/projects/flink/flink-docs-release-1.11/dev/table/sql/queries.html.

Q: How are Python and Scala supported?
Apache Flink's Table API supports Python and Scala through language integration using Python strings and Scala expressions. The operations supported are very similar to the SQL operations supported, including select, order, group, join, filter, and windowing. A full list of operations and samples are included in our documentation here.

Q: What versions of Apache Flink and Apache Zeppelin are supported?
To learn more about supported Apache Flink versions please see the Amazon Kinesis Data Analytics Release Notes Page. This page also includes the versions Apache Zeppelin, Apache Beam, Java, Python, and AWS SDKs that Kinesis Data Analytics supports.

Q: What integrations are supported in a Kinesis Data Analytics Studio application by default?

- Data sources: Amazon Managed Streaming for Apache Kafka (Amazon MSK), Amazon Kinesis Data Streams, Amazon S3
- Destinations, or sinks: Amazon MSK, Amazon Kinesis Data Streams, and Amazon S3

Q: Are custom integrations supported?
You can configure additional integrations with a few additional steps and lines of Apache Flink code (Python, Scala, or Java) to define connections with all Apache Flink supported integrations including destinations such as Amazon OpenSearch Service, Amazon ElastiCache for Redis, Amazon Aurora, Amazon Redshift, Amazon DynamoDB, Amazon Keyspaces, and more. You can attach executables for these custom connectors when you create or configure your Studio application.

Q: Should I develop with Kinesis Data Analytics Studio or Kinesis Data Analytics SQL?
We recommend getting started with Kinesis Data Analytics Studio as it offers a more comprehensive stream processing experience with exactly-once processing. Kinesis Data Analytics Studio offers stream processing application development in your language of choice (SQL, Python, and Scala), scales to GB/s of processing, supports long running computations over hours or even days, performs code updates within seconds, handles multiple input streams, and works with a variety of input streams including Amazon Kinesis Data Streams and Amazon MSK.

Building Kinesis Data Analytics SQL applications

For new projects, we recommend that you use the new Kinesis Data Analytics Studio over Kinesis Data Analytics for SQL Applications. Kinesis Data Analytics Studio combines ease of use with advanced analytical capabilities, enabling you to build sophisticated stream processing applications in minutes.

Configuring input for SQL applications

Q: What inputs are supported in a Kinesis Data Analytics SQL application?
SQL applications in Kinesis Data Analytics support two types of inputs: streaming data sources and reference data sources. A streaming data source is continuously generated data that is read into your application for processing. A reference data source is static data your application uses to enrich data coming in from streaming sources. Each application can have no more than one streaming data source and no more than one reference data source. An application continuously reads and processes new data from streaming data sources, including Amazon Kinesis Data Streams or Amazon Kinesis Data Firehose. An application reads a reference data source, including Amazon S3, in its entirety for use in enriching the streaming data source through SQL JOINs.

Q: What is a reference data source?
A reference data source is static data that your application uses to enrich data coming in from streaming sources. You store reference data as an object in your S3 bucket. When the SQL application starts, Kinesis Data Analytics reads the S3 object and creates an in-application SQL table to store the reference data. Your application code can then join it with an in- application stream. You can update the data in the SQL table by calling the UpdateApplication API.

Q: How do I set up a streaming data source in my SQL application?
A streaming data source can be an Amazon Kinesis data stream or an Amazon Kinesis Data Firehose delivery stream. Your Kinesis Data Analytics SQL application continuously reads new data from streaming data sources as it arrives in real time. The data is made accessible in your SQL code through an in-application stream. An in-application stream acts like a SQL table because you can create, insert, and select from it. However, the difference is that an in-application stream is continuously updated with new data from the streaming data source.

You can use the AWS Management Console to add a streaming data source. You can learn more about sources in the Configuring Application Input section of the Kinesis Data Analytics for SQL Developer Guide.

Q: How do I set up a reference data source in my SQL application?
A reference data source can be an Amazon S3 object. Your Kinesis Data Analytics SQL application reads the S3 object in its entirety when it starts running. The data is made accessible in your SQL code through a table. The most common use case for using a reference data source is to enrich the data coming from the streaming data source using a SQL JOIN.

Using the AWS CLI, you can add a reference data source by specifying the S3 bucket, object, IAM role, and associated schema. Kinesis Data Analytics loads this data when you start the application, and reloads it each time you make any update API call.

Q: What data formats are supported for SQL applications?
SQL applications in Kinesis Data Analytics can detect the schema and automatically parses UTF-8 encoded JSON and CSV records using the DiscoverInputSchema API. This schema is applied to the data read from the stream as part of the insertion into an in-application stream.

For other UTF-8 encoded data that does not use a delimiter, uses a different delimiter than CSV, or in cases were the discovery API did not fully discover the schema, you can define a schema using the interactive schema editor or use string manipulation functions to structure your data. For more information, see Using the Schema Discovery Feature and Related Editing in the Amazon Kinesis Data Analytics for SQL Developer Guide.

Q: How is my input stream exposed to my SQL code?
Kinesis Data Analytics for SQL applies your specified schema and inserts your data into one or more in-application streams for streaming sources, and a single SQL table for reference sources. The default number of in-application streams is the one that meets the needs of most of your use cases. You should increase this number if you find that your application is not keeping up with the latest data in your source stream as defined by CloudWatch metric MillisBehindLatest. The number of in-application streams required is impacted by both the amount of throughput in your source stream and your query complexity. The parameter for specifying the number of in-application streams that are mapped to your source stream is called input parallelism.

Authoring application code for SQL applications

Q: What does my SQL application code look like?
Application code is a series of SQL statements that process input and produce output. These SQL statements operate on in-application streams and reference tables. An in-application stream is like a continuously updating table on which you can perform the SELECT and INSERT SQL operations. Your configured sources and destinations are exposed to your SQL code through in-application streams. You can also create additional in-application streams to store intermediate query results.

You can use the following pattern to work with in-application streams:

- Always use a SELECT statement in the context of an INSERT statement. When you select rows, you insert results into another in-application stream.
- Use an INSERT statement in the context of a pump.
- You use a pump to make an INSERT statement continuous, and write to an in-application stream.

The following SQL code provides a simple, working application:

```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (
   ticker_symbol VARCHAR(4),
   change DOUBLE,
   price DOUBLE);

CREATE OR REPLACE PUMP "STREAM_PUMP" AS
 INSERT INTO "DESTINATION_SQL_STREAM"
   SELECT STREAM ticker_symbol, change, price
   FROM "SOURCE_SQL_STREAM_001";
```

For more information about application code, see Application Code in the Amazon Kinesis Data Analytics for SQL Developer Guide.

Q: How does Kinesis Data Analytics help me with writing SQL code?
Kinesis Data Analytics includes a library of analytics templates for common use cases including streaming filters, tumbling time windows, and anomaly detection. You can access these templates from the SQL editor in the AWS Management Console. After you create an application and navigate to the SQL editor, the templates are available in the upper-left corner of the console.

Q: How can I perform real-time anomaly detection in Kinesis Data Analytics?
Kinesis Data Analytics includes pre-built SQL functions for several advanced analytics including one for anomaly detection. You can simply make a call to this function from your SQL code for detecting anomalies in real-time. Kinesis Data Analytics uses the Random Cut Forest algorithm to implement anomaly detection. For more information on Random Cut Forests, see the Streaming Data Anomaly Detection whitepaper.

Configuring destinations in SQL applications

Q: What destinations are supported?
Kinesis Data Analytics for SQL supports up to three destinations per application. You can persist SQL results to Amazon S3, Amazon Redshift, and Amazon OpenSearch Service (through Amazon Kinesis Data Firehose), and Amazon Kinesis Data Streams. You can write to a destination not directly supported by Kinesis Data Analytics by sending SQL results to Amazon Kinesis Data Streams, and leveraging its integration with AWS Lambda to send to a destination of your choice.

Q: How do I set up a destination?
In your application code, you write the output of SQL statements to one or more in- application streams. Optionally, you can add an output configuration to your application to persist everything written to specific in-application streams to up to four external destinations. These external destinations can be an Amazon S3 bucket, Amazon Redshift table, Amazon OpenSearch Service domain (through Amazon Kinesis Data Firehose), and an Amazon Kinesis data stream. Each application supports up to four destinations, which can be any combination of the above. For more information, see Configuring Output Streams in the Amazon Kinesis Data Analytics for SQL Developer Guide.

Q: My preferred destination is not directly supported. How can I send SQL results to this destination?
You can use AWS Lambda to write to a destination that is not directly supported using Kinesis Data Analytics for SQL. We recommend that you write results to an Amazon Kinesis data stream, and then use AWS Lambda to read the processed results and send it to the destination of your choice. For more information, see the Example: AWS Lambda Integration in the Amazon Kinesis Data Analytics for SQL Developer Guide. Alternatively, you can use a Kinesis Data Firehose delivery stream to load the data into Amazon S3, and then trigger an AWS Lambda function to read that data and send it to the destination of your choice. For more information, see Using AWS Lambda with Amazon S3 in the AWS Lambda Developer Guide.

Q: What delivery model does Kinesis Data Analytics provide?
SQL applications in Kinesis Data Analytics uses an "at least once" delivery model for application output to the configured destinations. Kinesis Data Analytics applications take internal checkpoints, which are points in time when output records were delivered to the destinations and there was no data loss. The service uses the checkpoints as needed to ensure that your application output is delivered at least once to the configured destinations. For more information about the delivery model, see Configuring Application Output in the Amazon Kinesis Data Analytics for SQL Developer Guide.

Comparison to other stream processing solutions

Q: How does Amazon Kinesis Data Analytics differ from running my own application using the Amazon Kinesis Client Library?
The Amazon Kinesis Client Library (KCL) is a pre-built library that helps you build consumer applications for reading and processing data from an Amazon Kinesis data stream. The KCL handles complex issues such as adapting to changes in data stream volume, load balancing streaming data, coordinating distributed services, and processing data with fault-tolerance. The KCL enables you to focus on business logic while building applications.

With Kinesis Data Analytics, you can process and query real-time, streaming data. You use standard SQL to process your data streams, so you don't have to learn any new programming languages. You just point Kinesis Data Analytics to an incoming data stream, write your SQL queries, and then specify where you want the results loaded. Kinesis Data Analytics uses the KCL to read data from streaming data sources as one part of your underlying application. The service abstracts this from you, as well as many of the more complex concepts associated with using the KCL, such as checkpointing.

If you want a fully managed solution and you want to use SQL to process the data from your data stream, you should use Kinesis Data Analytics. Use the KCL if you need to build a custom processing solution whose requirements are not met by Kinesis Data Analytics, and you are able to manage the resulting consumer application.

Amazon EMR (previously called Amazon Elastic MapReduce)

Easily run and scale Apache Spark, Hive, Presto, and other big data workloads.

How it works

Amazon EMR is the industry-leading cloud big data solution for petabyte-scale data processing, interactive analytics, and machine learning using open-source frameworks such as Apache Spark, Apache Hive, and Presto.

Introducing Amazon EMR Serverless

Amazon EMR Serverless is a new option in Amazon EMR that makes it easy and cost-effective for data engineers and analysts to run applications built using open source big data frameworks such as Apache Spark, Hive or Presto, without having to tune, operate, optimize, secure or manage clusters.

Use cases

Perform big data analytics
Run large-scale data processing and what-if analysis using statistical algorithms and predictive models to uncover hidden patterns, correlations, market trends, and customer preferences.

Build scalable data pipelines
Extract data from a variety of sources, process it at scale, and make it available for applications and users.

Process real-time data streams
Analyze events from streaming data sources in real-time to create long-running, highly available, and fault-tolerant streaming data pipelines.

Accelerate data science and ML adoption
Analyze data using open-source ML frameworks such as Apache Spark MLlib, TensorFlow, and Apache MXNet. Connect to Amazon SageMaker Studio for large-scale model training, analysis, and reporting.

What is Amazon EMR?

Amazon EMR (previously called Amazon Elastic MapReduce) is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. Using these frameworks and related open-source projects, you can process data for analytics purposes and business intelligence workloads. Amazon EMR also lets you transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.

Amazon EMR FAQs

General

Q: What is Amazon EMR?

Amazon EMR is the industry-leading cloud big data platform for data processing, interactive analysis, and machine learning using open source frameworks such as Apache Spark, Apache Hive, and Presto. With EMR you can run petabyte-scale analysis at less than half of the cost of traditional on-premises solutions and over 1.7x faster than standard Apache Spark.

Q: Why should I use Amazon EMR?

Amazon EMR lets you focus on transforming and analyzing your data without having to worry about managing compute capacity or open-source applications, and saves you money. Using EMR, you can instantly provision as much or as little capacity as you like on Amazon EC2 and set up scaling rules to manage changing compute demand. You can set up CloudWatch alerts to notify you of changes in your infrastructure and take actions immediately. If you use Kubernetes, you can also use EMR to submit your workloads to Amazon EKS clusters.. Whether you use EC2 or EKS, you benefit from EMR's optimized runtimes which speed your analysis and save both time and money.

Q: How can I deploy and manage Amazon EMR?

You can deploy your workloads to EMR using Amazon EC2, Amazon Elastic Kubernetes Service (EKS), or on-premises AWS Outposts. You can run and manage your workloads withthe EMR Console, API, SDK or CLI and orchestrate them using Amazon Managed Workflows for Apache Airflow (MWAA) or AWS Step Functions. For an interactive experience you can use EMR Studio or SageMaker Studio.

Q: How can I get started with Amazon EMR?

To sign up for Amazon EMR, click the "Sign Up Now" button on the Amazon EMR detail page http://aws.amazon.com/emr. You must be signed up for Amazon EC2 and Amazon S3 to access Amazon EMR; if you are not already signed up for these services, you will be prompted to do so during the Amazon EMR sign-up process. After signing up, please refer to the Amazon EMR documentation, which includes our Getting Started Guide – the best place to get going with the service.

Q: How reliable is Amazon EMR?

Developing & debugging

Q: Where can I find code samples?

Check out the sample code in these Articles and Tutorials. If you use EMR Studio, you can explore the features using a set of notebook examples.

Q: How do I develop a data processing application?

You can develop, visualize and debug data science and data engineering applications written in R, Python, Scala, and PySpark in Amazon EMR Studio. You can also develop a data processing job on your desktop, for example, using Eclipse, Spyder, PyCharm, or RStudio, and run it on Amazon EMR. Additionally, you can select JupyterHub or Zeppelin in the software configuration when spinning up a new cluster and develop your application on Amazon EMR using one or more instances.

Q: What is the benefit of using the Command Line Tools or APIs vs. AWS Management Console?

The Command Line Tools or APIs provide the ability to programmatically launch and monitor progress of running clusters, to create additional custom functionality around clusters (such as sequences with multiple processing steps, scheduling, workflow, or monitoring), or to build value-added tools or applications for other Amazon EMR customers. In contrast, the AWS Management Console provides an easy-to-use graphical interface for launching and monitoring your clusters directly from a web browser.

Q: Can I add steps to a cluster that is already running?

Yes. Once the job is running, you can optionally add more steps to it via the AddJobFlowSteps API. The AddJobFlowSteps API will add new steps to the end of the current step sequence. You may want to use this API to implement conditional logic in your cluster or for debugging.

Q: Can I be notified when my cluster is finished?

You can sign up for up Amazon SNS and have the cluster post to your SNS topic when it is finished. You can also view your cluster progress on the AWS Management Console or you can use the Command Line, SDK, or APIs to get a status on the cluster.

Q: Can I terminate my cluster when my steps are finished?

Yes. You can terminate your cluster automatically when all your steps finish by turning the auto-terminate flag on.

Q: What OS versions are supported with Amazon EMR?

Amazon EMR 5.30.0 and later, and the Amazon EMR 6.x series are based on Amazon Linux 2. You can also specify a custom AMI that you create based on the Amazon Linux 2. This allows you to perform sophisticated pre-configuration for virtually any application. For more information, see Using a Custom AMI.

Q: Does Amazon EMR support third-party software packages?

Yes. You can use Bootstrap Actions to install third-party software packages on your cluster. You can also upload statically compiled executables using the Hadoop distributed cache mechanism. EMR 6.x supports Hadoop 3, which allows the YARN NodeManager to launch containers either directly on the EMR cluster host or inside a Docker container. Please see our documentation to learn more.

Q: What tools are available to me for debugging?

There are several tools you can use to gather information about your cluster to help determine what went wrong. If you use Amazon EMR studio, you can launch tools like Spark UI and YARN Timeline Service to simplify debugging. From the Amazon EMR Console, you can get off-cluster access to persistent application user interfaces for Apache Spark, Tez UI and the YARN timeline server, several on-cluster application user interfaces, and a summary view of application history in the Amazon EMR console for all YARN applications. You can also connect to your Master Node Using SSH and view cluster instances via these the web interfaces . For more information, see our documentation.

EMR Studio

Q: What is EMR Studio?

EMR Studio is an integrated development environment (IDE) that makes it easy for data scientists and data engineers to develop, visualize, and debug data engineering and data science applications written in R, Python, Scala, and PySpark.

It is a fully managed application with single sign-on, fully managed Jupyter Notebooks, automated infrastructure provisioning, and the ability to debug jobs without logging into the AWS Console or cluster. Data scientists and analysts can install custom kernels and libraries, collaborate with peers using code repositories like GitHub and BitBucket, or execute parameterized notebooks as part of scheduled workflows using orchestration services like Apache Airflow, AWS Step Functions, and Amazon Managed Workflows for Apache

Airflow. You can read orchestrating analytics jobs on Amazon EMR notesbooks using Amazon MWAA to learn more. EMR Studio kernels and applications run on EMR clusters, so you get the benefit of distributed data processing using the performance optimized Amazon EMR runtime for Apache Spark. Administrators can setup EMR Studio for analysts to run their applications on existing EMR clusters or create new clusters using pre-defined AWS CloudFormation templates for EMR.

Q: What can I do with EMR Studio?

With EMR Studio, you can log in directly to fully managed Jupyter notebooks using your corporate credentials without logging into the AWS console, start notebooks in seconds, get onboarded with sample notebooks, and perform your data exploration. You can also customize your environment by loading custom kernels and python libraries from notebooks. EMR Studio kernels and applications run on EMR clusters, so you get the benefit of distributed data processing using the performance optimized Amazon EMR runtime for Apache Spark. You can collaborate with peers by sharing notebooks via GitHub and other repositories. You can also run Notebooks directly as continuous integration and deployment pipelines. You can pass different parameter values to a notebook. You can also chain notebooks, and integrate notebooks into scheduled workflows using workflow orchestration services like Apache Airflow. Further, you can debug clusters and jobs using as few clicks as possible with native applications interfaces such as the Spark UI and the YARN Timeline service.

Q: How is EMR Studio different from EMR Notebooks?

There are five main differences.

1. There is no need to access AWS Management Console for EMR Studio. EMR Studio is hosted outside of the AWS Management Console. This is useful if you do not provide data scientists or data engineers access to the AWS Management Console.
2. You can use enterprise credentials from your identity provider using AWS IAM Identity Center (successor to AWS SSO) to log in to EMR Studio.
3. EMR Studio brings you a notebook first experience. EMR Studio kernels and applications run on EMR clusters, so you get the benefit of distributed data processing using the performance optimized Amazon EMR runtime for Apache Spark. Running code on a cluster is as simple as attaching the notebook to an existing cluster or provisioning a new one.
4. EMR Studio has a simplified user interface and abstracts hardware configurations. For example, you can setup cluster templates once and use the templates to start new clusters.
5. EMR Studio enables a simplified debugging experience so that you can access the native application user interfaces in one place using as few clicks as possible.

Q: How is EMR Studio different from SageMaker Studio?

You can use both EMR Studio and SageMaker Studio with Amazon EMR. EMR Studio provides an integrated development environment (IDE) that makes it easy for you to develop, visualize, and debug data engineering and data science applications written in R, Python, Scala, and PySpark. Amazon SageMaker Studio provides a single, web-based visual interface where you can perform all machine learning development steps. SageMaker Studio gives you complete access, control, and visibility into each step required to build, train, and deploy models. You can quickly upload data, create new notebooks, train and tune models, move back and forth between steps to adjust experiments, compare results, and deploy models to production all in one place, making you much more productive.

Q: How do I get started with EMR Studio?

Your administrator must first set up an EMR Studio. When you receive a unique sign-on URL for your Amazon EMR Studio from your administrator, you can log in to the Studio directly using your corporate credentials.

Q: Do I need to log in to the AWS Management Console to use EMR Studio?

No. After your administrator sets up an EMR Studio and provides the Studio access URL, your team can log in using corporate credentials. There's no need to log in to the AWS Management Console. In an EMR Studio, your team can perform tasks and access resources configured by your administrator.

Q: What identity providers are supported for the single sign-on experience in EMR Studio?

AWS IAM Identity Center (successor to AWS SSO) is the single sign-on service provider for EMR Studio. The list of identity providers supported by AWS IAM can be found in our documentation.

Q: What is a Workspace in EMR Studio?

Workspaces help you organize Jupyter Notebooks. All notebooks in a Workspace are saved to the same Amazon S3 location and run on the same cluster. You can also link a code repository like a GitHub repository to all notebooks in a workspace. You can create and configure a Workspace before attaching it to a cluster, but you should connect to a cluster before running a notebook.

Q: In EMR Studio, can I create a workspace or open a workspace without a cluster?

Yes, you can create or open a workspace without attaching it to a cluster. Only when you need to execute, you should connect them to a cluster. EMR Studio kernels and applications are executed on EMR clusters, so you get the benefit of distributed data processing using the performance optimized Amazon EMR runtime for Apache Spark.

Q: Can I install custom libraries to use in my notebook code?

All Spark queries run on your EMR cluster, so you need to install all runtime libraries that your Spark application uses on the cluster. You can easily install notebook-scoped libraries within a notebook cell. You can also install Jupyter Notebook kernels and Python libraries on a cluster master node either within a notebook cell or while connected using SSH to the master node of the cluster. For more information, see documentation. Additionally, you can use a bootstrap action or a custom AMI to install required libraries when you create a cluster. For more information, see Create Bootstrap Actions to Install Additional Software and Using a Custom AMI in the Amazon EMR Management Guide.

Q: Where are the notebooks saved?

Workspace together with notebook files in the workspace are saved automatically at regular intervals to the ipynb file format in the Amazon S3 location that you specify when you create the workspace. The notebook file has the same name as your notebook in the Amzon EMR Studio.

Q: How do I use version control with my notebook? Can I use repositories like GitHub?

You can associate Git-based repositories with your Amazon EMR Studio notebooks to save your notebooks in a version controlled environment.

Q: In EMR Studio, what compute resources can I run notebooks on?

With EMR Studio, you can run notebook code on Amazon EMR running on Amazon Elastic Compute Cloud (Amazon EC2) or Amazon EMR on Amazon Elastic Kubernetes Service (Amazon EKS). You can attach notebooks to either existing or new clusters. You can create EMR clusters in two ways in EMR Studio: create a cluster using a pre-configured cluster template via AWS Service Catalog, create a cluster by specifying cluster name, number of instances, and instance type.

Q: Can I re-attach a workspace with a different compute resource in EMR Studio?

Yes, you can open your workspace, choose *EMR Clusters* icon on the left, push *Detach* button, and then select a cluster from the *Select cluster* drop down list, and push *Attach* button.

Q: Where do I find all my workspaces in EMR Studio?

In EMR Studio, you may choose *Workspaces* tab on the left and view all workspaces created by you and other users in the same AWS account.

Q: What are the IAM policies needed to use EMR Studio?

Each EMR studio needs permissions to interoperate with other AWS services. To give your EMR Studios the necessary permissions, your administrators need to create an EMR Studio service role with the provided policies. They also need to specify a user role for EMR Studio that defines Studio-level permissions. When they add users and groups from AWS IAM Identity Center (successor to AWS SSO) to EMR Studio, they can assign a session policy to a user or group to apply fine-grained permission controls. Session policies help administrators refine user permissions without the need to create multiple IAM roles. For more information about session policies, see Policies and Permissions in the AWS Identity and Access Management User Guide.

Q: Is there any limitations on EMR clusters I can attach my workspace to in EMR Studio?

Yes. High Availability (Multi-master) clusters, Kerberized clusters, and AWS Lake Formation clusters are currently not supported.

Q: What is the cost of using Amazon EMR Studio?

Amazon EMR Studio is provided at no additional charge to you. Applicable charges for Amazon Simple Storage Service storage and for Amazon EMR clusters apply when you use EMR Studio.

EMR Notebooks

Q: What are EMR Notebooks?

We recommend that new customers use Amazon EMR Studio, not EMR Notebooks. EMR Notebooks provide a managed environment, based on Jupyter Notebook, that allows data scientists, analysts, and developers to prepare and visualize data, collaborate with peers, build applications, and perform interactive analysis using EMR clusters. Although we recommend that new customers use EMR Studio, EMR Notebooks is supported for compatibility.

Q: What can I do with EMR Notebooks?

You can use EMR Notebooks to build Apache Spark applications and run interactive queries on your EMR cluster with minimal effort. Multiple users can create serverless notebooks directly from the console, attach them to an existing shared EMR cluster, or provision a cluster directly from the console and immediately start experimenting with Spark. You can detach notebooks and re-attach them to new clusters. Notebooks are auto-saved to S3 buckets, and you can retrieve saved notebooks from the console to resume work. EMR Notebooks are prepackaged

with the libraries found in the Anaconda repository, allowing you to import and use these libraries in your notebooks code and use them to manipulate data and visualize results. Further, EMR notebooks have integrated Spark monitoring capabilities that you can use to monitor the progress of your Spark jobs and debug code from within the notebook.

Q: How do I get started with EMR Notebooks?

To get started with EMR Notebooks, open the EMR console and choose Notebooks in the navigation pane. From there, just choose Create Notebook, enter a name for your notebook, choose an EMR cluster or instantly create a new one, provide a service role for the notebook to use, and choose an S3 bucket where you want to save your notebook files and then click on Create Notebook. After the notebook shows a Ready status, choose Open to start the notebook editor.

Q: What EMR release versions are supported with EMR Notebooks?

EMR Notebooks can be attached to EMR clusters running EMR release 5.18.0 or later.

Q: What is the cost of using EMR Notebooks?

EMR notebooks are provided at no additional charge to you. You will be charged as usual for the attached EMR clusters in your account. You and find out more about the pricing for your cluster by visiting https://aws.amazon.com/emr/pricing/

Managing data

Q: How do I get my data into Amazon S3?

Amazon EMR provides several ways to get data onto a cluster. The most common way is to upload the data to Amazon S3 and use the built-in features of Amazon EMR to load the data onto your cluster. You can use the Distributed Cache feature of Hadoop to transfer files from a distributed file system to the local file system. For more details, see documentation. Alternatively, if you are migrating data from on premises to the cloud, you can use one of the Cloud Data Migration services from AWS.

Q: How do I get logs for terminated clusters?

Hadoop system logs as well as user logs will be placed in the Amazon S3 bucket which you specify when creating a cluster. Persistent application UIs are run off-cluster, Spark History Server, Tez UI and YARN timeline servers logs are available for 30 days after an application terminates.

Q: Do you compress logs?

No. At this time Amazon EMR does not compress logs as it moves them to Amazon S3.

Q: Can I load my data from the internet or somewhere other than Amazon S3?

Yes. You can use AWS Direct Connect to establish a private dedicated network connection to AWS. If you have large amounts of data, you can use AWS Import/Export. For more details refer to our documentation.

Security and Data Access Control

Q: How do I prevent other people from viewing my data during cluster execution?

Amazon EMR starts your instances in two Amazon EC2 security groups, one for the master and another for the other cluster nodes. The master security group has a port open for communication with the service. It also has the SSH port open to allow you to SSH into the instances, using the key specified at startup. The other nodes start in a separate security group, which only allows interaction with the master instance. By default both security groups are set up to not allow access from external sources including Amazon EC2 instances belonging to other customers. Since these are security groups within your account, you can reconfigure them using the standard EC2 tools or dashboard. Click here to learn more about EC2 security groups. Additionally, you can configure Amazon EMR block public access in each region that you use to prevent cluster creation if a rule allows public access on any port that you don't add to a list of exceptions.

Q: How secure is my data?

Amazon S3 provides authentication mechanisms to ensure that stored data is secured against unauthorized access. Unless the customer who is uploading the data specifies otherwise, only that customer can access the data. Amazon EMR customers can also choose to send data to Amazon S3 using the HTTPS protocol for secure transmission. In addition, Amazon EMR always uses HTTPS to send data between Amazon S3 and Amazon EC2. For added security, customers may encrypt the input data before they upload it to Amazon S3 (using any common data encryption tool); they then need to add a decryption step to the beginning of their cluster when Amazon EMR fetches the data from Amazon S3.

Q: Can I get a history of all EMR API calls made on my account for security or compliance auditing?

Yes. AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The AWS API call history produced by CloudTrail enables security analysis, resource change tracking, and compliance auditing. Learn more about CloudTrail at the AWS CloudTrail detail page, and turn it on via CloudTrail's AWS Management Console.

Q: How do I control what EMR users can access in Amazon S3?

By default, Amazon EMR application processes use EC2 instance profiles when they call other AWS services. For multi-tenant clusters, Amazon EMR offers three options to manage user access to Amazon S3 data.

1. Integration with AWS Lake Formation allows you to define and manage fine-grained authorization policies in AWS Lake Formation to access databases, tables, and columns in AWS Glue Data Catalog. You can enforce the authorization policies on jobs submitted through Amazon EMR Notebooks and Apache Zeppelin for interactive EMR Spark workloads, and send auditing events to AWS CloudTrail. By enabling this integration, you also enable federated Single Sign-On to EMR Notebooks or Apache Zeppelin from enterprise identity systems compatible with Security Assertion Markup Language (SAML) 2.0.
2. Native integration with Apache Ranger allows you to set up a new or an existing Apache Ranger server to define and manage fine-grained authorization policies for users to access databases, tables, and columns of Amazon S3 data via Hive Metastore. Apache Ranger is an open-source tool to enable, monitor, and manage comprehensive data security across the Hadoop platform.

   This native integration allows you to define three types of authorization policies on the Apache Ranger Policy Admin server. You can set table, column, and row level authorization for Hive, table and column level authorization for Spark, and prefix and object level authorization for Amazon S3. Amazon EMR automatically installs and

configures the corresponding Apache Ranger plugins on the cluster. These Ranger plugins sync up with the Policy Admin server for authorization polices, enforce data access control, and send auditing events to Amazon CloudWatch Logs.

3. Amazon EMR User Role Mapper allows you to leverage AWS IAM permissions to manage accesses to AWS resources. You can create mappings between users (or groups) and custom IAM roles. A user or group can only access the data permitted by the custom IAM role. This feature is currently available through AWS Labs.

Regions and Availability Zones

Q: How does Amazon EMR make use of Availability Zones?

Amazon EMR launches all nodes for a given cluster in the same Amazon EC2 Availability Zone. Running a cluster in the same zone improves performance of the jobs flows. By default, Amazon EMR chooses the Availability Zone with the most available resources in which to run your cluster. However, you can specify another Availability Zone if required. You also have the option to optimize your allocation for lowest-priced on demand instances, optimal spot capacity, or use On-Demand Capacity Reservations.

Q: In what Regions is this Amazon EMR available?

For a list of the supported Amazon EMR AWS regions, please visit the AWS Region Table for all AWS global infrastructure.

Q: Is Amazon EMR supported in AWS Local Zones?

EMR supports launching clusters in the Los Angeles AWS Local Zone. You can use EMR in the US West (Oregon) region to launch clusters into subnets associated with the Los Angeles AWS Local Zone.

Q: Which Region should I select to run my clusters?

When creating a cluster, typically you should select the Region where your data is located.

Q: Can I use EU data in a cluster running in the US region and vice versa?

Yes, you can. If you transfer data from one region to the other you will be charged bandwidth charges.

Q: What is different about the AWS GovCloud (US) region?

The AWS GovCloud (US) region is designed for US government agencies and customers. It adheres to US ITAR requirements. In GovCloud, EMR does not support spot instances or the enable-debugging feature. The EMR Management Console is not yet available in GovCloud.

Deployment options

Amazon EMR on Amazon EC2

Q: What is an Amazon EMR Cluster?

A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is called a node and has a role within the cluster, referred to as the node type. Amazon EMR also installs different software components on each node type, giving each

node a role in a distributed application like Apache Hadoop. Every cluster has a unique identifier that starts with "j-".

Q: What are node types in a cluster?

An Amazon EMR cluster has three types of nodes:

1. Master node: A node that manages the cluster by running software components to coordinate the distribution of data and tasks among other nodes for processing. The master node tracks the status of tasks and monitors the health of the cluster. Every cluster has a master node, and it's possible to create a single-node cluster with only the master node.
2. Core node: A node with software components that run tasks and store data in the Hadoop Distributed File System (HDFS) on your cluster. Multi-node clusters have at least one core node.
3. Task node: A node with software components that only runs tasks and does not store data in HDFS. Task nodes are optional.

Q: What is a cluster step?

A cluster step is a user-defined unit of processing, mapping roughly to one algorithm that manipulates the data. A step is a Hadoop MapReduce application implemented as a Java jar or a streaming program written in Java, Ruby, Perl, Python, PHP, R, or C++. For example, to count the frequency with which words appear in a document, and output them sorted by the count, the first step would be a MapReduce application which counts the occurrences of each word, and the second step would be a MapReduce application which sorts the output from the first step based on the counts.

Q: What are different cluster states?

STARTING – The cluster starts by configuring EC2 instances.
BOOTSTRAPPING – Bootstrap actions are being executed on the cluster.
RUNNING – A step for the cluster is currently being run.
WAITING – The cluster is currently active, but has no steps to run.
TERMINATING - The cluster is in the process of shutting down.
TERMINATED - The cluster was shut down without error.
TERMINATED_WITH_ERRORS - The cluster was shut down with errors.

Q: What are different step states?

PENDING – The step is waiting to be run.
RUNNING – The step is currently running.
COMPLETED – The step completed successfully.
CANCELLED – The step was cancelled before running because an earlier step failed or cluster was terminated before it could run.
FAILED – The step failed while running.

Launching a cluster

Q: How can I launch a cluster?

You can launch a cluster through the AWS Management Console by filling out a simple cluster request form. In the request form, you specify the name of your cluster, the location in Amazon S3 of your input data, your processing application, your desired data output location, and the number and type of Amazon EC2 instances you'd like to use. Optionally, you can specify a location to store your cluster log files and SSH Key to login to your cluster while it is running. Alternatively, you can launch a cluster using the RunJobFlow API or using the 'create' command in the Command Line Tools. For launching a cluster with EMR Studio, refer to the EMR Studio section above..

Q: How can I terminate a cluster?

At any time, you can terminate a cluster via the AWS Management Console by selecting a cluster and clicking the "Terminate" button. Alternatively, you can use the TerminateJobFlows API. If you terminate a running cluster, any results that have not been persisted to Amazon S3 will be lost and all Amazon EC2 instances will be shut down.

Q: Does Amazon EMR support multiple simultaneous cluster?

You can start as many clusters as you like. When you get started, you are limited to 20 instances across all your clusters. If you need more instances, complete the Amazon EC2 instance request form. Once your Amazon EC2 limit is raised, the new limit will be automatically applied to your Amazon EMR clusters.

Managing a cluster

Q: How does Amazon EMR use Amazon EC2 and Amazon S3?

You can upload your input data and a data processing application into Amazon S3. Amazon EMR then launches a number of Amazon EC2 instances that you specified. The service begins the cluster execution while pulling the input data from Amazon S3 using S3 URI scheme into the launched Amazon EC2 instances. Once the cluster is finished, Amazon EMR transfers the output data to Amazon S3, where you can then retrieve it or use as input in another cluster.

Q: How is a computation done in Amazon EMR?

Amazon EMR uses the Hadoop data processing engine to conduct computations implemented in the MapReduce programming model. The customer implements their algorithm in terms of map() and reduce() functions. The service starts a customer-specified number of Amazon EC2 instances, comprised of one master and multiple other nodes. Amazon EMR runs Hadoop software on these instances. The master node divides input data into blocks, and distributes the processing of the blocks to the other nodes. Each node then runs the map function on the data it has been allocated, generating intermediate data. The intermediate data is then sorted and partitioned and sent to processes which apply the reducer function to it locally on the nodes. Finally, the output from the reducer tasks is collected in files. A single "cluster" may involve a sequence of such MapReduce steps.

Q: Which Amazon EC2 instance types does Amazon EMR support?

See the EMR pricing page for details on latest available instance types and pricing per region.

Q: How long will it take to run my cluster?

The time to run your cluster will depend on several factors including the type of your cluster, the amount of input data, and the number and type of Amazon EC2 instances you choose for your cluster.

Q: If the master node in a cluster goes down, can Amazon EMR recover it?

Yes. You can launch an EMR cluster (version 5.23 or later) with three master nodes and support high availability of applications like YARN Resource Manager, HDFS Name Node, Spark, Hive, and Ganglia. Amazon EMR automatically fails over to a standby master node if the primary master node fails or if critical processes, like Resource Manager or Name Node, crash. Since the master node is not a potential single point of failure, you can run your long-lived EMR clusters without interruption. In the event of a failover, Amazon EMR automatically replaces the failed master node with a new master node with the same configuration and boot-strap actions.

Q: If another node goes down in a cluster, can Amazon EMR recover from it?

Yes. Amazon EMR is fault tolerant for node failures and continues job execution if a node goes down. Amazon EMR will also provision a new node when a core node fails. However, Amazon EMR will not replace nodes if all nodes in the cluster are lost.

Q: Can I SSH onto my cluster nodes?

Yes. You can SSH onto your cluster nodes and execute Hadoop commands directly from there. If you need to SSH into a specific node, you have to first SSH to the master node, and then SSH into the desired node.

Q: What is Amazon EMR Bootstrap Actions?

Bootstrap Actions is a feature in Amazon EMR that provides users a way to run custom set-up prior to the execution of their cluster. Bootstrap Actions can be used to install software or configure instances before running your cluster. You can read more about bootstrap actions in EMR's Developer Guide.

Q: How can I use Bootstrap Actions?

You can write a Bootstrap Action script in any language already installed on the cluster instance including Bash, Perl, Python, Ruby, C++, or Java. There are several pre-defined Bootstrap Actions available. Once the script is written, you need to upload it to Amazon S3 and reference its location when you start a cluster. Please refer to the Developer Guide for details on how to use Bootstrap Actions.

Q: How do I configure Hadoop settings for my cluster?

The EMR default Hadoop configuration is appropriate for most workloads. However, based on your cluster's specific memory and processing requirements, it may be appropriate to tune these settings. For example, if your cluster tasks are memory-intensive, you may choose to use fewer tasks per core and reduce your job tracker heap size. For this situation, a pre-defined Bootstrap Action is available to configure your cluster on startup. See the Configure Memory Intensive Bootstrap Action in the Developer's Guide for configuration details and usage instructions. An additional predefined bootstrap action is available that allows you to customize your cluster settings to any value of your choice. See the Configure Hadoop Bootstrap Action in the Developer's Guide for usage instructions.

Q: Can I modify the number of nodes in a running cluster?

Yes. Nodes can be of two types: (1) core nodes, which both host persistent data using Hadoop Distributed File System (HDFS) and run Hadoop tasks and (2) task nodes, which only run Hadoop tasks. While a cluster is running you may increase the number of core nodes and you may either increase or decrease the number of task nodes. This can be done through the API, Java SDK, or though the command line client. Please refer to the Resizing Running clusters section in the Developer's Guide for details on how to modify the size of your running cluster. You can also use EMR Managed Scaling.

Q: When would I want to use core nodes versus task nodes?

As core nodes host persistent data in HDFS and cannot be removed, core nodes should be reserved for the capacity that is required until your cluster completes. As task nodes can be added or removed and do not contain HDFS, they are ideal for capacity that is only needed on a temporary basis. You can launch task instance fleets on Spot Instances to increase capacity while minimizing costs.

Q: Why would I want to modify the number of nodes in my running cluster?

There are several scenarios where you may want to modify the number of nodes in a running cluster. If your cluster is running slower than expected, or timing requirements change, you can increase the number of core nodes to increase cluster performance. If different phases of your cluster have different capacity needs, you can start with a small number of core nodes and increase or decrease the number of task nodes to meet your cluster's varying capacity requirements. You can also used EMR Managed Scaling.

Q: Can I automatically modify the number of nodes between cluster steps?

Yes. You may include a predefined step in your workflow that automatically resizes a cluster between steps that are known to have different capacity needs. As all steps are guaranteed to run sequentially, this allows you to set the number of nodes that will execute a given cluster step.

Q: How can I allow other IAM users to access my cluster?

To create a new cluster that is visible to all IAM users within the EMR CLI: Add the --visible-to-all-users flag when you create the cluster. For example: elastic-mapreduce --create --visible-to-all-users. Within the Management Console, simply select "Visible to all IAM Users" on the Advanced Options pane of the Create cluster Wizard.

To make an existing cluster visible to all IAM users you must use the EMR CLI. Use --set-visible-to-all-users and specify the cluster identifier. For example: elastic-mapreduce --set-visible-to-all-users true --jobflow j-xxxxxxx. This can only be done by the creator of the cluster.

To learn more, see the Configuring User Permissions section of the EMR Developer Guide.

Tagging a cluster

Q: What Amazon EMR resources can I tag?

You can add tags to an active Amazon EMR cluster. An Amazon EMR cluster consists of Amazon EC2 instances, and a tag added to an Amazon EMR cluster will be propagated to each active Amazon EC2 instance in that cluster. You cannot add, edit, or remove tags from terminated clusters or terminated Amazon EC2 instances which were part of an active cluster.

Q: Does Amazon EMR tagging support resource-based permissions with IAM Users?

No, Amazon EMR does not support resource-based permissions by tag. However, it is important to note that propagated tags to Amazon EC2 instances behave as normal Amazon EC2 tags. Therefore, an IAM Policy for Amazon EC2 will act on tags propagated from Amazon EMR if they match conditions in that policy.

Q: How many tags can I add to a resource?

You can add up to ten tags on an Amazon EMR cluster.

Q: Do my Amazon EMR tags on a cluster show up on each Amazon EC2 instance in that cluster? If I remove a tag on my Amazon EMR cluster, will that tag automatically be removed from each associated EC2 instance?

Yes, Amazon EMR propagates the tags added to a cluster to that cluster's underlying EC2 instances. If you add a tag to an Amazon EMR cluster, it will also appear on the related Amazon EC2 instances. Likewise, if you remove a tag from an Amazon EMR cluster, it will also be removed from its associated Amazon EC2 instances. However, if you are using IAM policies for Amazon EC2 and plan to use Amazon EMR's tagging functionality, you should make sure that permission to use the Amazon EC2 tagging APIs CreateTags and DeleteTags is granted.

Q: How do I get my tags to show up in my billing statement to segment costs?

Select the tags you would like to use in your AWS billing report here. Then, to see the cost of your combined resources, you can organize your billing information based on resources that have the same tag key values.

Q: How do I tell which Amazon EC2 instances are part of an Amazon EMR cluster?

An Amazon EC2 instance associated with an Amazon EMR cluster will have two system tags:

- aws:elasticmapreduce:instance-group-role=CORE
    - Key = instance-group role ; Value = [CORE or TASK];
- aws:elasticmapreduce:job-flow-id=j-12345678
    - Key = job-flow-id ; Value = [JobFlowID]

Q: Can I edit tags directly on the Amazon EC2 instances?

Yes, you can add or remove tags directly on Amazon EC2 instances that are part of an Amazon EMR cluster. However, we do not recommend doing this, because Amazon EMR’s tagging system will not sync the changes you make to an associated Amazon EC2 instance directly. We recommend that tags for Amazon EMR clusters be added and removed from the Amazon EMR console, CLI, or API to ensure that the cluster and its associated Amazon EC2 instances have the correct tags.

EMR Serverless

General

Q: What is Amazon EMR Serverless?

Amazon EMR Serverless is a new deployment option in Amazon EMR that allows you to run big data frameworks such as Apache Spark and Apache Hive without configuring, managing, and scaling clusters.

Q: Who can use EMR Serverless?

Data engineers, analysts, and scientists can use EMR Serverless to build applications using open-source frameworks such as Apache Spark and Apache Hive. They can use these frameworks to transform data, run interactive SQL queries, and machine learning workloads.

Q: How do I get started with EMR Serverless?

You can use EMR Studio, AWS CLI, or APIs to submit jobs, track job status, and build your data pipelines to run on EMR Serverless. To get started with EMR Studio, sign into the AWS Management Console, navigate to Amazon EMR under the Analytics category, and select Amazon EMR Serverless. Follow the directions in the AWS Management Console, navigate to Amazon EMR under the Analytics category, and select Amazon EMR Serverless. Follow the directions in the Getting started guide to create your EMR Serverless application and submit jobs. You can refer to the Interacting with your application on the AWS CLI page to launch your applications and submit jobs using CLI. You can also find EMR Serverless examples and sample code in our GitHub repository.

Q: What open-source frameworks does EMR Serverless support?

EMR Serverless currently supports Apache Spark and Apache Hive engines. If you want support for additional frameworks such as Apache Presto or Apache Flink, please send a request to emr-feedback@amazon.com.

Q: In what Regions is EMR Serverless available?

EMR Serverless is available in the following AWS Regions: Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Paris), Europe (Stockholm), South America (São Paulo), US East (N. Virginia), US East (Ohio), US West (N. California), and US West (Oregon).

Q: What is the difference between Amazon EMR Serverless, Amazon EMR on EC2, Amazon EMR on AWS Outposts, and Amazon EMR on EKS?

Amazon EMR provides the option to run applications on EC2 based clusters, EKS clusters, Outposts, or Serverless. EMR on EC2 clusters are suitable for customers who need maximum control and flexibility over running their application. With EMR on EC2 clusters, customers can choose the EC2 instance type to meet application-specific performance needs, customize the Linux AMI, customize the EC2 instance configuration, customize and extend open-source frameworks, and install additional custom software on cluster instances. Amazon EMR on EKS is suitable for customers who want to standardize on EKS to manage clusters across applications or use different versions of an open-source framework on the same cluster. Amazon EMR on AWS Outposts is for customers who want to run EMR closer to their data center, within an Outpost. EMR Serverless is suitable for customers who want to avoid managing and operating clusters and prefer to run applications using open-source frameworks.
Q: What are some of the feature differences between EMR Serverless and Amazon EMR on EC2?

Feature	EMR Serverless	Amazon EMR on EC2	Amazon EMR on EKS
Support for latest open-source versions	Y	Y	Y

Resilience to Availability Zone failures	Y	N	Y
Automatically scale resources up and down as needed	Y	Y	Y
Encryption for data at rest	Y	Y	Y
Open-source frameworks supported	Spark and Hive	multiple	Spark
Support for fine-grained authorization using AWS Lake Formation	N	Y	N
Support for Apache Hudi and Apache Iceberg	Y	Y	Y
Integration with Apache Ranger for table- and column-level permission control	N	Y	N
Customize operating system images	N	Y	Y
Customize open-source framework installed	Y	Y	Y
Customize and load additional libraries and dependencies	Y	Y	Y

Run workloads from SageMaker Studio as part of machine learning (ML) workflow	N	Y	N
Connect to self-hosted Jupyter Notebooks	N	Y	Y
Build and orchestrate pipelines using Apache Airflow and Amazon Managed Workflows for Apache Airflow (MWAA)	Y	Y	Y
Build and orchestrate pipelines using AWS Step Functions	N	Y	Y

Q: What EMR releases are supported in EMR Serverless?

EMR Serverless supports EMR release labels 6.6 and above. With EMR Serverless, you get the same performance-optimized EMR runtime available in other EMR deployment options, which is 100% API-compatible with standard open-source frameworks.

Applications, workers, and jobs

Q: What is an application and how can I create it?

With Amazon EMR Serverless, you can create one or more EMR Serverless applications that use open-source analytics frameworks. To create an application, you must specify the following attributes: 1) the Amazon EMR release version for the open-source framework version you want to use and 2) the specific analytics engines that you want your application to use, such as Apache Spark 3.1 or Apache Hive 3.0. After you create an application, you can start running your data processing jobs or interactive requests to your application.

Q: What is a worker?

An EMR Serverless application internally uses workers to execute your workloads. When a job is submitted, EMR Serverless computes the resources needed for the job and schedules workers. EMR Serverless breaks down your workloads into tasks, provisions and sets up workers with the open-source framework, and decommissions them when the job completes. EMR Serverless automatically scales workers up or down depending on the workload and parallelism required at every stage of the job, thereby removing the need for you to estimate the number of

workers required to run your workloads. The default size of these workers is based on your application type and Amazon EMR release version. You can override these sizes when scheduling a job run.

Q: Can I specify the minimum and maximum number of workers that my jobs can use?

With EMR Serverless, you can specify the minimum and maximum number of concurrent workers and the vCPU and memory configuration for workers. You can also set the maximum capacity limits on the application's resources to control costs.

Q: When should I create multiple applications?

Consider creating multiple applications when doing any of the following:

1. Using different open-source frameworks
2. Using different versions of open-source frameworks for different use cases
3. Performing A/B testing when upgrading from one version to another
4. Maintaining separate logical environments for test and production scenarios
5. Providing separate logical environments for different teams with independent cost controls and usage tracking
6. Separating different lines-of-business applications

Q: Can I change default properties of an EMR Serverless application after it is created?

Yes, you can modify application properties such as initial capacity, maximum capacity limits, and network configuration using EMR Studio or the update-application API/CLI call.

Q: When should I create an application with a pre-initialized pool of workers?

An EMR Serverless application without pre-initialized workers takes up to 120 seconds to determine the required resources and provision them. EMR Serverless provides an optional feature that keeps workers initialized and ready to respond in seconds, effectively creating an on-call pool of workers for an application. This feature is called pre-initialized capacity and can be configured for each application by setting the initial-capacity parameter of an application.

Pre-initialized capacity allows jobs to start immediately, making it ideal for implementing time-sensitive jobs. You can specify the number of workers that you want to pre-initialize when you start an EMR Serverless application. Subsequently, when users submit jobs, the pre-initialized workers can be used to immediately start the jobs. If the job requires more workers than what you have chosen to pre-initialize, EMR Serverless automatically adds more workers (up to the maximum concurrent limit that you specify). After the job finishes, EMR Serverless automatically reverts back to maintaining the pre-initialized workers that you specified. Workers automatically shut down if they have been idle for 15 minutes. You can change the default idle timeout for your application using the updateApplication API or EMR Studio.

Q: How do I submit and manage jobs on EMR Serverless?

You can submit and manage EMR Serverless jobs using EMR Studio, SDK/CLI, or our Apache Airflow connectors.

Q: How can I include dependencies with jobs that I want to run on EMR Serverless?

For PySpark, you can package your Python dependencies using virtualenv and pass the archive file using the --archives option, which enables your workers to use the dependencies during the job run. For Scala or Java, you can package your dependencies as jars, upload them to Amazon S3, and pass them using the --jars or --packages options with your EMR Serverless job run.

Q: Do EMR Serverless Spark and Hive applications support user-defined functions (UDFs)?

EMR Serverless supports Java-based UDFs. You can package them as jars, upload them to S3, and use them in your Spark or HiveQL scripts.

Q: What worker configurations does EMR Serverless support?

Refer to the Supported Worker Configuration for details.

Q: Can I cancel an EMR Serverless job in case it is running longer than expected?

Yes, you can cancel a running EMR Serverless job from EMR Studio or by calling the cancelJobRun API/CLI.

Q: Can I add extra storage to the workers?

EMR Serverless comes with 20 GB of ephemeral storage on each worker. If you need more storage, you can customize this during job submission from 20 GB up to 200 GB.

Q: Where can I find code samples?

You can find EMR Serverless code samples in our GitHub repository.

Monitoring and debugging

Q: How do I monitor Amazon EMR Serverless applications and job runs?

Amazon EMR Serverless application- and job-level metrics are published every 30 seconds to Amazon CloudWatch.

Q: How do I launch Spark UI and Tez UI with EMR Serverless?

From EMR Studio, you can select a running or completed EMR Serverless job and then click on the Spark UI or Tez UI button to launch them.

Security and data control

Q: Can I access resources in my Amazon Virtual Private Cloud (VPC)?

Yes, you can configure Amazon EMR Serverless applications to access resources in your own VPC. See the Configuring VPC access section in the documentation to learn more.

Q: What kind of isolation can I get with an EMR Serverless application?

Each EMR Serverless application is isolated from other applications and runs on a secure Amazon VPC.

Amazon EMR on Amazon EKS

Q: What is Amazon EMR on Amazon EKS?
Amazon EMR on Amazon EKS is a deployment model of Amazon EMR that enables customers

to easily and cost-effectively process vast amounts of data. It utilizes hosted analytics frameworks running on the flexible Amazon EKS managed service in containers, with the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2), AWS Fargate, and Amazon Simple Storage Service (Amazon S3).

Q: Why should I use Amazon EMR on Amazon EKS?
Amazon EMR on Amazon EKS decouples the analytics job from the services and infrastructure that are processing the job by using a container-based approach. You can focus more on developing your application and less on operating the infrastructure as EMR on EKS dynamically configures the infrastructure based on the compute, memory, and application dependencies of the job. Infrastructure teams can centrally manage a common compute platform to consolidate EMR workloads with other container-based applications. Multiple teams, organizations, or business units can simultaneously and independently run their analytics processes on the shared infrastructure while maintaining isolation enabled by Amazon EKS and AWS Identity and Access Management (IAM).

Q: What are the benefits for users already running Apache Spark on Amazon EKS?
If you already run Apache Spark on Amazon EKS, you can get all of the benefits of Amazon EMR like automatic provisioning and scaling and the ability to use the latest fully managed versions of open source big data analytics frameworks. You get an optimized EMR runtime for Apache Spark with 3X faster performance than open source Apache Spark on EKS, a serverless data science experience with EMR Studio and Apache Spark UI, fine grained data access control, and support for data encryption.

Q: How does this feature relate to and work with other AWS services?
Amazon EKS provides customers with a managed experience for running Kubernetes on AWS, enabling you to add compute capacity using EKS Managed Node Groups or using AWS Fargate. Running EMR jobs on EKS can access their data on Amazon S3 while monitoring and logging can be integrated with Amazon CloudWatch. AWS Identity and Access Management (IAM) enables role based access control for both jobs and to dependent AWS services.

Q: How does Amazon EMR on Amazon EKS work?
Register your EKS cluster with Amazon EMR. Then, submit your Spark jobs to EMR using the CLI, SDK or EMR Studio. EMR requests the Kubernetes scheduler on EKS to schedule Pods. For each job that you run, EMR on EKS creates a container. The container contains Amazon Linux 2 base image with security updates, plus Apache Spark and associated dependencies to run Spark, plus your application-specific dependencies. Each Job runs in a pod. The Pod downloads this container and starts to execute it. If the container's image has been previously deployed to the node, then a cached image is used and the download is bypassed. Sidecar containers, such as log or metric forwarders, can be deployed to the pod. The Pod terminates after the job terminates. After the job terminates, you can still debug it using Spark UI.

Q: What AWS compute services can I use with Amazon EMR on EKS?
You can use Amazon EMR for EKS with both Amazon Elastic Compute Cloud (EC2) instances to support broader customization options, or the serverless AWS Fargate service to process your analytics without having to provision or manage EC2 instances. Application availability can automatically improve by spreading your analytics jobs across multiple AWS Availability Zones (AZs).

Q: How do I get started with EMR on EKS?
To get started, register your Amazon EKS cluster with Amazon EMR. After registration, reference this registration in your job definition (that includes application dependencies and framework parameters) by submitting your workloads to EMR for execution. With EMR on EKS, you can use different open source big data analytics frameworks, versions, and configurations for analytics applications running on the same EKS cluster. For more information, refer to our documentation.

Q: How do I submit analytics applications to EMR on EKS?
You submit analytics applications using AWS SDK / CLI, Amazon EMR Studio notebooks, and workflow orchestration services like Apache Airflow and Amazon Managed Workflows for Apache Airflow. AWS EMR on EKS's airflow plugin can be downloaded from S3. To install the emr-containers plugin for Amazon Managed Workflows for Apache Airflow, see our documentation.

Q: Can I use the same EMR release for EMR clusters and applications running on EKS?

Yes, you can use the same EMR release for applications that run on EMR clusters and applications that run on EKS.

Q: How do I troubleshoot analytics applications?
You can use the Amazon EMR Spark UI to diagnose and troubleshoot Spark applications. For all analytics applications, EMR provides access to application details, associated logs, and metrics for up to 30 days after they have completed. Jobs can be individually configured to send logs to an Amazon S3 location or Amazon CloudWatch.

Q: Can I see EMR applications in EKS?
Yes, EMR applications show up in the EKS console as Kubernetes jobs and deployments.

Q: Can I isolate multiple jobs or applications from each other on the same EKS cluster?
Yes, Kubernetes natively provides job isolation. Additionally, each job can be configured to run with its own execution-role to limit which AWS resources the job can access.

Q: How does EMR on EKS help reduce costs?
EMR on EKS reduces cost by removing the need to run dedicated clusters. You can use a common, shared EKS cluster to run analytics applications that require different versions of open source big data analytics frameworks. You can also use the same EKS cluster to run your other containerized non-EMR applications.

Q: How do you charge for EMR on EKS?
Amazon EMR on EKS pricing is calculated based on the vCPU and memory resources requested for the pod(s) that are running your job at per minute granularity.

Q: What are some of the differences between EMR on EKS and EMR on EC2?

Feature	EMR on EKS	EMR on EC2
Latest supported version of EMR	Y	Y
Multi-AZ Support for Jobs	Y	N
Multi-Tenant with non-big data workloads	Y	N
EMR version scope	job	cluster
Auto-Scaling Cluster	Y	Y
Managed Scaling	N	Y
Compute Providers	EC2, Fargate	EC2

Data Encryption	Y	Y
Kerberos Authentication	N	Y
Hosted Applications	Spark only	Multiple
AWS Lake Formation	N	Y
Apache Ranger Integration	N	Y
Custom AMI / Images	Y	Y
Integration with Sagemaker & Zeppelin	Y with Livy	Y
Self-Hosted Notebooks	N	Y
Integration with EMR Studio	Y	Y
Zeppelin, JEG	N	Y
Orchestration with Apache Airflow	Y	Y
Orchestration with AWS Stepfunctions	Y	Y

Q: What are Pod Templates?

EMR on EKS enables you to use Kubernetes Pod Templates to customize where and how your job runs in the Kubernetes cluster. Kubernetes Pod Templates provide a reusable design pattern or boilerplate for declaratively expressing how a Kubernetes pod should be deployed to your EKS cluster.

Q: Why should I use Pod Templates with my EMR on EKS job?

Pod Templates can provide more control over how your jobs are scheduled in Kubernetes. For example, you can reduce cost by running Spark driver tasks on Amazon EC2 Spot instances or only allowing jobs requiring SSDs to run on SSD enabled instances. Pod Templates with EMR on EKS to enables fine-grained control of how resources are allocated and running custom containers alongside your job. Therefore, resulting in reduced cost and increased performance of your jobs.

Q: What is a Pod?

Pods are one or more containers, with shared network and storage resources, that run on a Kubernetes worker node. EMR on EKS uses pods to run your job by scheduling Spark driver and executor tasks as individual pods.

Q: What are some use-cases for Pod Templates?

You can optimize both performance and cost by using Pod Templates. For example, you can save cost by defining jobs to run on EC2 Spot instances or increase performance by scheduling them on GPU or SSD-backed EC2 instances. Customers often need fine-grained workload

control in order to support multiple teams or organizations on EKS, and Pod Templates simplify running jobs on team designated node groups. In addition, you can deploy sidecar containers to run initialization code for your job or run common monitoring tools like Fluentd for log forwarding.

Q: Can I specify a different Pod Template for my Spark drivers and Spark executors?

You can, but is not required, to provide individual templates for drivers and executors. For example, you can configure nodeSelectors and tolerations to designate Spark drivers to run only on AWS EC2 On-Demand instances and Spark executors to run only on AWS Fargate instances. In your job submission, configure the spark properties spark.kubernetes.driver.podTemplateFile and spark.kubernetes.executor.podTemplateFile to reference the template's S3 location.

Q: What template values can I specify?

You can specify both Pod Level Fields (including Volumes, Pod Affinity, Init Containers, Node Selector) and Spark Main Container level fields (including EnvFrom, Working Directory, Lifecycle, Container Volume Mounts). The full list of allowed values is provided in our documentation.

Q: Where can I find more information about Pod Templates?

Amazon EKS already supports Pod Templates and for more information about Amazon EMR on EKS' support for Pod Templates, refer to our documentation and the Apache Spark Pod Template documentation.

Q: Why should I use Custom Images with EMR on EKS?

Without Custom Images, managing application dependencies with EMR on EKS required you to reference them at runtime from an external storage service such as Amazon S3. Now, with custom image support, you can create a self-contained docker image with the application and its' dependent libraries. You no longer need to maintain, update or version externally stored libraries and your big data applications can be developed using the same DevOps processes that your other containerized applications are using. Just point at your image and run it.

Q: What is a custom image?

A Custom Image is an EMR on EKS provided docker image ("base image") that contains the EMR runtime and connectors to other AWS services that you modify to include application dependencies or additional packages that your application requires. The new image can be stored in either Amazon Elastic Container Registry (ECR) or your own Docker container registry.

Q: What are some use-cases for Custom Images?

Customers can create a base image, add their corporate standard libraries, and then store it in Amazon Elastic Container Registry (Amazon ECR). Other customers can customize the image to include their application specific dependencies. The resulting immutable image can be vulnerability scanned, deployed to test and production environments. Examples of dependencies you can add include Java SDK, Python, or R libraries, you can add them to the image directly, just as with other containerized applications.

Q: What is included in the base image?

The same performance optimized Spark Runtime and components are included in the base image as you get when you submit a job without a custom image.

Q: When should I specify a different Custom Image for my Spark drivers and Spark executors?

You can specify separate images for your Spark drivers and executors when you want include different dependencies or libraries. Removing libraries that are not required in both images can result in a smaller image size and thus reduce job start time. You can specify a single image for both drivers and executors (spark.kubernetes.container.image) or specify a different image for drivers (spark.kubernetes.driver.container.image) and executors (spark.kubernetes.executor.container.image).

Q: Where can I find more information about Custom Images?

For more information about Amazon EMR on EKS' support for Custom Images, refer to our documentation or the Apache Spark Documentation.

Q: Is there an additional charge for Custom Images?

There is no charge for using the Custom Images feature.

Amazon EMR on AWS Outposts

Q: What is AWS Outposts?

AWS Outposts brings native AWS services, infrastructure, and operating models to virtually any data center, co-location space, or on-premises facility. Using EMR on Outposts, you can deploy, manage, and scale EMR clusters on-premises, just as you would in the cloud.

Q: When should I use EMR on Outposts?

If you have existing on-premises Apache Hadoop deployments and are struggling to meet capacity demands during peak utilization, you can use EMR on Outposts to augment your processing capacity without having to move data to the cloud. EMR on Outposts enables you to launch a new EMR cluster on-premises in minutes and connect to existing datasets in on-premises HDFS storage to meet this demand and maintain SLAs.

If you need to process data that needs to remain on-premises for governance, compliance, or other reasons, you can use EMR on Outposts to deploy and run applications like Apache Hadoop and Apache Spark on-premises, close to your data. This reduces the need to move large amounts of on-premises data to the cloud, reducing the overall time needed to process that data.

If you're in the process of migrating data and Apache Hadoop workloads to the cloud and want to start using EMR before your migration is complete, you can use AWS Outposts to launch EMR clusters that connect to your existing on-premises HDFS storage. You can then gradually migrate your data to Amazon S3 as part of an evolution to a cloud architecture.

Q: What EMR versions are supported with EMR on Outposts?

The minimum supported Amazon EMR release is 5.28.0.

Q: What EMR applications are available when using Outposts?

All applications in EMR release 5.28.0 and above are supported. See our release notes for a full list of EMR applications.

Q: What EMR features are not supported with EMR on Outposts?

- EC2 Spot instances are not available in AWS Outposts. When creating a cluster, you must choose EC2 On-Demand instances.
- A subset of EC2 instance types are available in AWS Outposts. For a list of supported instance types with EMR and Outposts, please see our documentation.
- When adding Amazon EBS volumes to instances, only the General Purpose SSD (GP2) storage type is supported in AWS Outposts.

Q: Can I use EMR clusters in an Outpost to read data from my existing on-premises Apache Hadoop clusters?

Workloads running on EMR in an Outpost can read and write data in existing HDFS storage, allowing you to easily integrate with existing on-premises Apache Hadoop deployments. This gives you the ability to augment your data processing needs using EMR without the need to migrate data.

Q: Can I choose where to store my data?

When an EMR cluster is launched in an Outpost, all of the compute and data storage resources are deployed in your Outpost. Data written locally to the EMR cluster is stored on local EBS volumes in your Outpost. Tools such as Apache Hive, Apache Spark, Presto, and other EMR applications can each be configured to write data locally in an Outpost, to external file system such as an existing HDFS installation, or to Amazon S3. Using EMR on Outposts, you have full control over storing your data in Amazon S3 or locally in your Outpost.

Q: Do any EMR features need uploaded data to S3?

When launching an EMR cluster in an Outpost, you have the option to enable logging. When logging is enabled, cluster logs will be uploaded to the S3 bucket that you specify. These logs are used to simplify debugging clusters after they have been terminated. When disabled, no logs will be uploaded to S3.

Q: What happens if my Outpost is out of capacity?

When launching a cluster in an Outpost, EMR will attempt to launch the number and type of EC2 On-Demand instances you’ve requested. If there is no capacity available on the Outpost, EMR will receive an insufficient capacity notice. EMR will retry for a period of time, and if no capacity becomes available, the cluster will fail to launch. The same process applies when resizing a cluster. If there is insufficient capacity on the Outpost for the requested instance types, EMR will be unable to scale up the cluster. You can easily set up Amazon CloudWatch alerts to monitor your capacity utilization on Outposts and receive alerts when instance capacity is lower than a desired threshold.

Q: What happens if network connectivity is interrupted between my Outpost and AWS?

If network connectivity between your Outpost and its AWS Region is lost, your clusters in Outposts will continue to run, but there will be actions you will be unable to take until connectivity is restored. Until connectivity is restored, you cannot create new clusters or take new actions on existing clusters. In case of instance failures, the instance will not be automatically replaced. Also, actions such as adding steps to a running cluster, checking step execution status, and sending CloudWatch metrics and events will be delayed until connectivity is restored.

We recommend that you provide reliable and highly available network connectivity between your Outpost and the AWS Region. If network connectivity between your Outpost and its AWS Region is lost for more than a few hours, clusters that have terminate protection enabled will

continue to run, and clusters that have terminate protection disabled may be terminated. If network connectivity will be impacted due to routine maintenance, we recommend proactively enabling terminate protection.

Using EBS volumes

Q: What can I do now that I could not do before?

Most EC2 instances have fixed storage capacity attached to an instance, known as an "instance store". You can now add EBS volumes to the instances in your Amazon EMR cluster, allowing you to customize the storage on an instance. The feature also allows you to run Amazon EMR clusters on EBS-Only instance families such as the M4 and C4.

Q: What are the benefits of adding EBS volumes to an instance running on Amazon EMR?

You will benefit by adding EBS volumes to an instance in the following scenarios:

1. Your processing requirements are such that you need a large amount of HDFS (or local) storage that what is available today on an instance. With support for EBS volumes, you will be able to customize the storage capacity on an instance relative to the compute capacity that the instance provides. Optimizing the storage on an instance will allow you to save costs.
2. You are running on an older generation instance family (such as the M1 and M2 family) and want to move to latest generation instance family but are constrained by the storage available per node on the next generation instance types. Now you can use any of the new generation instance type and add EBS volumes to optimize the storage. Internal benchmarks indicate that you can save cost and improve performance by moving from an older generation instance family (M1 or M2) to a new generation one (M4, C4 & R3). The Amazon EMR team recommends that you run your application to arrive at the right conclusion.
3. You want to use or migrate to the next-generation EBS-Only M4 and C4 family.

Q: Can I persist my data on an EBS volume after a cluster is terminated?

Currently, Amazon EMR will delete volumes once the cluster is terminated. If you want to persist data outside the lifecycle of a cluster, consider using Amazon S3 as your data store.

Q: What kind of EBS volumes can I attach to an instance?

Amazon EMR allows you to use different EBS Volume Types: General Purpose SSD (GP2), Magnetic and Provisioned IOPS (SSD).

Q: What happens to the EBS volumes once I terminate my cluster?

Amazon EMR will delete the volumes once the EMR cluster is terminated.

Q: Can I use an EBS with instances that already have an instance store?

Yes, you can add EBS volumes to instances that have an instance store.

Q: Can I attach an EBS volume to a running cluster?

No, currently you can only add EBS volumes when launching a cluster.

Q: Can I snapshot volumes from a cluster?

The EBS API allows you to Snapshot a cluster. However, Amazon EMR currently does not allow you to restore from a snapshot.

Q: Can I use encrypted EBS volumes?

You can encrypt EBS root device and storage volumes using AWS KMS as your key provider. For more information, see Local Disk Encryption.

Q: What happens when I remove an attached volume from a running cluster?

Removing an attached volume from a running cluster will be treated as a node failure. Amazon EMR will replace the node and the EBS volume with each of the same.

EMR workloads

Q: What is Apache Spark?

Apache Spark™ is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. Amazon EMR is the best place to deploy Apache Spark in the cloud, because it combines the integration and testing rigor of commercial Spark distributions with the scale, simplicity, and cost effectiveness of the cloud. It allows you to launch Spark clusters in minutes without needing to do node provisioning, cluster setup, Spark configuration, or cluster tuning. EMR features Amazon EMR runtime for Apache Spark, a performance-optimized runtime environment for Apache Spark that is active by default on Amazon EMR clusters. Amazon EMR runtime for Apache Spark can be over 3x faster than clusters without the EMR runtime, and has 100% API compatibility with standard Apache Spark. Learn more about Spark and Spark on Amazon EMR.

Q: What is Presto?

Presto is an open source, distributed SQL query engine, designed from the ground up for fast analytic queries against data of any size. With Amazon EMR, you can launch Presto clusters in minutes without needing to do node provisioning, cluster setup, Presto configuration, or cluster tuning. EMR enables you to provision one, hundreds, or thousands of compute instances in minutes. Presto has two community projects –PrestoDB and PrestoSQL. Amazon EMR supports both projects. Learn more about Presto and Presto on Amazon EMR.

Using Hive

Q: What is Apache Hive?

Hive is an open source data warehouse and analytics package that runs on top of Hadoop. Hive is operated by a SQL-based language called Hive QL that allows users to structure, summarize, and query data sources stored in Amazon S3. Hive QL goes beyond standard SQL, adding first-class support for map/reduce functions and complex extensible user-defined data types like Json and Thrift. This capability allows processing of complex and even unstructured data sources such as text documents and log files. Hive allows user extensions via user-defined functions written in Java and deployed via storage in Amazon S3. You can learn more about Apache Hive here.

Q: What can I do with Hive running on Amazon EMR?

Using Hive with Amazon EMR, you can implement sophisticated data-processing applications with a familiar SQL-like language and easy to use tools available with Amazon EMR. With

Amazon EMR, you can turn your Hive applications into a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence tasks.

Q: How is Hive different than traditional RDBMS systems?

Traditional RDBMS systems provide transaction semantics and ACID properties. They also allow tables to be indexed and cached so that small amounts of data can be retrieved very quickly. They provide for fast update of small amounts of data and for enforcement of referential integrity constraints. Typically they run on a single large machine and do not provide support for executing map and reduce functions on the table, nor do they typically support acting over complex user defined data types.

In contrast, Hive executes SQL-like queries using MapReduce. Consequently, it is optimized for doing full table scans while running on a cluster of machines and is therefore able to process very large amounts of data. Hive provides partitioned tables, which allow it to scan a partition of a table rather than the whole table if that is appropriate for the query it is executing.

Traditional RDMS systems are best for when transactional semantics and referential integrity are required and frequent small updates are performed. Hive is best for offline reporting, transformation, and analysis of large data sets; for example, performing click stream analysis of a large website or collection of websites.

One of the common practices is to export data from RDBMS systems into Amazon S3 where offline analysis can be performed using Amazon EMR clusters running Hive.

Q: How can I get started with Hive running on Amazon EMR?

The best place to start is to review our written documentation located here.

Q: Are there new features in Hive specific to Amazon EMR?

Yes. Refer to our documentation for further details:

- You can launch an EMR cluster with multiple master nodes to support high availability for Apache Hive. Amazon EMR automatically fails over to a standby master node if the primary master node fails or if critical processes, like Resource Manager or Name Node, crash. This means that you can run Apache Hive on EMR clusters without interruption.
- Amazon EMR allows you to define EMR Managed Scaling for Apache Hive clusters to help you optimize your resource usage. With EMR Managed Scaling, you can automatically resize your cluster for best performance at the lowest possible cost. With EMR Managed Scaling you specify the minimum and maximum compute limits for your clusters and Amazon EMR automatically resizes them for best performance and resource utilization. EMR Managed Scaling continuously samples key metrics associated with the workloads running on clusters.
- Amazon EMR 6.0.0 adds support for Hive LLAP, providing an average performance speedup of 2x over EMR 5.29. You can learn more here.
- You can now use S3 Select with Hive on Amazon EMR to improve performance. S3 Select allows applications to retrieve only a subset of data from an object, which reduces the amount of data transferred between Amazon EMR and Amazon S3.
- Amazon EMR also enables fast performance on complex Apache Hive queries. EMR uses Apache Tez by default, which is significantly faster than Apache MapReduce. Apache MapReduce uses multiple phases, so a complex Apache Hive query would get broken down into four or five jobs. Apache Tez is designed for more complex queries,

so that same job on Apache Tez would run in one job, making it significantly faster than Apache MapReduce.

- With Amazon EMR, you have the option to leave the metastore as local or externalize it. EMR provides integration with the AWS Glue Data Catalog, Amazon Aurora, Amazon RDS and AWS Lake Formation. Amazon EMR can pull information directly from Glue or Lake Formation to populate the metastore.
- You can load table partitions automatically from Amazon S3. Previously, to import a partitioned table you needed a separate alter table statement for each individual partition in the table. Amazon EMR a now includes a new statement type for the Hive language: “alter table recover partitions.” This statement allows you to easily import tables concurrently into many clusters without having to maintain a shared meta-data store. Use this functionality to read from tables into which external processes are depositing data, for example log files.
- You can write data directly to Amazon S3. When writing data to tables in Amazon S3, the version of Hive installed in Amazon EMR writes directly to Amazon S3 without the use of temporary files. This produces a significant performance improvement but it means that HDFS and S3 from a Hive perspective behave differently. You cannot read and write within the same statement to the same table if that table is located in Amazon S3. If you want to update a table located in S3, then create a temporary table in the cluster’s local HDFS filesystem, write the results to that table, and then copy them to Amazon S3.
- You can access resources located in Amazon S3. The version of Hive installed in Amazon EMR allows you to reference resources such as scripts for custom map and reduce operations or additional libraries located in Amazon S3 directly from within your Hive script (e.g., add jar s3://elasticmapreduce/samples/hive-ads/libs/jsonserde.jar).

Q: What types of Hive clusters are supported?

There are two types of clusters supported with Hive: interactive and batch. In an interactive mode a customer can start a cluster and run Hive scripts interactively directly on the master node. Typically, this mode is used to do ad hoc data analyses and for application development. In batch mode, the Hive script is stored in Amazon S3 and is referenced at the start of the cluster. Typically, batch mode is used for repeatable runs such as report generation.

Q: How can I launch a Hive cluster?

Both batch and interactive clusters can be started from AWS Management Console, EMR command line client, or APIs. Please refer to the Hive section in the Release Guide for more details on launching a Hive cluster.

Q: When should I use Hive vs. PIG?

Hive and PIG both provide high level data-processing languages with support for complex data types for operating on large datasets. The Hive language is a variant of SQL and so is more accessible to people already familiar with SQL and relational databases. Hive has support for partitioned tables which allow Amazon EMR clusters to pull down only the table partition relevant to the query being executed rather than doing a full table scan. Both PIG and Hive have query plan optimization. PIG is able to optimize across an entire scripts while Hive queries are optimized at the statement level.

Ultimately the choice of whether to use Hive or PIG will depend on the exact requirements of the application domain and the preferences of the implementers and those writing queries.

Q: What version of Hive does Amazon EMR support?

For latest version of Hive on Amazon EMR, please refer to documentation.

Q: Can I write to a table from two clusters concurrently

No. Hive does not support concurrently writing to tables. You should avoid concurrently writing to the same table or reading from a table while you are writing to it. Hive has non-deterministic behavior when reading and writing at the same time or writing and writing at the same time.

Q: Can I share data between clusters?

Yes. You can read data in Amazon S3 within a Hive script by having 'create external table' statements at the top of your script. You need a create table statement for each external resource that you access.

Q: Should I run one large cluster, and share it amongst many users or many smaller clusters?

Amazon EMR provides a unique capability for you to use both methods. On the one hand one large cluster may be more efficient for processing regular batch workloads. On the other hand, if you require ad-hoc querying or workloads that vary with time, you may choose to create several separate cluster tuned to the specific task sharing data sources stored in Amazon S3. You can use EMR Managed Scaling to optimize resource usage.

Q: Can I access a script or jar resource which is on my local file system?

No. You must upload the script or jar to Amazon S3 or to the cluster's master node before it can be referenced. For uploading to Amazon S3 you can use tools including s3cmd, jets3t or S3Organizer.

Q: Can I run a persistent cluster executing multiple Hive queries?

Yes. You run a cluster in a manual termination mode so it will not terminate between Hive steps. To reduce the risk of data loss we recommend periodically persisting all of your important data in Amazon S3. It is good practice to regularly transfer your work to a new cluster to test your process for recovering from master node failure.

Q: Can multiple users execute Hive steps on the same source data?

Yes. Hive scripts executed by multiple users on separate clusters may contain create external table statements to concurrently import source data residing in Amazon S3.

Q: Can multiple users run queries on the same cluster?

Yes. In the batch mode, steps are serialized. Multiple users can add Hive steps to the same cluster; however, the steps will be executed serially. In interactive mode, several users can be logged on to the same cluster and execute Hive statements concurrently.

Q: Can data be shared between multiple AWS users?

Yes. Data can be shared using standard Amazon S3 sharing mechanism described here.

Q: Does Hive support access from JDBC?

Yes. Hive provides JDBC drive, which can be used to programmatically execute Hive statements. To start a JDBC service in your cluster you need to pass an optional parameter in

the Amazon EMR command line client. You also need to establish an SSH tunnel because the security group does not permit external connections.

Q: What is your procedure for updating packages on EMR AMIs?

On first boot, the Amazon Linux AMIs for EMR connect to the Amazon Linux AMI yum repositories to install security updates. When you use a custom AMI, you can disable this feature, but we don't recommend this for security reasons.

Q: Can I update my own packages on EMR clusters?

Yes. You can use Bootstrap Actions to install updates to packages on your clusters.

Q: Can I process DynamoDB data using Hive?

Yes. Simply define an external Hive table based on your DynamoDB table. You can then use Hive to analyze the data stored in DynamoDB and either load the results back into DynamoDB or archive them in Amazon S3. For more information please visit our Developer Guide.

Using Hudi

Q: What is Apache Hudi?

Apache Hudi is an open-source data management framework used to simplify incremental data processing and data pipeline development. Apache Hudi enables you to manage data at the record-level in Amazon S3 to simplify Change Data Capture (CDC) and streaming data ingestion, and provides a framework to handle data privacy use cases requiring record level updates and deletes. Data sets managed by Apache Hudi are stored in S3 using open storage formats, and integrations with Presto, Apache Hive, Apache Spark, and AWS Glue Data Catalog give you near real-time access to updated data using familiar tools.

Q: When should I use Apache Hudi?

Apache Hudi helps you with uses cases requiring record-level data management on S3. There are five common use-cases that benefit from these abilities:

1. Complying with data privacy laws that require organizations to remove user data, or update user preferences when users choose to change their preferences as to how their data can be used. Apache Hudi gives you the ability to perform record-level insert, update, and delete operations on your data stored in S3, using open source data formats such as Apache Parquet, and Apache Avro.
2. Consuming real time data streams and applying change data capture logs from enterprise systems. Many organizations require Enterprise Data Warehouses (EDW) and Operational Data Stores (ODS) data to be available in Amazon S3 so it's accessible to SQL engines like Apache Hive and Presto for data processing and analytics. Apache Hudi simplifies applying change logs, and gives users near real-time access to data.
3. Reinstating late arriving, or incorrect data. Late arriving, or incorrect data requires the data to be restated, and existing data sets updated to incorporate new, or updated records. Apache Hudi allows you to "upsert" records into an existing data set, relying on the framework to insert or update records based on their presence in the data set.
4. Tracking change to data sets and providing the ability to rollback changes. With Apache Hudi, each change to a data set is tracked as a commit, and can be easily rolled back, allowing you to find specific changes to a data set and "undo" them.

5. Simplifying file management on S3. To make sure data files are efficiently sized, customers have to build custom solutions that monitor and re-write many small files into fewer large files. With Apache Hudi, data files on S3 are managed, and users can simply configure an optimal file size to store their data and Hudi will merge files to create efficiently sized files.
6. Writing deltas to a target Hudi dataset. Hudi datasets can be pulled incrementally, which means you can get ALL and ONLY the updated & new rows since a specified instant time.

Q: How do I create an Apache Hudi data set?

Apache Hudi data sets are created using Apache Spark. Creating a data set is as simple as writing an Apache Spark DataFrame. The metadata for Apache Hudi data sets can optionally be stored in the AWS Glue Data Catalog or the Hive metastore to simplify data discovery and for integrating with Apache Hive and Presto.

Q: How does Apache Hudi manage data sets?

When creating a data set with Apache Hudi, you can choose what type of data access pattern the data set should be optimized for. For read-heavy use cases, you can choose the “Copy on Write” data management strategy to optimize for frequent reads of the data set. This strategy organizes data using columnar storage formats, and merges existing data and new updates when the updates are written. For write-heavy workloads, Apache Hudi uses the “Merge on Read” data management strategy which organizes data using a combination of columnar and row storage formats, where updates are appended to a file in row based storage format, while the merge is performed at read time to provide the updated results.

Q: How do I write to an Apache Hudi data set?

Changes to Apache Hudi data sets are made using Apache Spark. With Apache Spark, Apache Hudi data sets are operated on using the Spark DataSource API, enabling you to read and write data. DataFrame containing newly added data or updates to existing data can be written using the same DataSource API". You can also use the Hudi DeltaStreamer utility.

Q: How do I read from an Apache Hudi data set?

You can read data using either Apache Spark, Apache Hive, Presto, Amazon Redshift Spectrum or Amazon Athena. When you create a data set, you have the option to publish the metadata of that data set in either the AWS Glue Data Catalog, or the Hive metastore. If you choose to publish the metadata in a metastore, your data set will look just like an ordinary table, and you can query that table using Apache Hive and Presto.

Q: What considerations or limitations should I be aware of when using Apache Hudi?

For a full list of consideration and limitations when using Apache Hudi on Amazon EMR, please refer to our Amazon EMR documentation.

Q: How does my existing data work with Apache Hudi?

If you have existing data that you want to now manage with Apache Hudi, you can easily convert your Apache Parquet data to Apache Hudi data sets using an import tool provided with Apache Hudi on Amazon EMR, or you can use Hudi DeltaStreamer utility, or Apache Spark to rewrite your existing data as an Apache Hudi data set.

Using Impala

Q: What is Impala?

Impala is an open source tool in the Hadoop ecosystem for interactive, ad hoc querying using SQL syntax. Instead of using MapReduce, it leverages a massively parallel processing (MPP) engine similar to that found in traditional relational database management systems (RDBMS). With this architecture, you can query your data in HDFS or HBase tables very quickly, and leverage Hadoop's ability to process diverse data types and provide schema at runtime. This lends Impala to interactive, low-latency analytics. In addition, Impala uses the Hive metastore to hold information about the input data, including the partition names and data types. Also, Impala on Amazon EMR requires AMIs running Hadoop 2.x or greater. Click here to learn more about Impala.

Q: What can I do with Impala running on Amazon EMR?

Similar to using Hive with Amazon EMR, leveraging Impala with Amazon EMR can implement sophisticated data-processing applications with SQL syntax. However, Impala is built to perform faster in certain use cases (see below). With Amazon EMR, you can use Impala as a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence. Here are three use cases:

- Use Impala instead of Hive on long-running clusters to perform ad hoc queries. Impala reduces interactive queries to seconds, making it an excellent tool for fast investigation. You could run Impala on the same cluster as your batch MapReduce workflows, use Impala on a long-running analytics cluster with Hive and Pig, or create a cluster specifically tuned for Impala queries.
- Use Impala instead of Hive for batch ETL jobs on transient Amazon EMR clusters. Impala is faster than Hive for many queries, which provides better performance for these workloads. Like Hive, Impala uses SQL, so queries can easily be modified from Hive to Impala.
- Use Impala in conjunction with a third party business intelligence tool. Connect a client ODBC or JDBC driver with your cluster to use Impala as an engine for powerful visualization tools and dashboards.

Both batch and interactive Impala clusters can be created in Amazon EMR. For instance, you can have a long-running Amazon EMR cluster running Impala for ad hoc, interactive querying or use transient Impala clusters for quick ETL workflows.

Q: How is Impala different than traditional RDBMSs?

Traditional relational database systems provide transaction semantics and database atomicity, consistency, isolation, and durability (ACID) properties. They also allow tables to be indexed and cached so that small amounts of data can be retrieved very quickly, provide for fast updates of small amounts of data, and for enforcement of referential integrity constraints. Typically, they run on a single large machine and do not provide support for acting over complex user defined data types. Impala uses a similar distributed query system to that found in RDBMSs, but queries data stored in HDFS and uses the Hive metastore to hold information about the input data. As with Hive, the schema for a query is provided at runtime, allowing for easier schema changes. Also, Impala can query a variety of complex data types and execute user defined functions. However, because Impala processes data in-memory, it is important to understand the hardware limitations of your cluster and optimize your queries for the best performance.

Q: How is Impala different than Hive?

Impala executes SQL queries using a massively parallel processing (MPP) engine, while Hive executes SQL queries using MapReduce. Impala avoids Hive's overhead from creating MapReduce jobs, giving it faster query times than Hive. However, Impala uses significant memory resources and the cluster's available memory places a constraint on how much memory any query can consume. Hive is not limited in the same way, and can successfully process larger data sets with the same hardware. Generally, you should use Impala for fast, interactive queries, while Hive is better for ETL workloads on large datasets. Impala is built for speed and is great for ad hoc investigation, but requires a significant amount of memory to execute expensive queries or process very large datasets. Because of these limitations, Hive is recommended for workloads where speed is not as crucial as completion. Click here to view some performance benchmarks between Impala and Hive.

Q: Can I use Hadoop 1?

No, Impala requires Hadoop 2, and will not run on a cluster with an AMI running Hadoop 1.x.

Q: What instance types should I use for my Impala cluster?

For the best experience with Impala, we recommend using memory-optimized instances for your cluster. However, we have shown that there are performance gains over Hive when using standard instance types as well. We suggest reading our Performance Testing and Query Optimization section in the Amazon EMR Developer's Guide to better estimate the memory resources your cluster will need with regards to your dataset and query types. The compression type, partitions, and the actual query (number of joins, result size, etc.) all play a role in the memory required. You can use the EXPLAIN statement to estimate the memory and other resources needed for an Impala query.

Q: What happens if I run out of memory on a query?

If you run out of memory, queries fail and the Impala daemon installed on the affected node shuts down. Amazon EMR then restarts the daemon on that node so that Impala will be ready to run another query. Your data in HDFS on the node remains available, because only the daemon running on the node shuts down, rather than the entire node itself. For ad hoc analysis with Impala, the query time can often be measured in seconds; therefore, if a query fails, you can discover the problem quickly and be able to submit a new query in quick succession.

Q: Does Impala support user defined functions?

Yes, Impala supports user defined functions (UDFs). You can write Impala specific UDFs in Java or C++. Also, you can modify UDFs or user-defined aggregate functions created for Hive for use with Impala. For information about Hive UDFs, click here.

Q: Where is the data stored for Impala to query?

Impala queries data in HDFS or in HBase tables.

Q: Can I run Impala and MapReduce at the same time on a cluster?

Yes, you can set up a multitenant cluster with Impala and MapReduce. However, you should be sure to allot resources (memory, disk, and CPU) to each application using YARN on Hadoop 2.x. The resources allocated should be dependent on the needs for the jobs you plan to run on each application.

Q: Does Impala support ODBC and JDBC drivers?

While you can use ODBC drivers, Impala is also a great engine for third-party tools connected through JDBC. You can download and install the Impala client JDBC driver from

http://elasticmapreduce.s3.amazonaws.com/libs/impala/1.2.1/impala-jdbc-1.2.1.zip. From the client computer where you have your business intelligence tool installed, connect the JDBC driver to the master node of an Impala cluster using SSH or a VPN on port 21050. For more information, see Open an SSH Tunnel to the Master Node.

Using Pig

Q: What is Apache Pig?

Pig is an open source analytics package that runs on top of Hadoop. Pig is operated by a SQL-like language called Pig Latin, which allows users to structure, summarize, and query data sources stored in Amazon S3. As well as SQL-like operations, Pig Latin also adds first-class support for map/reduce functions and complex extensible user defined data types. This capability allows processing of complex and even unstructured data sources such as text documents and log files. Pig allows user extensions via user-defined functions written in Java and deployed via storage in Amazon S3.

Q: What can I do with Pig running on Amazon EMR?

Using Pig with Amazon EMR, you can implement sophisticated data-processing applications with a familiar SQL-like language and easy to use tools available with Amazon EMR. With Amazon EMR, you can turn your Pig applications into a reliable data warehouse to execute tasks such as data analytics, monitoring, and business intelligence tasks.

Q: How can I get started with Pig running on Amazon EMR?

The best place to start is to review our written documentation located here.

Q: Are there new features in Pig specific to Amazon EMR?

Yes. There are three new features which make Pig even more powerful when used with Amazon EMR, including:

a/ Accessing multiple filesystems. By default a Pig job can only access one remote file system, be it an HDFS store or S3 bucket, for input, output and temporary data. EMR has extended Pig so that any job can access as many file systems as it wishes. An advantage of this is that temporary intra-job data is always stored on the local HDFS, leading to improved performance.

b/ Loading resources from S3. EMR has extended Pig so that custom JARs and scripts can come from the S3 file system, for example “REGISTER s3:///my-bucket/piggybank.jar”

c/ Additional Piggybank function for String and DateTime processing.

Q: What types of Pig clusters are supported?

There are two types of clusters supported with Pig: interactive and batch. In an interactive mode a customer can start a cluster and run Pig scripts interactively directly on the master node. Typically, this mode is used to do ad hoc data analyses and for application development. In batch mode, the Pig script is stored in Amazon S3 and is referenced at the start of the cluster. Typically, batch mode is used for repeatable runs such as report generation.

Q: How can I launch a Pig cluster?

Both batch and interactive clusters can be started from AWS Management Console, EMR command line client, or APIs.

Q: What version of Pig does Amazon EMR support?

Amazon EMR supports multiple versions of Pig.

Q: Can I write to a S3 bucket from two clusters concurrently

Yes, you are able to write to the same bucket from two concurrent clusters.

Q: Can I share input data in S3 between clusters?

Yes, you are able to read the same data in S3 from two concurrent clusters.

Q: Can data be shared between multiple AWS users?

Yes. Data can be shared using standard Amazon S3 sharing mechanism described here http://docs.amazonwebservices.com/AmazonS3/latest/index.html?S3_ACLs.html

Q: Should I run one large cluster, and share it amongst many users or many smaller clusters?

Amazon EMR provides a unique capability for you to use both methods. On the one hand one large cluster may be more efficient for processing regular batch workloads. On the other hand, if you require ad-hoc querying or workloads that vary with time, you may choose to create several separate cluster tuned to the specific task sharing data sources stored in Amazon S3.

Q: Can I access a script or jar resource which is on my local file system?

No. You must upload the script or jar to Amazon S3 or to the cluster's master node before it can be referenced. For uploading to Amazon S3 you can use tools including s3cmd, jets3t or S3Organizer.

Q: Can I run a persistent cluster executing multiple Pig queries?

Yes. You run a cluster in a manual termination mode so it will not terminate between Pig steps. To reduce the risk of data loss we recommend periodically persisting all important data in Amazon S3. It is good practice to regularly transfer your work to a new cluster to test you process for recovering from master node failure.

Q: Does Pig support access from JDBC?

No. Pig does not support access through JDBC.

Using HBase

Q: What is Apache HBase?

HBase is an open source, non-relational, distributed database modeled after Google's BigTable. It was developed as part of Apache Software Foundation's Hadoop project and runs on top of Hadoop Distributed File System(HDFS) to provide BigTable-like capabilities for Hadoop. HBase provides you a fault-tolerant, efficient way of storing large quantities of sparse data using column-based compression and storage. In addition, HBase provides fast lookup of data because data is stored in-memory instead of on disk. HBase is optimized for sequential write operations, and it is highly efficient for batch inserts, updates, and deletes. HBase works seamlessly with Hadoop, sharing its file system and serving as a direct input and output to Hadoop jobs. HBase also integrates with Apache Hive, enabling SQL-like queries over HBase tables, joins with Hive-based tables, and support for Java Database Connectivity (JDBC). You can learn more about Apache HBase here.

Q: Are there new features in HBase specific to Amazon EMR?

With Amazon EMR, you can you can use HBase on Amazon S3 to store a cluster's HBase root directory and metadata directly to Amazon S3 and create read replicas and snapshots. Please see our documentation to learn more.

Q: Which versions of HBase are supported on Amazon EMR?

You can look at the latest HBase versions supported on Amazon EMR here.

Kinesis connector

Q: What does EMR Connector to Kinesis enable?

The connector enables EMR to directly read and query data from Kinesis streams. You can now perform batch processing of Kinesis streams using existing Hadoop ecosystem tools such as Hive, Pig, MapReduce, Hadoop Streaming, and Cascading.

Q: What does the EMR connector to Kinesis enable that I couldn't have done before?

Reading and processing data from a Kinesis stream would require you to write, deploy and maintain independent stream processing applications. These take time and effort. However, with this connector, you can start reading and analyzing a Kinesis stream by writing a simple Hive or Pig script. This means you can analyze Kinesis streams using SQL! Of course, other Hadoop ecosystem tools could be used as well. You don't need to developed or maintain a new set of processing applications.

Q: Who will find this functionality useful?

The following types of users will find this integration useful:

- Hadoop users who are interested in utilizing the extensive set of Hadoop ecosystem tools to analyze Kinesis streams.
- Kinesis users who are looking for an easy way to get up and running with stream processing and ETL of Kinesis data.
- Business analysts and IT professionals who would like to perform ad-hoc analysis of data in Kinesis streams using familiar tools like SQL (via Hive) or scripting languages like Pig.

Q: What are some use cases for this integration?

The following are representative use cases are enabled by this integration:

- Streaming Log Analysis: You can analyze streaming web logs to generate a list of top 10 error type every few minutes by region, browser, and access domains.
- Complex Data Processing Workflows: You can join Kinesis stream with data stored in S3, Dynamo DB tables, and HDFS. You can write queries that join clickstream data from Kinesis with advertising campaign information stored in a DynamoDB table to identify the most effective categories of ads that are displayed on particular websites.
- Ad-hoc Queries: You can periodically load data from Kinesis into HDFS and make it available as a local Impala table for fast, interactive, analytic queries.

Q: What EMR AMI version do I need to be able to use the connector?

You need to use EMR's AMI version 3.0.4 and later.

Q: Is this connector a stand-alone tool?

No, it is a built in component of the Amazon distribution of Hadoop and is present on EMR AMI versions 3.0.4 and later. Customer simply needs to spin up a cluster with AMI version 3.0.4 or later to start using this feature.

Q: What data format is required to allow EMR to read from a Kinesis stream?

The EMR Kinesis integration is not data format-specific. You can read data in any format. Individual Kinesis records are presented to Hadoop as standard records that can be read using any Hadoop MapReduce framework. Individual frameworks like Hive, Pig and Cascading have built in components that help with serialization and deserialization, making it easy for developers to query data from many formats without having to implement custom code. For example, in Hive users can read data from JSON files, XML files and SEQ files by specifying the appropriate Hive SerDe when they define a table. Pig has a similar component called Loadfunc/Evalfunc and Cascading has a similar component called a Tap. Hadoop users can leverage the extensive ecosystem of Hadoop adapters without having to write format-specific code. You can also implement custom deserialization formats to read domain specific data in any of these tools.

Q: How do I analyze a Kinesis stream using Hive in EMR?

Create a table that references a Kinesis stream. You can then analyze the table like any other table in Hive. Please see our tutorials for page more details.

Q: Using Hive, how do I create queries that combine Kinesis stream data with other data source?

First create a table that references a Kinesis stream. Once a Hive table has been created, you can join it with tables mapping to other data sources such as Amazon S3, Amazon Dynamo DB, and HDFS. This effectively results in joining data from Kinesis stream to other data sources.

Q: Is this integration only available for Hive?

No, you can use Hive, Pig, MapReduce, Hadoop Streaming, and Cascading.

Q: How do I setup scheduled jobs to run on a Kinesis stream?

The EMR Kinesis input connector provides features that help you configure and manage scheduled periodic jobs in traditional scheduling engines such as Cron. For example, you can develop a Hive script that runs every N minutes. In the configuration parameters for a job, you can specify a Logical Name for the job. The Logical Name is a label that will inform the EMR Kinesis input connector that individual instances of the job are members of the same periodic schedule. The Logical Name allows the process to take advantage of iterations, which are explained next.

Since MapReduce is a batch processing framework, to analyze a Kinesis stream using EMR, the continuous stream is divided in to batches. Each batch is called an Iteration. Each Iteration is assigned a number, starting with 0. Each Iteration's boundaries are defined by a start sequence number and end sequence number. Iterations are then processed sequentially by EMR.

In the event of an attempt's failure, the EMR Kinesis input connector will re-try the iteration within the Logical Name from the known start sequence number of the iteration. This functionality ensures that successive attempts on the same iteration will have precisely the same input records from the Kinesis stream as the previous attempts. This guarantees idempotent (consistent) processing of a Kinesis stream.

You can specify Logical Names and Iterations as runtime parameters in your respective Hadoop tools. For example, in the tutorial section “Running queries with checkpoints”, the code sample shows a scheduled Hive query that designates a Logical Name for the query and increments the iteration with each successive run of the job.

Additionally, a sample cron scheduling script is provided in the tutorials.

Q: Where is the metadata for Logical Names and Iterations stored?

The metadata that allows the EMR Kinesis input connector to work in scheduled periodic workflows is stored in Amazon DynamoDB. You must provision an Amazon Dynamo DB table and specify it as an input parameter to the Hadoop Job. It is important that you configure appropriate IOPS for the table to enable this integration. Please refer to the getting started tutorial for more information on setting up your Amazon Dynamo DB table.

Q: What happens when an iteration processing fails?

Iterations identifiers are user-provided values that map to specific boundary (start and end sequence numbers) in a Kinesis stream. Data corresponding to these boundaries is loaded in the Map phase of the MapReduce job. This phase is managed by the framework and will be automatically re-run (three times by default) in case of job failure. If all the retries fail, you would still have options to retry the processing starting from last successful data boundary or past data boundaries. This behavior is controlled by providing kinesis.checkpoint.iteration.no parameter during processing. Please refer to the getting started tutorial for more information on how this value is configured for different tools in the Hadoop ecosystem.

Q: Can I run multiple queries on the same iteration?

Yes, you can specify a previously run iteration by setting the kinesis.checkpoint.iteration.no parameter in successive processing. The implementation ensures that successive runs on the same iteration will have precisely the same input records from the Kinesis stream as the previous runs.

Q: What happens if records in an Iteration expire from the Kinesis stream?

In the event that the beginning sequence number and/or end sequence number of an iteration belong to records that have expired from the Kinesis steam, the Hadoop job will fail. You would need to use a different Logical Name to process data from the beginning of the Kinesis stream.

Q: Can I push data from EMR into Kinesis stream?

No. The EMR Kinesis connector currently does not support writing data back into a Kinesis stream.

Q: Does the EMR Hadoop input connector for Kinesis enable continuous stream processing?

The Hadoop MapReduce framework is a batch processing system. As such, it does not support continuous queries. However there is an emerging set of Hadoop ecosystem frameworks like Twitter Storm and Spark Streaming that enable to developers build applications for continuous stream processing. A Storm connector for Kinesis is available at on GitHub here and you can find a tutorial explaining how to setup Spark Streaming on EMR and run continuous queries here.

Additionally, developers can utilize the Kinesis client library to develop real-time stream processing applications. You can find more information on developing custom Kinesis applications in the Kinesis documentation here.

Q: Can I specify access credential to read a Kinesis stream that is managed in another AWS account?

Yes. You can read streams from another AWS account by specifying the appropriate access credentials of the account that owns the Kinesis stream. By default, the Kinesis connector utilizes the user-supplied access credentials that are specified when the cluster is created. You can override these credentials to access streams from other AWS Accounts by setting the kinesis.accessKey and kinesis.secretKey parameters. The following examples show how to set the kinesis.accessKey and kinesis.secretKey parameters in Hive and Pig.

Code sample for Hive:

```
...
STORED BY
'com.amazon.emr.kinesis.hive.KinesisStorageHandler'
TBLPROPERTIES(
"kinesis.accessKey"="AwsAccessKey",
"kinesis.secretKey"="AwsSecretKey",
);
```

Code sample for Pig:

```
…
raw_logs = LOAD 'AccessLogStream' USING com.amazon.emr.kinesis.pig.Kin
esisStreamLoader('kinesis.accessKey=AwsAccessKey', 'kinesis.secretKey=AwsSecretKey'
) AS (line:chararray);
```

Q: Can I run multiple parallel queries on a single Kinesis Stream? Is there a performance impact?

Yes, a customer can run multiple parallel queries on the same stream by using separate logical names for each query. However, reading from a shard within a Kinesis stream is subjected to a rate limit of 2MB/sec. Thus, if there are N parallel queries running on the same stream, each one would get roughly (2/N) MB/sec egress rate per shard on the stream. This may slow down the processing and in some cases fail the queries as well.

Q: Can I join and analyze multiple Kinesis streams in EMR?

Yes, for example in Hive, you can create two tables mapping to two different Kinesis streams and create joins between the tables.

Q: Does the EMR Kinesis connector handle Kinesis scaling events, such as merge and split events?

Yes. The implementation handles split and merge events. The Kinesis connector ties individual Kinesis shards (the logical unit of scale within a Kinesis stream) to Hadoop MapReduce map tasks. Each unique shard that exists within a stream in the logical period of an Iteration will result in exactly one map task. In the event of a shard split or merge event, Kinesis will provision new unique shard Ids. As a result, the MapReduce framework will provision more map tasks to read from Kinesis. All of this is transparent to the user.

Q: What happens if there are periods of “silence” in my stream?

The implementation allows you to configure a parameter called kinesis.nodata.timeout. For example, consider a scenario where kinesis.nodata.timeout is set to 2 minutes and you want to run a Hive query every 10 minutes. Additionally, consider some data has been written to the stream since the last iteration (10 minutes ago). However, currently no new records are arriving, i.e. there is a silence in the stream. In this case, when the current iteration of the query launches,

the Kinesis connector would find that no new records are arriving. The connector will keep polling the stream for 2 minutes and if no records arrive for that interval then it will stop and process only those records that were already read in the current batch of stream. However, if new records start arriving before kinesis.nodata.timeout interval is up, then the connector will wait for an additional interval corresponding to a parameter called kinesis.iteration.timeout. Please look at the tutorials to see how to define these parameters.

Q: How do I debug a query that continues to fail in each iteration?

In the event of a processing failure, you can utilize the same tools they currently do when debugging Hadoop Jobs. Including the Amazon EMR web console, which helps identify and access error logs. More details on debugging an EMR job can be found here.

Q: What happens if I specify a DynamoDB table that I don't have access to?

The job would fail and the exception would show up in error logs for the job.

Q: What happens if job doesn't fail but checkpointing to DynamoDB fails?

The job would fail and the exception would show up in error logs for the job.

Q: How do I maximize the read throughput from Kinesis stream to EMR?

Throughput from Kinesis stream increases with instance size used and record size in the Kinesis stream. We recommend that you use m1.xlarge and above for both master and core nodes for this feature.

Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that makes it easy to quickly and cost-effectively process vast amounts of data.Amazon EMR is the industry-leading cloud big data platform for processing vast amounts of data using open source tools such as Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi, and Presto. Amazon EMR makes it easy to set up, operate, and scale your big data environments by automating time-consuming tasks like provisioning capacity and tuning clusters and uses Hadoop, an open source framework, to distribute your data and processing across a resizable cluster of Amazon EC2 instances. Amazon EMR is used in a variety of applications, including log analysis, web indexing, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics. Customers launch millions of Amazon EMR clusters every year.

Benefits

Easy to use

You can use EMR Studio, an integrated development environment (IDE), to easily develop, visualize, and debug data engineering and data science applications written in R, Python, Scala, and PySpark.

Low cost

EMR pricing is simple and predictable: You pay a per-instance rate for every second used, with a one-minute minimum charge. You can launch a 10-node EMR cluster for as little as ¥0.187

per hour. You can save the cost of the instances by selecting Amazon EC2 Spot for transient workloads and Reserved Instances for long-running workloads.

Elastic

Unlike the rigid infrastructure of on-premises clusters, EMR decouples compute and storage, giving you the ability to scale each independently and take advantage of the tiered storage of Amazon S3. With EMR, you can provision one, hundreds, or thousands of compute instances or containers to process data at any scale. The number of instances can be increased or decreased automatically using Auto Scaling (which manages cluster sizes based on utilization) and you only pay for what you use.

Reliable

Spend less time tuning and monitoring your cluster. EMR is tuned for the cloud and constantly monitors your cluster — retrying failed tasks and automatically replacing poorly performing instances. Clusters are highly available and automatically failover in the event of a node failure. EMR provides the latest stable open source software releases, so you don't have to manage updates and bug fixes, which leads to fewer issues and less effort to maintain your environment.

Secure

EMR automatically configures EC2 firewall settings, controlling network access to instances and launches clusters in an Amazon Virtual Private Cloud (VPC). Server-side encryption or client-side encryption can be used with the Amazon Key Management Service or your own customer-managed keys. EMR makes it easy to enable other encryption options, like in-transit and at-rest encryption, and strong authentication with Kerberos. You can use Amazon Lake Formation or Apache Ranger to apply fine-grained data access controls for databases, tables, and columns.

Flexible

You have complete control over your EMR clusters and your individual EMR jobs. You can launch EMR clusters with custom Amazon Linux AMIs and easily configure the clusters using scripts to install additional third party software packages. EMR enables you to reconfigure applications on running clusters on the fly without the need to relaunch clusters. Also, you can customize the execution environment for individual jobs by specifying the libraries and runtime dependencies in a Docker container and submit them with your job.

Amazon Managed Streaming for Apache Kafka (Amazon MSK)

Securely stream data with a fully managed, highly available Apache Kafka service.

Fully managed, highly available, and secure Apache Kafka service.

Amazon MSK is a fully managed service that makes it easy for you to build and run applications that use Apache Kafka to process streaming data. Apache Kafka is an open-source platform for building real-time streaming data pipelines and applications. With Amazon MSK, you can use native Apache Kafka APIs to populate data lakes, stream changes to and from databases, and power machine learning and analytics applications.

Apache Kafka clusters are challenging to setup, scale, and manage in production. When you run Apache Kafka on your own, you need to provision servers, configure Apache Kafka manually, replace servers when they fail, orchestrate server patches and upgrades, architect the cluster for high availability, ensure data is durably stored and secured, setup monitoring and alarms, and carefully plan scaling events to support load changes. Amazon MSK makes it easy for you to build and run production applications on Apache Kafka without needing Apache Kafka infrastructure management expertise. That means you spend less time managing infrastructure and more time building applications.
With a few clicks in the Amazon MSK console you can create highly available Apache Kafka clusters with settings and configuration based on Apache Kafka's deployment best practices. Amazon MSK automatically provisions and runs your Apache Kafka clusters. Amazon MSK continuously monitors cluster health and automatically replaces unhealthy nodes with no downtime to your application. In addition, Amazon MSK secures your Apache Kafka cluster by encrypting data at rest.

How it works

Amazon MSK makes it easy to ingest and process streaming data in real time with fully managed Apache Kafka.

Use cases

Ingest and process log and event streams
Capture events with MSK, and then express your stream processing logic within Apache Zeppelin notebooks to derive insights from data streams in milliseconds.

Run centralized state or data buses
Use Amazon MSK and the Apache Kafka log structure to form real-time, centralized, and privately accessible data buses.

Power your event-driven systems
Ingest and respond to digital changes occurring throughout your applications and business infrastructure in real time.

Benefits

Fully compatible
Amazon MSK runs and manages Apache Kafka for you. This makes it easy for you to migrate and run your existing Apache Kafka applications on Amazon Web Services without changes to the application code. By using Amazon MSK, you maintain open source compatibility and can continue to use familiar custom and community-built tools such as MirrorMaker and Prometheus.

Fully managed
Amazon MSK lets you focus on creating your streaming applications without having to worry about the operational overhead of managing your Apache Kafka environment. Amazon MSK manages the provisioning, configuration, and maintenance of Apache Kafka clusters and Apache ZooKeeper nodes for you. Amazon MSK also shows key Apache Kafka performance metrics in the Amazon Web Services console.

Highly available
Amazon MSK creates an Apache Kafka cluster and offers multi-AZ replication within an Amazon Web Services Region. Amazon MSK continuously monitors cluster health, and if a component fails, Amazon MSK will automatically replace it.

Highly secure
Amazon MSK provides multiple levels of security for your Apache Kafka clusters including VPC network isolation, Amazon IAM for control-plane API authorization, encryption at rest, TLS encryption in-transit.

How it works

Apache Kafka is a streaming data store that decouples applications producing streaming data (producers) into its data store from applications consuming streaming data (consumers) from its data store. Organizations use Apache Kafka as a data source for applications that continuously analyze and react to streaming data.

With a few clicks in the Amazon MSK Console Amazon MSK provisions your Apache Kafka cluster and manages Apache Kafka upgrades so you are always using the most secure and the fastest version of Apache Kafka. Once you configure your clusters, your applications can stream data from producers to a topic, where this data is read in real-time by consumers.

What is Amazon MSK?

Amazon Managed Streaming for Apache Kafka (Amazon MSK) is a fully managed service that enables you to build and run applications that use Apache Kafka to process streaming data. Amazon MSK provides the control-plane operations, such as those for creating, updating, and deleting clusters. It lets you use Apache Kafka data-plane operations, such as those for producing and consuming data. It runs open-source versions of Apache Kafka. This means existing applications, tooling, and plugins from partners and the Apache Kafka community are supported without requiring changes to application code. You can use Amazon MSK to create clusters that use any of the Apache Kafka versions listed under Supported Apache Kafka versions.

The following diagram provides an overview of how Amazon MSK works.

The diagram demonstrates the interaction between the following components:

- **Broker nodes** — When creating an Amazon MSK cluster, you specify how many broker nodes you want Amazon MSK to create in each Availability Zone. In the example cluster shown in this diagram, there's one broker per Availability Zone. Each Availability Zone has its own virtual private cloud (VPC) subnet.
- **ZooKeeper nodes** — Amazon MSK also creates the Apache ZooKeeper nodes for you. Apache ZooKeeper is an open-source server that enables highly reliable distributed coordination.
- **Producers, consumers, and topic creators** — Amazon MSK lets you use Apache Kafka data-plane operations to create topics and to produce and consume data.
- **Cluster Operations** You can use the AWS Management Console, the AWS Command Line Interface (AWS CLI), or the APIs in the SDK to perform control-plane operations. For example, you can create or delete an Amazon

MSK cluster, list all the clusters in an account, view the properties of a cluster, and update the number and type of brokers in a cluster.

Amazon MSK detects and automatically recovers from the most common failure scenarios for clusters so that your producer and consumer applications can continue their write and read operations with minimal impact. When Amazon MSK detects a broker failure, it mitigates the failure or replaces the unhealthy or unreachable broker with a new one. In addition, where possible, it reuses the storage from the older broker to reduce the data that Apache Kafka needs to replicate. Your availability impact is limited to the time required for Amazon MSK to complete the detection and recovery. After a recovery, your producer and consumer apps can continue to communicate with the same broker IP addresses that they used before the failure.

Amazon Managed Streaming for Apache Kafka (Amazon MSK) FAQs

General

Q: What is Amazon MSK?
Amazon Managed Streaming for Apache Kafka (Amazon MSK) is an AWS streaming data service that manages Apache Kafka infrastructure and operations, making it easy for developers and DevOps managers to run Apache Kafka applications and Kafka Connect connectors on AWS, without the need to become experts in operating Apache Kafka. Amazon MSK operates, maintains, and scales Apache Kafka clusters, provides enterprise-grade security features out of the box, and has built-in AWS integrations that accelerate development of streaming data applications. To get started, you can migrate existing Apache Kafka workloads and Kafka Connect connectors into Amazon MSK, or with a few clicks, you can build new ones from scratch. There are no data transfer charges for in-cluster traffic, and no commitments or upfront payments required. You only pay for the resources that you use.

Q: What is Apache Kafka?
Apache Kafka is an open-source, high performance, fault-tolerant, and scalable platform for building real-time streaming data pipelines and applications. Apache Kafka is a streaming data store that decouples applications producing streaming data (producers) into its data store from applications consuming streaming data (consumers) from its data store. Organizations use Apache Kafka as a data source for applications that continuously analyze and react to streaming data. Learn more about Apache Kafka.

Q. What is Kafka Connect?
Kafka Connect, an open-source component of Apache Kafka, is a framework for connecting Apache Kafka with external systems, such as databases, key-value stores, search indexes, and file systems.

Q: What is streaming data?
Streaming data is a continuous stream of small records or events (a record or event is typically a few kilobytes) generated by thousands of machines, devices, websites, and applications. Streaming data includes a wide variety of data such as log files generated by customers using

your mobile or web applications, ecommerce purchases, in-game player activity, information from social networks, financial trading floors, geospatial services, security logs, metrics, and telemetry from connected devices or instrumentation in data centers. Streaming data services like Amazon MSK and Amazon Kinesis Data Streams make it easy for you to continuously collect, process, and deliver streaming data. Learn more about streaming data.

Q: What are Apache Kafka's primary capabilities?
Apache Kafka has three key capabilities:

- Apache Kafka stores streaming data in a fault-tolerant way, providing a buffer between producers and consumers. It stores events as a continuous series of records and preserves the order in which the records were produced.
- Apache Kafka allows many data producers—e.g. websites, internet of things (IoT) devices, Amazon Elastic Compute Cloud (Amazon EC2) instances—to continuously publish streaming data and categorize it using Apache Kafka topics. Multiple data consumers (e.g. machine learning applications, AWS Lambda functions, or microservices) read from these topics at their own rate, similar to a message queue or enterprise messaging system.
- Data consumers process data from Apache Kafka topics on a first-in-first-out basis, preserving the order data was produced.

Q: What are the key concepts of Apache Kafka?
Apache Kafka stores records in topics. Data producers write records to topics and consumers read records from topics. Each record in Apache Kafka consists of a key, a value, and a timestamp and sometimes header metadata. Apache Kafka partitions topics and replicates these partitions across multiple nodes called brokers. Apache Kafka runs as a cluster on one or more brokers, and brokers can be located in multiple AWS availability zones to create a highly available cluster. Apache Kafka relies on Apache ZooKeeper to coordinate cluster tasks and can maintain state for resources interacting with an Apache Kafka cluster.

Q: When should I use Apache Kafka?
Apache Kafka supports real-time applications that transform, deliver, and react to streaming data, and can be used to build real-time streaming data pipelines that reliably send data between multiple systems or applications.

Q: What does Amazon MSK do?
Amazon MSK makes it easy to get started and run open-source versions of Apache Kafka on AWS with high availability and security. Amazon MSK also offers integrations with AWS services without the operational overhead of running an Apache Kafka cluster. Amazon MSK allows you to use open-source versions of Apache Kafka while the service manages the setup, provisioning, AWS integrations, and on-going maintenance of Apache Kafka clusters.

With a few clicks in the console, you can create an Amazon MSK cluster. From there, Amazon MSK replaces unhealthy brokers, automatically replicates data for high availability, manages Apache ZooKeeper nodes, automatically deploys hardware patches as needed, manages the integrations with AWS services, makes important metrics visible through the console, and supports Apache Kafka version upgrades so you can take advantage of improvements to the open-source version of Apache Kafka.

Q: What Apache Kafka versions does Amazon MSK support?
For supported Kafka versions, see the Amazon MSK documentation.

Q: Are Apache Kafka APIs compatible with Amazon MSK?
Yes, all data plane and admin APIs are natively supported by Amazon MSK.

Q: Is the Apache Kafka AdminClient supported by Amazon MSK?
Yes.

Q: Does Amazon MSK support schema registration?
Yes, Apache Kafka clients can use the AWS Glue Schema Registry, a serverless feature of AWS Glue, at no additional charge. Visit the Schema Registry user documentation to get started and to learn more.

MSK Serverless

Q: What is MSK Serverless?
MSK Serverless is a cluster type for Amazon MSK that makes it easy for you to run Apache Kafka clusters without having to manage compute and storage capacity. With MSK Serverless, you can run your applications without having to provision, configure, or optimize clusters, and you pay for the data volume you stream and retain.

Q: Does MSK Serverless automatically balance partitions within a cluster?
Yes. MSK Serverless fully manages partitions, including monitoring and moving them to even load across a cluster.

Q: How much data throughput capacity does MSK Serverless support?
MSK Serverless provides up to 200 MBps of write capacity and 400 MBps of read capacity per cluster. Additionally, to ensure sufficient throughput availability for all partitions in a cluster, MSK Serverless allocates up to 5 MBps of instant write capacity and 10 MBps of instant read capacity per partition.

Q: What security features does MSK Serverless offer?
MSK Serverless encrypts all traffic in transit and all data at rest using service-managed keys issued through AWS Key Management Service (KMS). Clients connect to MSK Serverless over a private connection using AWS PrivateLink without exposing your traffic to the public internet. Additionally, MSK Serverless offers IAM Access Control, which you can use to manage client authentication and client authorization to Apache Kafka resources such as topics.

Q: How can producers and consumers access my MSK Serverless clusters?
When you create a MSK Serverless cluster, you provide subnets of one or more Amazon Virtual Private Clouds (VPCs) that host the clients of the cluster. Clients hosted in any of these VPCs will be able to connect to the MSK Serverless cluster using its broker bootstrap string.

Q: Which regions is MSK Serverless available in?
Please refer to the MSK pricing page for up-to-date regional availability.

Q: Which authentication types does MSK Serverless support?
MSK Serverless currently supports AWS IAM for client authentication and authorization. Your clients can assume an AWS IAM role for authentication, and you can enforce access control using an associated IAM policy.

Q: How do I process data in my MSK Serverless cluster?
You can use any Apache Kafka compatible tools to process data in your MSK Serverless cluster topics. MSK Serverless integrates with Amazon Kinesis Data Analytics for Apache Flink for stateful stream processing and AWS Lambda for event processing. You can also use Kafka Connect sink connectors to send data to any desired destination. (MSK Connect support for MSK Serverless is coming soon.)

Q: How does MSK Serverless ensure high availability?
When you create a partition, MSK Serverless creates 2 replicas of it and places them in different

availability zones. Additionally, MSK serverless automatically detects and recovers failed backend resources to maintain high availability.

Data production and consumption

Q: Can I use Apache Kafka APIs to get data in and out of Apache Kafka?
Yes, Amazon MSK supports the native Apache Kafka producer and consumer APIs. Your application code does not need to change when clients begin to work with clusters within Amazon MSK.

Q: Can I use Apache Kafka Connect, Apache Kafka Streams, or any other ecosystem component of Apache Kafka with Amazon MSK?
Yes, you can use any component that leverages the Apache Kafka producer and consumer APIs, and the Apache Kafka Admin Client. Tools that upload .jar files into Apache Kafka clusters are currently not compatible with Amazon MSK, including Confluent Control Center, Confluent Auto Data Balancer, and Uber uReplicator.

Migrating to Amazon MSK

Q: Can I migrate data within my existing Apache Kafka cluster to Amazon MSK?
Yes, you can use third-party tools or open-source tools like MirrorMaker, supported by Apache Kafka, to replicate data from clusters into an Amazon MSK cluster. Here is an Amazon MSK migration lab to help you complete a migration.

Version upgrades

Q: Are Apache Kafka version upgrades supported?
Yes, Amazon MSK supports fully managed in-place Apache Kafka version upgrades for provisioned clusters. To learn more about upgrading your Apache Kafka version and high availability best practices, see the version upgrades documentation.

Clusters

Q: How do I create my first Amazon MSK cluster?
You can create your first cluster with a few clicks in the AWS management console or using the AWS SDKs. First, in the Amazon MSK console select an AWS region to create an Amazon MSK cluster in. Choose a name for your cluster, the Virtual Private Cloud (VPC) you want to run the cluster with, and the subnets for each AZ. If you are creating a provisioned cluster, you will also be able to pick a broker instance type, quantity of brokers per AZ, and storage per broker.

Q: What resources are within a cluster?
Provisioned clusters contain broker instances, provisioned storage, and abstracted Apache ZooKeeper nodes. Serverless clusters are a resource in of themselves, which abstract away all underlying resources.

Q: What types of broker instances can I provision within an Amazon MSK cluster?
For provisioned clusters, you can choose EC2 T3.small or instances within the EC2 M5 instance family. For serverless clusters, brokers are completely abstracted.

Q: Does Amazon MSK offer Reserved Instance pricing?
No, not at this time.

Q: Do I need to provision and pay for broker boot volumes?
No, each broker you provision includes boot volume storage managed by the Amazon MSK service.

Q: When I create an Apache Kafka cluster, do the underlying resources (e.g. Amazon EC2 instances) show up in my EC2 console?
Some resources, like elastic network interfaces (ENIs), will show up in your Amazon EC2 account. Other Amazon MSK resources will not show up in your EC2 account as these are managed by the Amazon MSK service.

Q: What do I need to provision within an Amazon MSK cluster?
For provisioned clusters, you need to provision broker instances and broker storage with every cluster you create. You may optionally provision storage throughput for storage volumes, which can be used to seamlessly scale I/O without having to provision additional brokers. You do not need to provision Apache ZooKeeper nodes as these resources are included at no additional charge with each cluster you create. For serverless clusters, you just create a cluster as a resource.

Q: What is the default broker configuration for a cluster?
Unless otherwise specified, Amazon MSK uses the same defaults specified by the open-source version of Apache Kafka. The default settings for both cluster types are documented here.
Q: Can I provision brokers such that they are imbalanced across AZs (e.g. 3 in us-east-1a, 2 in us-east-1b, 1 in us-east-1c)?
No, Amazon MSK enforces the best practice of balancing broker quantities across AZs within a cluster.

Q: How does data replication work in Amazon MSK?
Amazon MSK uses Apache Kafka's leader-follower replication to replicate data between brokers. Amazon MSK makes it easy to deploy clusters with multi-AZ replication and gives you the option to use a custom replication strategy by topic. By default with each of the replication options, leader and follower brokers will be deployed and isolated using the replication strategy specified. For example, if you select a three AZ broker replication strategy with 1 broker per AZ cluster, Amazon MSK will create a cluster of three brokers (one broker in three AZs in a region), and by default (unless you choose to override the topic replication factor) the topic replication factor will also be three.
Q: Can I change the default broker configurations or upload a cluster configuration to Amazon MSK?
Yes, Amazon MSK allows you to create custom configurations and apply them to new and existing clusters. For more information on custom configurations, see the configuration documentation.

Q: What configuration properties am I able to customize?
The configurations properties that you can customize are documented here.

Q: What is the default configuration of a new topic?
Amazon MSK uses Apache Kafka's default configuration unless otherwise specified here.

Topics

Q: How do I create topics?
Once your Apache Kafka cluster has been created, you can create topics using the Apache Kafka APIs. All topic and partition level actions and configurations are performed using Apache Kafka APIs. The following command is an example of creating a topic using Apache Kafka APIs and the configuration details available for your cluster:

```
bin/kafka-topics.sh --create —bootstrap-server <BootstrapBrokerString> --replication-factor 3 --partitions 1 --topic TopicName
```

Networking

Q: Does Amazon MSK run in an Amazon VPC?
Yes, Amazon MSK always runs within an Amazon VPC managed by the Amazon MSK service. Amazon MSK resources will be available to your own Amazon VPC, subnet, and security group you select when the cluster is setup. IP addresses from your VPC are attached to your Amazon MSK resources through elastic network interfaces (ENIs), and all network traffic stays within the AWS network and is not accessible to the internet by default.

Q: How will the brokers in my Amazon MSK cluster be made accessible to clients within my VPC?
The brokers in your cluster will be made accessible to clients in your VPC through ENIs appearing in your account. The Security Groups on the ENIs will dictate the source and type of ingress and egress traffic allowed on your brokers.

Q: Is it possible to connect to my cluster over the public Internet?
Yes, Amazon MSK offers an option to securely connect to the brokers of Amazon MSK clusters running Apache Kafka 2.6.0 or later versions over the internet. By enabling public access, authorized clients external to a private Amazon Virtual Private Cloud (VPC) can stream encrypted data in and out of specific Amazon MSK clusters. You can enable public access for MSK clusters after a cluster has been created at no additional cost, but standard AWS data transfer costs for cluster ingress and egress apply. To learn more about turning on public access, see the public access documentation.

Q: Is the connection between my clients and an Amazon MSK cluster private?
By default, the only way data can be produced and consumed from an Amazon MSK cluster is over a private connection between your clients in your VPC and the Amazon MSK cluster. However, if you turn on public access for your Amazon MSK cluster and connect to your MSK cluster using the public bootstrap-brokers string, the connection, though authenticated, authorized and encrypted, is no longer considered private. We recommend that you configure the cluster's security groups to have inbound TCP rules that allow public access from your trusted IP address and make these rules as restrictive as possible if you turn on public access.

Connecting to the VPC

Q: How do I connect to my Amazon MSK cluster outside of the VPC?
There are several methods to connect to your Amazon MSK clusters outside of your VPC. The easiest way is to turn on public connectivity over the internet to the brokers of MSK clusters running Apache Kafka 2.6.0 or later versions. For security reasons, you can't turn on public access while creating an MSK cluster. However, you can update an existing cluster to make it publicly accessible. You can also create a new cluster and then update it to make it publicly accessible. To learn more about turning on public access, see the public access documentation. If you want to connect to your MSK cluster from inside AWS network but outside the cluster's Amazon VPC, you can use VPC Peering, AWS Direct Connect, AWS Transit Gateway, VPN or a REST Proxy. To learn more about setting up these options, see Access from within AWS documentation.

Encryption

Q: Can I encrypt data in my Amazon MSK cluster?
Yes, Amazon MSK uses Amazon Elastic Block Store (Amazon EBS) server-side encryption and AWS Key Management Service (AWS KMS) keys to encrypt storage volumes.

Q: Is data encrypted in transit between brokers within an Amazon MSK cluster?
Yes, by default new clusters have encryption in-transit enabled via TLS for inter-broker

communication. For provisioned clusters, you can opt out of using encryption in transit when a cluster is created.

Q: Is data encrypted in transit between my Apache Kafka clients and the Amazon MSK service?
Yes, by default in-transit encryption is set to TLS only for clusters created from the CLI or AWS Management Console. Additional configuration is required for clients to communicate with clusters using TLS encryption. For provisioned clusters, you can change the default encryption setting by selecting the TLS/plaintext or plaintext settings. Read more about MSK Encryption.

Q: Is data encrypted in transit as it moves between brokers and Apache ZooKeeper nodes in an Amazon MSK cluster?
Yes, Amazon MSK clusters running Apache Kafka version 2.5.1 or greater support TLS in-transit encryption between Kafka brokers and ZooKeeper nodes.

Access Management

Q: How do I control cluster authentication and Apache Kafka API authorization?
For serverless clusters, you can use IAM Access Control for both authentication and authorization. For provisioned clusters, you have three options: 1) AWS Identity and Access Management (IAM) Access Control for both AuthN/Z (recommended), 2) TLS certificate authentication (CA) for AuthN and access control lists for AuthZ, and 3) SASL/SCRAM for AuthN and access control lists for AuthZ. Amazon MSK recommends using IAM Access Control. It is the easiest to use and, because it defaults to least privilege access, the most secure option.

Q: How does authorization work in Amazon MSK?
If you are using IAM Access Control, Amazon MSK uses the policies you write and its own authorizer to authorize actions. If you are using TLS certificate authentication or SASL/SCRAM, Apache Kafka uses access control lists (ACLs) for authorization. To enable ACLs you must enable client authentication using either TLS certificates or SASL/SCRAM.

Q: How can I authenticate and authorize a client at the same time?
If you are using IAM Access Control, Amazon MSK will authenticate and authorize for you without any additional set up. If you are using TLS authentication, you can use the Dname of clients TLS certificates as the principal of the ACL to authorize client requests. If you are using SASL/SCRAM, you can use the username as the principal of the ACL to authorize client requests.

Q: How do I control service API actions?
You can control service API actions using AWS Identity and Access Management (IAM).

Q: Can I enable IAM Access Control for an existing cluster?
Yes, you can enable IAM Access Control for an existing cluster from the AWS console or by using the UpdateSecurity API.

Q: Can I use IAM Access Control outside of Amazon MSK?
No, IAM Access Control is only available for Amazon MSK clusters.

Monitoring, metrics, logging, and tagging

Q: How do I monitor the performance of my clusters or topics?
You can monitor the performance of your clusters using the Amazon MSK console, Amazon CloudWatch console, or via JMX and host metrics using Open Monitoring with Prometheus, an open-source monitoring solution.

Q: What is the cost for the different CloudWatch monitoring levels?
The cost of monitoring your cluster using Amazon CloudWatch is dependent on the monitoring

level and the size of your Apache Kafka cluster. Amazon CloudWatch charges per metric per month and includes a free tier. For details on the number of metrics exposed for each monitoring level, see Amazon MSK monitoring documentation.

Q: What monitoring tools are compatible with Open Monitoring with Prometheus?
Tools that are designed to read from Prometheus exporters are compatible with Open Monitoring, like: Datadog, Lenses, New Relic, Sumo Logic, or a Prometheus server. For details on Open Monitoring, see Amazon MSK Open Monitoring documentation.

Q: How do I monitor the health and performance of clients?
You can use any client-side monitoring supported by the Apache Kafka version you are using.

Q: Can I tag Amazon MSK resources?
Yes, you can tag Amazon MSK clusters from the AWS Command Line Interface (AWS CLI) or Console.

Q: How do I monitor consumer lag?
Topic-level consumer lag metrics are available as part of the default set of metrics that Amazon MSK publishes to Amazon CloudWatch for all clusters. No additional setup is required to get these metrics. For provisioned clusters, you can also get partition-level consumer lag metrics (partition dimension). To do so, enable enhanced monitoring (PER_PARTITION_PER_TOPIC) on your cluster. Alternatively, you can enable Open Monitoring on your cluster, and use a Prometheus server, to capture partition level metrics from the brokers in your cluster. Consumer lag metrics are available at port 11001, just as other Kafka metrics.

Q: How much does it cost to publish the consumer lag metric to Amazon CloudWatch?
Topic level metrics are included in the default set of Amazon MSK metrics, which are free of charge. Partition level metrics are charged as per Amazon CloudWatch pricing.

Q: How do I access Apache Kafka broker Logs?
You can enable broker log delivery for provisioned clusters. You can deliver broker logs to Amazon CloudWatch Logs, Amazon Simple Storage Service (S3), and Amazon Kinesis Data Firehose. Kinesis Data Firehose supports Amazon OpenSearch Service among other destinations. To learn how to enable this feature, see the Amazon MSK Logging Documentation.

Q: What is the logging level for broker logs?
Amazon MSK provides INFO level logs for all brokers within a provisioned cluster.

Q: How do I access Apache ZooKeeper Logs?
You can request Apache ZooKeeper logs through a support ticket.

Q: Can I log the use of Apache Kafka resource APIs, like create topic?
Yes, if you use IAM Access Control, the use of Apache Kafka resource APIs is logged to AWS CloudTrail.

Apache ZooKeeper

Q: What is Apache ZooKeeper?
From https://zookeeper.apache.org: “Apache ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications,” including Apache Kafka.

Q: Does Amazon MSK use Apache ZooKeeper?
Yes, Amazon MSK uses Apache ZooKeeper and manages Apache ZooKeeper within each

cluster as a part of the Amazon MSK service. Apache ZooKeeper nodes are included with each cluster at no additional cost.

Integrations

Q: What AWS services does Amazon MSK integrate with?
Amazon MSK integrates with:

- Amazon Virtual Private Cloud (Amazon VPC) for network isolation and security
- Amazon CloudWatch for metrics
- AWS Key Management Service (AWS KMS) for storage volume encryption
- Amazon IAM for authentication and authorization of Apache Kafka and service APIs
- AWS Lambda for MSK event sourcing
- AWS IoT for IoT event sourcing
- AWS Glue Schema Registry for controlling the evolution of schemas used by Apache Kafka applications
- AWS CloudTrail for AWS API logs
- AWS Certificate Manager for Private CAs used for client TLS authentication
- AWS CloudFormation for describing and provisioning Amazon MSK clusters using code
- Amazon Kinesis Data Analytics for fully managed Apache Flink applications that process streaming data
- Amazon Kinesis Data Analytics Studio for interactive streaming SQL on Apache Kafka
- AWS Secrets Manager for client credentials used for SASL/SCRAM authentication

Amazon MSK Serverless integrates with:

- Amazon VPC for network isolation and security
- Amazon CloudWatch for metrics
- Amazon IAM for authentication and authorization of Apache Kafka and service APIs
- AWS Glue Schema Registry for controlling the evolution of schemas used by Apache Kafka applications
- AWS CloudTrail for AWS API logs
- AWS PrivateLink for private connectivity

Scaling

Q: How can I scale up storage in my cluster?
You can scale up storage in your provisioned clusters using the AWS Management Console or the AWS CLI. You can also use tiered storage to virtually store unlimited data on your cluster without having to add brokers for storage. In serverless clusters, storage is scaled seamlessly based on your usage.

Q: How does tiered storage work?
Apache Kafka stores data in files called log segments. As each segment is complete, based on the size configured at cluster or topic level, it is copied to the low-cost storage tier. Data is held

in performance optimized storage for a specified retention time, or size, and then deleted. There is a separate time and size limit setting for the low-cost storage, which will be longer than the primary storage tier. If clients request data from segments stored in the low-cost tier, the broker will read the data from it and serve the data in the same way as if it is being served from the primary storage.

Q: How can I automatically expand storage in my cluster?
You can create an auto-scaling storage policy using the AWS Management Console or by creating an AWS Application Auto Scaling policy using the AWS CLI or APIs.

Q: Can I scale the number of brokers in an existing cluster?
Yes. You can scale out, or increase, the number of brokers for provisioned Amazon MSK clusters.

Q: Can I scale the broker size in an existing cluster?
Yes. You can choose to scale to a smaller or larger broker type on your provisioned Amazon MSK clusters.

Q: How do I balance partitions across brokers?
You can use Cruise Control for automatically rebalancing partitions to manage I/O heat. See the Cruise Control documentation for more information. Alternatively, you can use the Kafka Admin API kafka-reassign-partitions.sh to reassign partitions across brokers. In serverless clusters, Amazon MSK automatically balances partitions.

Compliance

Q: What compliance programs are in scope for Amazon MSK?
Amazon MSK is compliant or eligible for the following programs:

- HIPAA eligible
- PCI
- ISO
- SOC 1,2,3

AWS Glue, AWS Glue DataBrew, and AWS Glue Elastic Views

AWS Glue

Discover, prepare, and integrate all your data at any scale.

AWS Glue DataBrew

AWS Glue Elastic Views

AWS Glue

How it works

AWS Glue is a serverless data integration service that makes it easier to discover, prepare, move, and integrate data from multiple sources for analytics, machine learning (ML), and application development.

Event-driven ETL
AWS Glue can run your extract, transform, and load (ETL) jobs as new data arrives. For example, you can configure AWS Glue to initiate your ETL jobs to run as soon as new data becomes available in Amazon Simple Storage Service (S3).

AWS Glue Data Catalog

You can use the Data Catalog to quickly discover and search multiple AWS datasets without moving the data. Once the data is cataloged, it is immediately available for search and query using Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.

No-code ETL jobs

AWS Glue Studio makes it easier to visually create, run, and monitor AWS Glue ETL jobs. You can build ETL jobs that move and transform data using a drag-and-drop editor, and AWS Glue automatically generates the code.

Self-service data preparation

With AWS Glue DataBrew, you can explore and experiment with data directly from your data lake, data warehouses, and databases, including Amazon S3, Amazon Redshift, AWS Lake Formation, Amazon Aurora, and Amazon Relational Database Service (RDS). You can choose from over 250 prebuilt transformations in DataBrew to automate data preparation tasks such as filtering anomalies, standardizing formats, and correcting invalid values.

Why AWS Glue?

Preparing your data to obtain quality results is the first step in an analytics or ML project. AWS Glue is a serverless data integration service that makes data preparation simpler, faster, and cheaper. You can discover and connect to over 70 diverse data sources, manage your data in a centralized data catalog, and visually create, run, and monitor ETL pipelines to load data into your data lakes.

Use cases

Simplify ETL pipeline development
Remove infrastructure management with automatic provisioning and worker management, and consolidate all your data integration needs into a single service.

Discover data efficiently
Quickly identify data across multiple AWS datasets, and then make it instantly available for querying and transforming.

Interactively explore, experiment on, and process data
Using AWS Glue interactive sessions, data engineers can interactively explore and prepare data using the integrated development environment (IDE) or notebook of their choice.

Support various processing frameworks and workloads
More easily support various data processing frameworks, such as ETL and ELT, and various workloads, including batch, micro-batch, and streaming.

What is AWS Glue?

AWS Glue is a serverless data integration service that makes it easy for analytics users to discover, prepare, move, and integrate data from multiple sources. You can

use it for analytics, machine learning, and application development. It also includes additional productivity and data ops tooling for authoring, running jobs, and implementing business workflows.

With AWS Glue, you can discover and connect to more than 70 diverse data sources and manage your data in a centralized data catalog. You can visually create, run, and monitor extract, transform, and load (ETL) pipelines to load data into your data lakes. Also, you can immediately search and query cataloged data using Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.

AWS Glue consolidates major data integration capabilities into a single service. These include data discovery, modern ETL, cleansing, transforming, and centralized cataloging. It's also serverless, which means there's no infrastructure to manage. With flexible support for all workloads like ETL, ELT, and streaming in one service, AWS Glue supports users across various workloads and types of users.

Also, AWS Glue makes it easy to integrate data across your architecture. It integrates with AWS analytics services and Amazon S3 data lakes. AWS Glue has integration interfaces and job-authoring tools that are easy to use for all users, from developers to business users, with tailored solutions for varied technical skill sets. With the ability to scale on demand, AWS Glue helps you focus on high-value activities that maximize the value of your data. It scales for any data size, and supports all data types and schema variances. To increase agility and optimize costs, AWS Glue provides built-in high availability and pay-as-you-go billing.

AWS Glue Studio

AWS Glue Studio is a graphical interface that makes it easy to create, run, and monitor data integration jobs in AWS Glue. You can visually compose data transformation workflows and seamlessly run them on the Apache Spark–based serverless ETL engine in AWS Glue. For more information, see [What is AWS Glue Studio](#).

With AWS Glue Studio, you can create and manage jobs that gather, transform, and clean data. You can also use AWS Glue Studio to troubleshoot and edit job scripts.

AWS Glue features

AWS Glue features fall into three major categories:

- Discover and organize data
- Transform, prepare, and clean data for analysis
- Build and monitor data pipelines

Discover and organize data

- **Unify and search across multiple data stores** – Store, index, and search across multiple data sources and sinks by cataloging all your data in AWS.
- **Automatically discover data** – Use AWS Glue crawlers to automatically infer schema information and integrate it into your AWS Glue Data Catalog.
- **Manage schemas and permissions** – Validate and control access to your databases and tables.
- **Connect to a wide variety of data sources** – Tap into multiple data sources, both on premises and on AWS, using AWS Glue connections to build your data lake.

Transform, prepare, and clean data for analysis

- **Visually transform data with a drag-and-drop interface** – Define your ETL process in the drag-and-drop job editor and automatically generate the code to extract, transform, and load your data.
- **Build complex ETL pipelines with simple job scheduling** – Invoke AWS Glue jobs on a schedule, on demand, or based on an event.
- **Clean and transform streaming data in transit** – Enable continuous data consumption, and clean and transform it in transit. This makes it available for analysis in seconds in your target data store.
- **Deduplicate and cleanse data with built-in machine learning** – Clean and prepare your data for analysis without becoming a machine learning expert by using the FindMatches feature. This feature deduplicates and finds records that are imperfect matches for each other.
- **Built-in job notebooks** – AWS Glue Studio job notebooks provide serverless notebooks with minimal setup in AWS Glue Studio so you can get started quickly.
- **Edit, debug, and test ETL code** – With AWS Glue interactive sessions, you can interactively explore and prepare data. You can explore, experiment on, and process data interactively using the IDE or notebook of your choice.
- **Define, detect, and remediate sensitive data** – AWS Glue sensitive data detection lets you define, identify, and process sensitive data in your data pipeline and in your data lake.

Build and monitor data pipelines

- **Automatically scale based on workload** – Dynamically scale resources up and down based on workload. This assigns workers to jobs only when needed.

- **Automate jobs with event-based triggers** – Start crawlers or AWS Glue jobs with event-based triggers, and design a chain of dependent jobs and crawlers.
- **Run and monitor jobs** – Run your AWS Glue jobs, and then monitor them with automated monitoring tools, the Apache Spark UI, AWS Glue job run insights, and AWS CloudTrail.
- **Define workflows for ETL and integration activities** – Define workflows for ETL and integration activities for multiple crawlers, jobs, and triggers.

Accessing AWS Glue

You can create, view, and manage your AWS Glue jobs using the following interfaces:

- **AWS Glue console** – Provides a web interface for you to create, view, and manage your AWS Glue jobs. To access the console, see AWS Glue console.
- **AWS Glue Studio** – Provides a graphical interface for you to create and edit your AWS Glue jobs visually. For more information, see What is AWS Glue Studio.
- **AWS Glue section of the AWS CLI Reference** – Provides AWS CLI commands that you can use with AWS Glue. For more information, see AWS CLI Reference for AWS Glue.
- **AWS Glue API** – Provides a complete API reference for developers. For more information, see AWS Glue API.

Related services

Users of AWS Glue also use:

- **AWS Lake Formation** – A service that is an authorization layer that provides fine-grained access control to resources in the AWS Glue Data Catalog.
- **AWS Glue Data Brew** – A visual data preparation tool that you can use to clean and normalize data without writing any code.

AWS Glue DataBrew

AWS Glue DataBrew is a new visual data preparation tool that makes it easy for data analysts and data scientists to clean and normalize data to prepare it for analytics and machine learning. You can choose from over 250 pre-built transformations to automate data preparation tasks, all without the need to write

any code. You can automate filtering anomalies, converting data to standard formats, and correcting invalid values, and other tasks. After your data is ready, you can immediately use it for analytics and machine learning projects. You only pay for what you use - no upfront commitment.

What is AWS Glue DataBrew?

AWS Glue DataBrew is a visual data preparation tool that enables users to clean and normalize data without writing any code. Using DataBrew helps reduce the time it takes to prepare data for analytics and machine learning (ML) by up to 80 percent, compared to custom developed data preparation. You can choose from over 250 ready-made transformations to automate data preparation tasks, such as filtering anomalies, converting data to standard formats, and correcting invalid values.

Using DataBrew, business analysts, data scientists, and data engineers can more easily collaborate to get insights from raw data. Because DataBrew is serverless, no matter what your technical level, you can explore and transform terabytes of raw data without needing to create clusters or manage any infrastructure.

With the intuitive DataBrew interface, you can interactively discover, visualize, clean, and transform raw data. DataBrew makes smart suggestions to help you identify data quality issues that can be difficult to find and time-consuming to fix. With DataBrew preparing your data, you can use your time to act on the results and iterate more quickly. You can save transformation as steps in a recipe, which you can update or reuse later with other datasets, and deploy on a continuing basis.

The following image shows how DataBrew works at a high level.

To use DataBrew, you create a project and connect to your data. In the project workspace, you see your data displayed in a grid-like visual interface. Here, you can explore the data and see value distributions and charts to understand its profile.

To prepare the data, you can choose from more than 250 point-and-click transformations. These include removing nulls, replacing missing values, fixing schema inconsistencies, creating columns based on functions, and many more. You can also use transformations to apply natural language processing (NLP) techniques to split sentences into phrases. Immediate previews show a portion of your data before and after transformation, so you can modify your recipe before applying it to the entire dataset.

After DataBrew has run your recipe on your dataset, the output is stored in Amazon Simple Storage Service (Amazon S3). After your cleansed, prepared dataset is in Amazon S3, another of your data storage or data management systems can ingest it.

AWS Glue Elastic Views

AWS Glue Elastic Views is a new capability of AWS Glue that makes it easy to build materialized views that combine and replicate data across multiple data stores without you having to write custom code. With AWS Glue Elastic Views, you can use familiar Structured Query Language (SQL) to quickly create a virtual table—a materialized view—from multiple different source data stores. AWS Glue Elastic Views copies data from each source data store and creates a replica in a target data store. AWS Glue Elastic Views continuously monitors for changes to data in your source data stores, and provides updates to the materialized views in your target data stores automatically, ensuring data accessed through the materialized view is always up-to-date.

AWS Glue Elastic Views supports many AWS databases and data stores, including Amazon DynamoDB, Amazon S3, Amazon Redshift, and Amazon Elasticsearch Service, with support for Amazon Relational Database Service, Amazon Aurora, and others to follow. AWS Glue Elastic Views is serverless and scales capacity up or down automatically based on demand, so there's no infrastructure to manage.

Amazon Glue FAQs

General

Q: What is Amazon Glue?

Amazon Glue is a serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development. Amazon Glue provides all of the capabilities needed for data integration, so you can start analyzing your data and putting it to use in minutes instead of months. AMAZON Glue provides both visual and code-based interfaces to make data integration easier. Users can easily find and access data using the AMAZON Glue Data Catalog. Data engineers and ETL (extract, transform, and load)

developers can create and run ETL workflows. Data analysts and data scientists can use AMAZON Glue DataBrew to visually enrich, clean, and normalize data without writing code.

Q. What are the main components of Amazon Glue?

Amazon Glue consists of a Data Catalog which is a central metadata repository; an ETL engine that can automatically generate Scala or Python code; a flexible scheduler that handles dependency resolution, job monitoring, and retries; and Amazon Glue DataBrew for cleaning and normalizing data with a visual interface. Together, these automate much of the undifferentiated heavy lifting involved with discovering, categorizing, cleaning, enriching, and moving data, so you can spend more time analyzing your data.

Q: When should I use Amazon Glue?

You should use Amazon Glue to discover properties of the data you own, transform it, and prepare it for analytics. Glue can automatically discover both structured and semi-structured data stored in your data lake on Amazon S3, data warehouse in Amazon Redshift, and various databases running on Amazon. It provides a unified view of your data via the Glue Data Catalog that is available for ETL, querying and reporting using services like Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. Glue automatically generates Scala or Python code for your ETL jobs that you can further customize using tools you are already familiar with. You can use Amazon Glue DataBrew to visually clean up and normalize data without writing code. Amazon Glue is serverless, so there are no compute resources to configure and manage.

Q: How does Amazon Glue relate to Amazon Lake Formation?

A: Lake Formation leverages a shared infrastructure with AMAZON Glue, including console controls, ETL code creation and job monitoring, a common data catalog, and a serverless architecture. While Amazon Glue is still focused on these types of functions, Lake Formation encompasses Amazon Glue features AND provides additional capabilities designed to help build, secure, and manage a data lake. See the Amazon Lake Formation pages for more details.

Amazon Glue Data Catalog

Q: What is the Amazon Glue Data Catalog?

The Amazon Glue Data Catalog is a central repository to store structural and operational metadata for all your data assets. For a given data set, you can store its table definition, physical location, add business relevant attributes, as well as track how this data has changed over time.

The Amazon Glue Data Catalog is Apache Hive Metastore compatible and is a drop-in replacement for the Apache Hive Metastore for Big Data applications running on Amazon EMR. For more information on setting up your EMR cluster to use Amazon Glue Data Catalog as an Apache Hive Metastore, click here.

The Amazon Glue Data Catalog also provides out-of-box integration with Amazon EMR, and Amazon Redshift Spectrum. Once you add your table definitions to the Glue Data Catalog, they are available for ETL and also readily available for querying in Amazon EMR, and Amazon Redshift Spectrum so that you can have a common view of your data between these services.

Q: How do I get my metadata into the Amazon Glue Data Catalog?

Amazon Glue provides a number of ways to populate metadata into the Amazon Glue Data Catalog. Glue crawlers scan various data stores you own to automatically infer schemas and

partition structure and populate the Glue Data Catalog with corresponding table definitions and statistics. You can also schedule crawlers to run periodically so that your metadata is always up-to-date and in-sync with the underlying data. Alternately, you can add and update table details manually by using the Amazon Glue Console or by calling the API. You can also run Hive DDL statements via a Hive client on an Amazon EMR cluster. Finally, if you already have a persistent Apache Hive Metastore, you can perform a bulk import of that metadata into the Amazon Glue Data Catalog by using our import script.

Q: What are Amazon Glue crawlers?

An Amazon Glue crawler connects to a data store, progresses through a prioritized list of classifiers to extract the schema of your data and other statistics, and then populates the Glue Data Catalog with this metadata. Crawlers can run periodically to detect the availability of new data as well as changes to existing data, including table definition changes. Crawlers automatically add new tables, new partitions to existing table, and new versions of table definitions. You can customize Glue crawlers to classify your own file types.

Q: How do I import data from my existing Apache Hive Metastore to the Amazon Glue Data Catalog?

You simply run an ETL job that reads from your Apache Hive Metastore, exports the data to an intermediate format in Amazon S3, and then imports that data into the Amazon Glue Data Catalog.

Q: Do I need to maintain my Apache Hive Metastore if I am storing my metadata in the Amazon Glue Data Catalog?

No. Amazon Glue Data Catalog is Apache Hive Metastore compatible. You can point to the Glue Data Catalog endpoint and use it as an Apache Hive Metastore replacement. For more information on how to configure your cluster to use Amazon Glue Data Catalog as an Apache Hive Metastore, please read our documentation here.

Amazon Glue Schema Registry

Q: What is the Amazon Glue Schema Registry?

Amazon Glue Schema Registry, a serverless feature of Amazon Glue, enables you to validate and control the evolution of streaming data using schemas registered in Apache Avro and JSON Schema data formats, at no additional charge. Through Apache-licensed serializers and deserializers, the Schema Registry integrates with Java applications developed for Apache Kafka, Amazon Managed Streaming for Apache Kafka (MSK), Amazon Kinesis Data Streams, Apache Flink, Amazon Kinesis Data Analytics for Apache Flink, and Amazon Lambda. When data streaming applications are integrated with the Schema Registry, you can improve data quality and safeguard against unexpected changes using compatibility checks that govern schema evolution. Additionally, you can create or update Amazon Glue tables and partitions using Apache Avro schemas stored within the registry.

Q: Why should I use Amazon Glue Schema Registry?

With the Amazon Glue Schema Registry, you can:

1. Validate schemas. When data streaming applications are integrated with Amazon Glue Schema Registry, schemas used for data production are validated against schemas within a central registry, allowing you to centrally control data quality.
2. Safeguard schema evolution. You can set rules on how schemas can and cannot evolve using one of eight compatibility modes.
3. Improve data quality. Serializers validate schemas used by data producers against those stored in the registry, improving data quality when it originates and reducing downstream issues from unexpected schema drift.
4. Save costs. Serializers convert data into a binary format and can compress it before it is delivered, reducing data transfer and storage costs.
5. Improve processing efficiency. In many cases, a data stream contains records of different schemas. The Schema Registry enables applications that read from data streams to selectively process each record based on the schema without having to parse its contents, which increases processing efficiency.

Q: What data format, client language, and integrations are supported by Amazon Glue Schema Registry?

The Schema Registry supports Apache Avro and JSON Schema data formats and Java client applications. We plan to continue expanding support for other data formats and non-Java clients. The Schema Registry integrates with applications developed for Apache Kafka, Amazon Managed Streaming for Apache Kafka (MSK), Amazon Kinesis Data Streams, Apache Flink, Amazon Kinesis Data Analytics for Apache Flink, and Amazon Lambda.

Q: What kinds of evolution rules does Amazon Glue Schema Registry support?

The following compatibility modes are available for you to manage your schema evolution: Backward, Backward All, Forward, Forward All, Full, Full All, None, and Disabled. Visit the Schema Registry user documentation to learn more about compatibility rules.

Q: How does Amazon Glue Schema Registry maintain high availability for my applications?

The Schema Registry storage and control plane is designed for high availability and is backed by the Amazon Glue SLA, and the serializers and deserializers leverage best-practice caching techniques to maximize schema availability within clients.

Q: Is Amazon Glue Schema Registry open-source?

Amazon Glue Schema Registry storage is an Amazon Web Services service, while the serializers and deserializers are Apache-licensed open-source components.

Q: Does Amazon Glue Schema Registry provide encryption at rest and in-transit?

Yes, your clients communicate with the Schema Registry via API calls which encrypt data in-transit using TLS encryption over HTTPS. Schemas stored in the Schema Registry are always encrypted at rest using a service-managed KMS key.

Q: How can I privately connect to Amazon Glue Schema Registry?

You can use Amazon PrivateLink to connect your data producer's VPC to Amazon Glue by defining an interface VPC endpoint for Amazon Glue. When you use a VPC interface endpoint, communication between your VPC and Amazon Glue is conducted entirely within the Amazon Web Services network. For more information, please visit the user documentation.

Q: How can I monitor my Amazon Glue Schema Registry usage?

Amazon CloudWatch metrics are available as part of CloudWatch's free tier. You can access these metrics in the CloudWatch Console. Visit the Amazon Glue Schema Registry user documentation for more information.

Q: Does Amazon Glue Schema Registry provide tools to manage user authorization?

Yes, the Schema Registry supports both resource-level permissions and identity-based IAM policies.

Q: How do I migrate from an existing schema registry to the Amazon Glue Schema Registry?

Steps to migrate from a third-party schema registry to Amazon Glue Schema Registry are available in the user documentation.

Extract, transform, and load (ETL)

Q: What programming language can I use to write my ETL code for Amazon Glue?

You can use either Scala or Python.

Q: How can I customize the ETL code generated by Amazon Glue?

Amazon Glue's ETL script recommendation system generates Scala or Python code. It leverages Glue's custom ETL library to simplify access to data sources as well as manage job execution. You can find more details about the library in our documentation. You can write ETL code using Amazon Glue's custom library or write arbitrary code in Scala or Python by using inline editing via the Amazon Glue Console script editor, downloading the auto-generated code, and editing it in your own IDE. You can also start with one of the many samples hosted in our Github repository and customize that code.

Q: Can I import custom libraries as part of my ETL script?

Yes. You can import custom Python libraries and Jar files into your Amazon Glue ETL job. For more details, please check our documentation here.

Q: Can I bring my own code?

Yes. You can write your own code using Amazon Glue's ETL library, or write your own Scala or Python code and upload it to a Glue ETL job. For more details, please check our documentation here.

Q: How can I develop my ETL code using my own IDE?

You can create and connect to development endpoints that offer ways to connect your notebooks and IDEs.

Q: How can I build end-to-end ETL workflow using multiple jobs in Amazon Glue?

In addition to the ETL library and code generation, Amazon Glue provides a robust set of orchestration features that allow you to manage dependencies between multiple jobs to build

end-to-end ETL workflows. Amazon Glue ETL jobs can either be triggered on a schedule or on a job completion event. Multiple jobs can be triggered in parallel or sequentially by triggering them on a job completion event. You can also trigger one or more Glue jobs from an external source such as an Amazon Lambda function.

Q: How does Amazon Glue monitor dependencies?

Amazon Glue manages dependencies between two or more jobs or dependencies on external events using triggers. Triggers can watch one or more jobs as well as invoke one or more jobs. You can either have a scheduled trigger that invokes jobs periodically, an on-demand trigger, or a job completion trigger.

Q: How does Amazon Glue handle errors?

Amazon Glue monitors job event metrics and errors, and pushes all notifications to Amazon CloudWatch. With Amazon CloudWatch, you can configure a host of actions that can be triggered based on specific notifications from Amazon Glue. For example, if you get an error or a success notification from Glue, you can trigger an Amazon Lambda function. Glue also provides default retry behavior that will retry all failures three times before sending out an error notification.

Q: Can I run my existing ETL jobs with Amazon Glue?

Yes. You can run your existing Scala or Python code on Amazon Glue. Simply upload the code to Amazon S3 and create one or more jobs that use that code. You can reuse the same code across multiple jobs by pointing them to the same code location on Amazon S3.

Q: How can I use Amazon Glue to ETL streaming data?

Amazon Glue ETL is batch oriented, and you can schedule your ETL jobs at a minimum of 5 min intervals. While it can process micro-batches, it does not handle streaming data. If your use case requires you to ETL data while you stream it in, you can perform the first leg of your ETL using Amazon Kinesis Data Firehose, and then store data to either Amazon S3 or Amazon Redshift and trigger a Glue ETL job to pick up that dataset and continue applying additional transformations to that data.

Q: Do I have to use both Amazon Glue Data Catalog and Glue ETL to use the service?

No. While we do believe that using both the Amazon Glue Data Catalog and ETL provides an end-to-end ETL experience, you can use either one of them independently without using the other.

Amazon Glue DataBrew

Q: What is Amazon Glue DataBrew?

Amazon Glue DataBrew is a visual data preparation tool that makes it easy for data analysts and data scientists to prepare data with an interactive, point-and-click visual interface without writing code. With Glue DataBrew, you can easily visualize, clean, and normalize terabytes, and even petabytes of data directly from your data lake, data warehouses, and databases, including

Amazon S3, Amazon Redshift, Amazon Aurora, and Amazon RDS. Amazon Glue DataBrew is generally available in the regions where Amazon Glue is available .

Q: Who can use Amazon Glue DataBrew?

Amazon Glue DataBrew is built for users who need to clean and normalize data for analytics and machine learning. Data analysts and data scientists are the primary users. For data analysts, examples of job functions are business intelligence analysts, operations analysts, market intelligence analysts, legal analysts, financial analysts, economists, quants, or accountants. For data scientists, examples of job functions are materials scientists, bioanalytical scientists, and scientific researchers.

Q: What types of transformations are supported in Amazon Glue DataBrew?

You can choose from over 250 built-in transformations to combine, pivot, and transpose the data without writing code. Amazon Glue DataBrew also automatically recommends transformations such as filtering anomalies, correcting invalid, incorrectly classified, or duplicate data, normalizing data to standard date and time values, or generating aggregates for analyses. For complex transformations, such as converting words to a common base or root word, Glue DataBrew provides transformations that use advanced machine learning techniques such as Natural Language Processing (NLP). You can group multiple transformations together, save them as recipes, and apply the recipes directly to the new incoming data.

Q: What file formats does Amazon Glue DataBrew support?

For input data, Amazon Glue DataBrew supports commonly used file formats, such as comma-separated values (.csv), JSON and nested JSON, Apache Parquet and nested Apache Parquet, and Excel sheets. For output data, Amazon Glue DataBrew supports comma-separated values (.csv), JSON, Apache Parquet, Apache Avro, Apache ORC and XML.

Q: Can I try Amazon Glue DataBrew for free?

Yes. Sign up for an Amazon Free Tier account, then visit the Amazon Glue DataBrew Management Console, and get started instantly for free. If you are a first-time user of Glue DataBrew, the first 40 interactive sessions are free.

Q: Do I need to use Amazon Glue Data Catalog or Amazon Lake Formation to use Amazon Glue DataBrew?

No. You can use Amazon Glue DataBrew without using either Amazon Glue Data Catalog or Amazon Lake Formation. If you use Glue Data Catalog to store schema and metadata, Glue DataBrew automatically infers schema from the Glue Data Catalog. If your data is centralized and secured in Amazon Lake Formation, DataBrew users can use all data sets available to them from its centralized data catalog.

Q: Can I retain a record of all changes made to my data?

Yes. You can visually track all the changes made to your data in the Amazon Glue DataBrew Management Console. The visual view makes it easy to trace the changes and relationships made to the datasets, projects and recipes, and all other associated jobs. In addition, Glue DataBrew keeps all account activities as logs in the Amazon CloudTrail.

Amazon Web Services Product Integrations

Q: When should I use Amazon Glue vs. Amazon EMR?

Amazon Glue works on top of the Apache Spark environment to provide a scale-out execution environment for your data transformation jobs. Amazon Glue infers, evolves, and monitors your ETL jobs to greatly simplify the process of creating and maintaining jobs. Amazon EMR provides you with direct access to your Hadoop environment, affording you lower-level access and greater flexibility in using tools beyond Spark.

Q: When should I use Amazon Glue vs Amazon Database Migration Service?

Amazon Database Migration Service (DMS) helps you migrate databases to Amazon Web Services easily and securely. For use cases which require a database migration from on-premises to Amazon Web Services or database replication between on-premises sources and sources on Amazon Web Services, we recommend you use Amazon DMS. Once your data is in Amazon Web Services, you can use Amazon Glue to move and transform data from your data source into another database or data warehouse, such as Amazon Redshift.

Security and availability

Q: How does Amazon Glue keep my data secure?

We provide server side encryption for data at rest and SSL for data in motion.

Q: What are the service limits associated with Amazon Glue?

Please refer our documentation to learn more about service limits.

Q: How many DPUs (Data Processing Units) are allocated to the development endpoint?

A development endpoint is provisioned with 5 DPUs by default. You can configure a development endpoint with a minimum of 2 DPUs and a maximum of 5 DPUs.

Q: How do I scale the size and performance of my Amazon Glue ETL jobs?

You can simply specify the number of DPUs (Data Processing Units) you want to allocate to your ETL job. A Glue ETL job requires a minimum of 2 DPUs. By default, Amazon Glue allocates 10 DPUs to each ETL job.

Q: How do I monitor the execution of my Amazon Glue jobs?

The Amazon Glue provides status of each job and pushes all notifications to Amazon CloudWatch. You can setup SNS notifications via CloudWatch actions to be informed of job failures or completions.

AWS Glue Data Catalog

Q: What is the AWS Glue Data Catalog?

The AWS Glue Data Catalog is a central repository to store structural and operational metadata for all your data assets. For a given data set, you can store its table definition, physical location, add business relevant attributes, as well as track how this data has changed over time.

The AWS Glue Data Catalog is Apache Hive Metastore compatible and is a drop-in replacement for the Apache Hive Metastore for Big Data applications running on Amazon EMR. For more

information on setting up your EMR cluster to use AWS Glue Data Catalog as an Apache Hive Metastore, click here.

The AWS Glue Data Catalog also provides out-of-box integration with Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum. Once you add your table definitions to the Glue Data Catalog, they are available for ETL and also readily available for querying in Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum so that you can have a common view of your data between these services.

Q: How do I get my metadata into the AWS Glue Data Catalog?

AWS Glue provides a number of ways to populate metadata into the AWS Glue Data Catalog. Glue crawlers scan various data stores you own to automatically infer schemas and partition structure and populate the Glue Data Catalog with corresponding table definitions and statistics. You can also schedule crawlers to run periodically so that your metadata is always up-to-date and in-sync with the underlying data. Alternately, you can add and update table details manually by using the AWS Glue Console or by calling the API. You can also run Hive DDL statements via the Amazon Athena Console or a Hive client on an Amazon EMR cluster. Finally, if you already have a persistent Apache Hive Metastore, you can perform a bulk import of that metadata into the AWS Glue Data Catalog by using our import script.

Q: What are AWS Glue crawlers?

An AWS Glue crawler connects to a data store, progresses through a prioritized list of classifiers to extract the schema of your data and other statistics, and then populates the Glue Data Catalog with this metadata. Crawlers can run periodically to detect the availability of new data as well as changes to existing data, including table definition changes. Crawlers automatically add new tables, new partitions to existing table, and new versions of table definitions. You can customize Glue crawlers to classify your own file types.

Q: How do I import data from my existing Apache Hive Metastore to the AWS Glue Data Catalog?

You simply run an ETL job that reads from your Apache Hive Metastore, exports the data to an intermediate format in Amazon S3, and then imports that data into the AWS Glue Data Catalog.

Q: Do I need to maintain my Apache Hive Metastore if I am storing my metadata in the AWS Glue Data Catalog?

No. AWS Glue Data Catalog is Apache Hive Metastore compatible. You can point to the Glue Data Catalog endpoint and use it as an Apache Hive Metastore replacement. For more information on how to configure your cluster to use AWS Glue Data Catalog as an Apache Hive Metastore, please read our documentation here.

Q. If I am already using Amazon Athena or Amazon Redshift Spectrum and have tables in Amazon Athena's internal data catalog, how can I start using the AWS Glue Data Catalog as my common metadata repository?

Before you can start using AWS Glue Data Catalog as a common metadata repository between Amazon Athena, Amazon Redshift Spectrum, and AWS Glue, you must upgrade your

Amazon Athena data catalog to AWS Glue Data Catalog. The steps required for the upgrade are detailed here.

Q: What analytics services use the AWS Glue Data Catalog?

The metadata stored in the AWS Glue Data Catalog can be readily accessed from Glue ETL, Amazon Athena, Amazon EMR, Amazon Redshift Spectrum, and third-party services.

AWS Glue Schema Registry

Q: What is the AWS Glue Schema Registry?

AWS Glue Schema Registry, a serverless feature of AWS Glue, enables you to validate and control the evolution of streaming data using schemas registered in Apache Avro and JSON Schema data formats, at no additional charge. Through Apache-licensed serializers and deserializers, the Schema Registry integrates with Java applications developed for Apache Kafka, Amazon Managed Streaming for Apache Kafka (MSK), Amazon Kinesis Data Streams, Apache Flink, Amazon Kinesis Data Analytics for Apache Flink, and AWS Lambda. When data streaming applications are integrated with the Schema Registry, you can improve data quality and safeguard against unexpected changes using compatibility checks that govern schema evolution. Additionally, you can create or update AWS Glue tables and partitions using Apache Avro schemas stored within the registry.

Q: Why should I use AWS Glue Schema Registry?

With the AWS Glue Schema Registry, you can:

1. Validate schemas. When data streaming applications are integrated with AWS Glue Schema Registry, schemas used for data production are validated against schemas within a central registry, allowing you to centrally control data quality.
2. Safeguard schema evolution. You can set rules on how schemas can and cannot evolve using one of eight compatibility modes.
3. Improve data quality. Serializers validate schemas used by data producers against those stored in the registry, improving data quality when it originates and reducing downstream issues from unexpected schema drift.
4. Save costs. Serializers convert data into a binary format and can compress it before it is delivered, reducing data transfer and storage costs.
5. Improve processing efficiency. In many cases, a data stream contains records of different schemas. The Schema Registry enables applications that read from data streams to selectively process each record based on the schema without having to parse its contents, which increases processing efficiency.

Q: What data format, client language, and integrations are supported by AWS Glue Schema Registry?

The Schema Registry supports Apache Avro and JSON Schema data formats and Java client applications. We plan to continue expanding support for other data formats and non-Java clients. The Schema Registry integrates with applications developed for Apache Kafka, Amazon Managed Streaming for Apache Kafka (MSK), Amazon Kinesis Data Streams, Apache Flink, Amazon Kinesis Data Analytics for Apache Flink, and AWS Lambda.

Q: What kinds of evolution rules does AWS Glue Schema Registry support?

The following compatibility modes are available for you to manage your schema evolution: Backward, Backward All, Forward, Forward All, Full, Full All, None, and Disabled. Visit the Schema Registry user documentation to learn more about compatibility rules.

Q: How does AWS Glue Schema Registry maintain high availability for my applications?

The Schema Registry storage and control plane is designed for high availability and is backed by the AWS Glue SLA, and the serializers and deserializers leverage best-practice caching techniques to maximize schema availability within clients.

Q: Is AWS Glue Schema Registry open-source?

AWS Glue Schema Registry storage is an AWS service, while the serializers and deserializers are Apache-licensed open-source components.

Q: Does AWS Glue Schema Registry provide encryption at rest and in-transit?

Yes, your clients communicate with the Schema Registry via API calls which encrypt data in-transit using TLS encryption over HTTPS. Schemas stored in the Schema Registry are always encrypted at rest using a service-managed KMS key.

Q: How can I privately connect to AWS Glue Schema Registry?

You can use AWS PrivateLink to connect your data producer's VPC to AWS Glue by defining an interface VPC endpoint for AWS Glue. When you use a VPC interface endpoint, communication between your VPC and AWS Glue is conducted entirely within the AWS network. For more information, please visit the user documentation.

Q: How can I monitor my AWS Glue Schema Registry usage?

AWS CloudWatch metrics are available as part of CloudWatch's free tier. You can access these metrics in the CloudWatch Console. Visit the AWS Glue Schema Registry user documentation for more information.

Q: Does AWS Glue Schema Registry provide tools to manage user authorization?

Yes, the Schema Registry supports both resource-level permissions and identity-based IAM policies.

Q: How do I migrate from an existing schema registry to the AWS Glue Schema Registry?

Steps to migrate from a third-party schema registry to AWS Glue Schema Registry are available in the user documentation.

Extract, transform, and load (ETL)

Q: Does AWS Glue have a no-code interface for visual ETL?

Yes. AWS Glue Studio offers a graphical interface for authoring Glue jobs to process your data. After you define the flow of your data sources, transformations and targets in the visual interface, AWS Glue studio will generate Apache Spark code on your behalf.
Q: What programming language can I use to write my ETL code for AWS Glue?

You can use either Scala or Python.

Q: How can I customize the ETL code generated by AWS Glue?

AWS Glue's ETL script recommendation system generates Scala or Python code. It leverages Glue's custom ETL library to simplify access to data sources as well as manage job execution. You can find more details about the library in our documentation. You can write ETL code using AWS Glue's custom library or write arbitrary code in Scala or Python by using inline editing via the AWS Glue Console script editor, downloading the auto-generated code, and editing it in your own IDE. You can also start with one of the many samples hosted in our Github repository and customize that code.

Q: Can I import custom libraries as part of my ETL script?

Yes. You can import custom Python libraries and Jar files into your AWS Glue ETL job. For more details, please check our documentation here.

Q: Can I bring my own code?

Yes. You can write your own code using AWS Glue's ETL library, or write your own Scala or Python code and upload it to a Glue ETL job. For more details, please check our documentation here.

Q: How can I develop my ETL code using my own IDE?

You can create and connect to development endpoints that offer ways to connect your notebooks and IDEs.

Q: How can I build end-to-end ETL workflow using multiple jobs in AWS Glue?

In addition to the ETL library and code generation, AWS Glue provides a robust set of orchestration features that allow you to manage dependencies between multiple jobs to build end-to-end ETL workflows. AWS Glue ETL jobs can either be triggered on a schedule or on a job completion event. Multiple jobs can be triggered in parallel or sequentially by triggering them on a job completion event. You can also trigger one or more Glue jobs from an external source such as an AWS Lambda function.

Q: How does AWS Glue monitor dependencies?

AWS Glue manages dependencies between two or more jobs or dependencies on external events using triggers. Triggers can watch one or more jobs as well as invoke one or more jobs. You can either have a scheduled trigger that invokes jobs periodically, an on-demand trigger, or a job completion trigger.

Q: How does AWS Glue handle ETL errors?

AWS Glue monitors job event metrics and errors, and pushes all notifications to Amazon CloudWatch. With Amazon CloudWatch, you can configure a host of actions that can be triggered based on specific notifications from AWS Glue. For example, if you get an error or a success notification from Glue, you can trigger an AWS Lambda function. Glue also provides default retry behavior that will retry all failures three times before sending out an error notification.

Q: Can I run my existing ETL jobs with AWS Glue?

Yes. You can run your existing Scala or Python code on AWS Glue. Simply upload the code to Amazon S3 and create one or more jobs that use that code. You can reuse the same code across multiple jobs by pointing them to the same code location on Amazon S3.

Q: How can I use AWS Glue to ETL streaming data?

AWS Glue supports ETL on streams from Amazon Kinesis Data Streams, Apache Kafka, and Amazon MSK. Add the stream to the Glue Data Catalog and then choose it as the data source when setting up your AWS Glue job.

Q: Do I have to use both AWS Glue Data Catalog and Glue ETL to use the service?

No. While we do believe that using both the AWS Glue Data Catalog and ETL provides an end-to-end ETL experience, you can use either one of them independently without using the other.

Q: When should I use AWS Glue Streaming and when should I use Amazon Kinesis Data Analytics?

Both AWS Glue and Amazon Kinesis Data Analytics can be used to process streaming data. AWS Glue is recommended when your use cases are primarily ETL and when you want to run jobs on a serverless Apache Spark-based platform. Amazon Kinesis Data Analytics is recommended when your use cases are primarily analytics and when you want to run jobs on a serverless Apache Flink-based platform.

Streaming ETL in AWS Glue enables advanced ETL on streaming data using the same serverless, pay-as-you-go platform that you currently use for your batch jobs. AWS Glue generates customizable ETL code to prepare your data while in flight and has built-in functionality to process streaming data that is semi-structured or has an evolving schema. Use Glue to apply both its built-in and Spark-native transforms to data streams and load them into your data lake or data warehouse.

Amazon Kinesis Data Analytics enables you to build sophisticated streaming applications to analyze streaming data in real time. It provides a serverless Apache Flink runtime that automatically scales without servers and durably saves application state. Use Amazon Kinesis Data Analytics for real-time analytics and more general stream data processing.

Q: When should I use AWS Glue and when should I use Amazon Kinesis Data Firehose?

Both AWS Glue and Amazon Kinesis Data Firehose can be used for streaming ETL. AWS Glue is recommended for complex ETL, including joining streams, and partitioning the output in Amazon S3 based on the data content. Amazon Kinesis Data Firehose is recommended when your use cases focus on data delivery and preparing data to be processed after it is delivered.

Streaming ETL in AWS Glue enables advanced ETL on streaming data using the same serverless, pay-as-you-go platform that you currently use for your batch jobs. AWS Glue generates customizable ETL code to prepare your data while in flight and has built-in functionality to process streaming data that is semi-structured or has an evolving schema. Use Glue to apply complex transforms to data streams, enrich records with information from other streams and persistent data stores, and then load records into your data lake or data warehouse.

Streaming ETL in Amazon Kinesis Data Firehose enables you to easily capture, transform, and deliver streaming data. Amazon Kinesis Data Firehose provides ETL capabilities including

serverless data transformation through AWS Lambda and format conversion from JSON to Parquet. It includes ETL capabilities that are designed to make data easier to process after delivery, but does not include the advanced ETL capabilities that AWS Glue supports.

Deduplicate data

Q: What kind of problems does the FindMatches ML Transform solve?

FindMatches generally solves Record Linkage and Data Deduplication problems. Deduplication is what you have to do when you are trying to identify records in a database which are conceptually “the same”, but for which you have separate records. This problem is trivial if duplicate records can be identified by a unique key (for instance if products can be uniquely identified by a UPC Code), but becomes very challenging when you have to do a “fuzzy match”.

Record linkage is basically the same problem as data deduplication under the hood, but this term usually means that you are doing a “fuzzy join” of two databases that do not share a unique key rather than deduplicating a single database. As an example, consider the problem of matching a large database of customers to a small database of known fraudsters. FindMatches can be used on both record linkage and deduplication problems.

For instance, AWS Glue's FindMatches ML Transform can help you with the following problems:

Linking patient records between hospitals so that doctors have more background information and are better able to treat patients by using FindMatches on separate databases that both contain common fields such as name, birthday, home address, phone number, etc.

Deduplicating a database of movies containing columns like “title”, “plot synopsis”, “year of release”, “run time”, and “cast”. For instance, the same movie might be variously identified as “Star Wars”, “Star Wars: A New Hope”, and “Star Wars: Episode IV—A New Hope (Special Edition)”.

Automatically group all related products together in your storefront by identifying equivalent items in an apparel product catalog where you want to define “equivalent” to mean that they are the same ignoring differences in size and color. Hence “Levi 501 Blue Jeans, size 34x34” is defined to be the same as “Levi 501 Jeans--black, Size 32x31”.

Q: How does AWS Glue deduplicate my data?

AWS Glue's FindMatches ML Transform makes it easy to find and link records that refer to the same entity but don’t share a reliable identifier. Before FindMatches, developers would commonly solve data-matching problems deterministically, by writing huge numbers of hand-tuned rules. FindMatches uses machine learning algorithms behind the scenes to learn how to match records according to each developer's own business criteria. FindMatches first identifies records for the customer to label as to whether they match or do not match and then uses machine learning to create an ML Transform. Customers can then execute this Transform on their database to find matching records or they can ask FindMatches to give them additional records to label to push their ML Transform to higher levels of accuracy.

Q: What are ML Transforms?

ML Transforms provide a destination for creating and managing machine-learned transforms. Once created and trained, these ML Transforms can then be executed in standard AWS Glue scripts. Customers select a particular algorithm (for example, the FindMatches ML Transform) and input datasets and training examples, and the tuning parameters needed by that algorithm.

AWS Glue uses those inputs to build an ML Transform that can be incorporated into a normal ETL Job workflow.

Q: How do ML Transforms work?

AWS Glue includes specialized ML-based dataset transformation algorithms customers can use to create their own ML Transforms. These include record de-duplication and match finding.

Customers start by navigating to the ML Transforms tab in the console (or using the ML Transforms service endpoints or accessing ML Transforms training via CLI) to create their first ML transform model. The ML Transforms tab provides a user-friendly view for management of user transforms. ML Transforms require distinct workflow requirements from other transforms, including the need for separate training, parameter tuning, and execution workflows; the need for estimating the quality metrics of generated transformations; and the need to manage and collect additional truth labels for training and active learning.

To create an ML transform via the console, customers first select the transform type (such as Record Deduplication or Record Matching) and provide the appropriate data sources previously discovered in Data Catalog. Depending on the transform, customers may then be asked to provide ground truth label data for training or additional parameters. Customers can monitor the status of their training jobs and view quality metrics for each transform. (Quality metrics are reported using a hold-out set of the customer-provided label data.)

Once satisfied with the performance, customers can promote ML Transforms models for use in production. ML Transforms can then be used during ETL workflows, both in code autogenerated by the service and in user-defined scripts submitted with other jobs, similar to pre-built transforms offered in other AWS Glue libraries.

Q: Can I see a presentation on using AWS Glue (and AWS Lake Formation) to find matches and deduplicate records?

A: Yes, the full recording of the AWS Online Tech Talk, "Fuzzy Matching and Deduplicating Data with ML Transforms for AWS Lake Formation" is available here.

AWS Glue DataBrew

Q: What is AWS Glue DataBrew?

AWS Glue DataBrew is a visual data preparation tool that makes it easy for data analysts and data scientists to prepare data with an interactive, point-and-click visual interface without writing code. With Glue DataBrew, you can easily visualize, clean, and normalize terabytes, and even petabytes of data directly from your data lake, data warehouses, and databases, including Amazon S3, Amazon Redshift, Amazon Aurora, and Amazon RDS. AWS Glue DataBrew is generally available today in US East (N. Virginia), US East (Ohio), US West (Oregon), EU (Ireland), EU (Frankfurt), Asia Pacific (Sydney), and Asia Pacific (Tokyo).

Q: Who can use AWS Glue DataBrew?

AWS Glue DataBrew is built for users who need to clean and normalize data for analytics and machine learning. Data analysts and data scientists are the primary users. For data analysts, examples of job functions are business intelligence analysts, operations analysts, market intelligence analysts, legal analysts, financial analysts, economists, quants, or accountants. For

data scientists, examples of job functions are materials scientists, bioanalytical scientists, and scientific researchers.

Q: What types of transformations are supported in AWS Glue DataBrew?

You can choose from over 250 built-in transformations to combine, pivot, and transpose the data without writing code. AWS Glue DataBrew also automatically recommends transformations such as filtering anomalies, correcting invalid, incorrectly classified, or duplicate data, normalizing data to standard date and time values, or generating aggregates for analyses. For complex transformations, such as converting words to a common base or root word, Glue DataBrew provides transformations that use advanced machine learning techniques such as Natural Language Processing (NLP). You can group multiple transformations together, save them as recipes, and apply the recipes directly to the new incoming data.

Q: What file formats does AWS Glue DataBrew support?

For input data, AWS Glue DataBrew supports commonly used file formats, such as comma-separated values (.csv), JSON and nested JSON, Apache Parquet and nested Apache Parquet, and Excel sheets. For output data, AWS Glue DataBrew supports comma-separated values (.csv), JSON, Apache Parquet, Apache Avro, Apache ORC and XML.

Q: Can I try AWS Glue DataBrew for free?

Yes. Sign up for an AWS Free Tier account, then visit the AWS Glue DataBrew Management Console, and get started instantly for free. If you are a first-time user of Glue DataBrew, the first 40 interactive sessions are free.

Q: Do I need to use AWS Glue Data Catalog or AWS Lake Formation to use AWS Glue DataBrew?

No. You can use AWS Glue DataBrew without using either the AWS Glue Data Catalog or AWS Lake Formation. However, if you use either the AWS Glue Data Catalog or AWS Lake Formation, DataBrew users can select the data sets available to them from their centralized data catalog.

Q: Can I retain a record of all changes made to my data?

Yes. You can visually track all the changes made to your data in the AWS Glue DataBrew Management Console. The visual view makes it easy to trace the changes and relationships made to the datasets, projects and recipes, and all other associated jobs. In addition, Glue DataBrew keeps all account activities as logs in the AWS CloudTrail.

AWS Glue Flex Jobs

Q: What is Glue Flex?

AWS Glue Flex is a flexible execution job class that allows you to reduce the cost of your non-urgent data integration workloads (e.g., pre-production jobs, testing, data loads, etc.) by up to 35%. Glue has two job execution classes: standard and flexible. The standard execution class is ideal for time-sensitive workloads that require fast job startup and dedicated resources. The flexible execution-class is appropriate for non-urgent jobs whose start and completion times

may vary. AWS Glue Flex can reduce the cost of your non-time-sensitive workloads (e.g. nightly batch ETL jobs, weekend jobs, one-time bulk data ingestion jobs, etc.).

Q: How do AWS Glue's standard and flexible execution classes differ?

AWS Glue's standard and flexible execution classes have different execution properties. With the standard execution class, jobs start immediately and have dedicated resources while running. Flexible execution class jobs run on non-dedicated compute resources in AWS that can be reclaimed while a job is running, and their start and completion times vary. As a result, the two execution-classes are appropriate for different workloads. The standard execution class is ideal for time-sensitive workloads that require fast job startup and dedicated resources. The flexible execution class is less expensive and suitable for non-urgent jobs where variance in start and completion times is acceptable.

Q: How do I get started with AWS Glue Flex flexible execution class jobs?

The flexible execution class is available for Glue Spark jobs. To use the flexible execution class, you simply change the default setting of the execution class parameter from "STANDARD to "FLEX". You can do this via Glue Studio or CLI. Visit the AWS Glue _user documentation_ for more information.

Q: What types of data integration and ETL workloads are not appropriate for AWS Glue Flex flexible execution class?

AWS Glue Flex flexible execution class is not appropriate for time-sensitive workloads that require consistent job start and run times, or for jobs that must complete execution by a specific time. AWS Glue Flex is also not recommended for long-running data integration workloads because they are more likely to get interrupted, resulting in frequent cancellations.

Q: How often should I expect jobs running with AWS Glue Flex flexible execution class to be interrupted?

The availability and interruption frequency of AWS Glue Flex depends on several factors, including the Region and Availability Zone (AZ), time of day, day of week. Resource availability determines whether Glue Flex jobs will start at all. While the interruption rate can be between 5-10% during peak hours, we expect the interruption rate of Glue Flex jobs to be about 5% or the failure rate of Glue Flex jobs due to interruption to be under 5%.
Q: Is the flexible execution class always available?

Yes, you can always choose the flexible execution class to run your Glue jobs. However, the ability of AWS Glue to execute these jobs is based on the availability of non-dedicated AWS capacity and the number of workers selected for your job. It is possible that, during peak times, Glue may not have adequate capacity for your job. In that case, your job will not start. You can specify a timeout value after which Glue will cancel the job. The longer the timeout value, the greater the chance that your job will be executed.

Q: What happens if an AWS Glue Flex job is interrupted during execution?

If a Glue Flex job is interrupted because there are no longer sufficient workers to complete the job based on the number of workers specified, the job will fail. Glue will retry failed jobs up to the specified maximum number of retries on the job definition before cancelling the job. You

should not use flexible execution class for any job that has a downstream dependency on other systems or processes.

Q: What types of AWS Glue jobs are supported by the flexible execution class?

The flexible execution class supports only Glue Spark jobs. Pythonshell and streaming are not supported.AWS Glue Flex is supported by Glue version 3.0 and later. The flexible execution class does not currently support streaming workloads.

AWS Product Integrations

Q: When should I use AWS Glue vs. AWS Data Pipeline?

AWS Glue provides a managed ETL service that runs on a serverless Apache Spark environment. This allows you to focus on your ETL job and not worry about configuring and managing the underlying compute resources. AWS Glue takes a data first approach and allows you to focus on the data properties and data manipulation to transform the data to a form where you can derive business insights. It provides an integrated data catalog that makes metadata available for ETL as well as querying via Amazon Athena and Amazon Redshift Spectrum.

AWS Data Pipeline provides a managed orchestration service that gives you greater flexibility in terms of the execution environment, access and control over the compute resources that run your code, as well as the code itself that does data processing. AWS Data Pipeline launches compute resources in your account allowing you direct access to the Amazon EC2 instances or Amazon EMR clusters.

Furthermore, AWS Glue ETL jobs are Scala or Python based. If your use case requires you to use an engine other than Apache Spark or if you want to run a heterogeneous set of jobs that run on a variety of engines like Hive, Pig, etc., then AWS Data Pipeline would be a better choice.
Q: When should I use AWS Glue vs. Amazon EMR?

AWS Glue works on top of the Apache Spark environment to provide a scale-out execution environment for your data transformation jobs. AWS Glue infers, evolves, and monitors your ETL jobs to greatly simplify the process of creating and maintaining jobs. Amazon EMR provides you with direct access to your Hadoop environment, affording you lower-level access and greater flexibility in using tools beyond Spark.

Q: When should I use AWS Glue vs AWS Database Migration Service?

AWS Database Migration Service (DMS) helps you migrate databases to AWS easily and securely. For use cases which require a database migration from on-premises to AWS or database replication between on-premises sources and sources on AWS, we recommend you use AWS DMS. Once your data is in AWS, you can use AWS Glue to move, combine, replicate, and transform data from your data source into another database or data warehouse, such as Amazon Redshift.

Q: When should I use AWS Glue vs AWS Batch?

AWS Batch enables you to easily and efficiently run any batch computing job on AWS regardless of the nature of the job. AWS Batch creates and manages the compute resources in your AWS account, giving you full control and visibility into the resources being used. AWS Glue is a fully-managed ETL service that provides a serverless Apache Spark environment to run your ETL jobs. For your ETL use cases, we recommend you explore using AWS Glue. For other batch-oriented use cases, including some ETL use cases, AWS Batch might be a better fit.

Security and availability

Q: How does AWS Glue keep my data secure?

We provide server-side encryption for data at rest and SSL for data in motion.

Q: What are the service limits associated with AWS Glue?

Please refer our documentation to learn more about service limits.

Q: What regions is AWS Glue in?

Please refer to the AWS Region Table for details of AWS Glue service availability by region.

Q: How many DPUs (Data Processing Units) are allocated to the development endpoint?

A development endpoint is provisioned with 5 DPUs by default. You can configure a development endpoint with a minimum of 2 DPUs and a maximum of 5 DPUs.

Q: How do I scale the size and performance of my AWS Glue ETL jobs?

You can simply specify the number of DPUs (Data Processing Units) you want to allocate to your ETL job. A Glue ETL job requires a minimum of 2 DPUs. By default, AWS Glue allocates 10 DPUs to each ETL job.

Q: How do I monitor the execution of my AWS Glue jobs?

AWS Glue provides the status of each job and pushes all notifications to Amazon CloudWatch. You can set up SNS notifications via CloudWatch actions to be informed of job failures or completions.

AWS Data Exchange

Easily find, subscribe to, and use third-party data in the cloud.

- Extensive dataset catalog
- Better data technology with AWS integration
- Streamlined data procurement and governance
- Easy to use for data files, tables, and APIs

AWS Data Exchange advantage

AWS Data Exchange is on a mission to increase speed to value for third-party datasets in the cloud.

There is no other place where customers can find data files, data tables, and data APIs from a vast portfolio of third-party datasets. We continuously innovate to make the world's third-party data easy to find in one data catalog, simple to subscribe to with consistent pricing options, and seamless to use with AWS data and analytics and machine learning services.

AWS Data Exchange dataset catalog categories

- Automotive
- ESG & sustainability
- Gaming
- Location
- Manufacturing
- Public sector
- Telecoms

Why AWS Data Exchange?

There is no other place to find data files, data tables, and data APIs from an extensive selection of data sets and data categories.

AWS Data Exchange is the world's most comprehensive service for third-party data sets

AWS Data Exchange is the only data marketplace with more than 3,500 products from over 300 providers delivered —through files, APIs, or Amazon Redshift queries— directly to the data lakes, applications, analytics, and machine learning models that use it.

With AWS Data Exchange, you can streamline all third-party data consumption, from existing subscriptions —which you can migrate at no additional cost to you or the provider— to future data subscriptions, in one place.

As an AWS service, AWS Data Exchange is secure and compliant, integrated with AWS and third-party tools and services, and offers consolidated billing and subscription management.

AWS Data Exchange benefits

Extensive Data Set Selection

- 3,500+ data sets from 300+ data providers
- Over 1,000 free data products and custom data products
- Automatic access to new data

Streamlined Data Procurement & Governance

- One place to exchange data publicly or privately
- Migrate existing subscriptions at no additional cost
- Simplified contracts and consolidated secure billing

Better Data Technology

- Native integration of data into AWS
- Data encrypted at rest and in transit
- Integration with AWS Identity & Access Management (IAM)

Ease of Use for Data Analytics & Machine Learning

- Cloud native support for data files, tables, and APIs
- Self-service options
- Secure and compliant solutions

How AWS Data Exchange Works

AWS Data Exchange supports Data Files, Data Tables, and Data APIs.

Data Files
Ingest data files with Amazon S3
Automatically export new or updated data sets to Amazon S3:

- Use data in a range of applications
- 'Set and forget' export preference to any Amazon S3 bucket
- Configure export jobs to begin automatically after subscription

Data Tables
Query data tables with AWS Data Exchange for Amazon Redshift

- Query data directly from vendor's databases
- No ETL – data is structure and ready for analysis
- Access the last data with automatic refreshes

Data APIs
Call APIs for data
Use AWS-native authentication and governance, AWS SDKs, and consistent API documentation

- Synchronous APIs for frequent retrieval of small amounts of data
- Simplified discovery, set up, and invocation
- Downloadable SDKs and consistent documentation

What is AWS Data Exchange?

AWS Data Exchange is a service that makes it easy for AWS customers to find, subscribe to, and use third-party data in the AWS Cloud.

As a subscriber, you can find and subscribe to thousands of products from qualified data providers. Then, you can use the AWS Data Exchange console or APIs to create, view, manage, and access data sets for use across a variety of AWS analytics and machine learning services. Anyone with an AWS account can be an AWS Data Exchange subscriber. For information about becoming a subscriber, see Subscribing to data products on AWS Data Exchange.

For providers, AWS Data Exchange eliminates the need to build and maintain any data delivery, entitlement, or billing technology. Providers in AWS Data Exchange have a secure, transparent, and reliable channel to reach AWS customers and grant existing customers their subscriptions more efficiently. The process for becoming an AWS Data Exchange provider requires a few steps to determine eligibility. For more information, see Providing data products on AWS Data Exchange.

In addition, anyone, with or without an AWS account, can find and use publicly available data sets that are part of the Open Data on AWS program. For more information, see Using Open Data on AWS data sets.

What is an AWS Data Exchange product?

A product is the unit of exchange in AWS Data Exchange that is published by a provider and made available for use to subscribers. When a provider publishes a product, that product is listed on the AWS Data Exchange product catalog as well as AWS Marketplace after being reviewed by AWS against our guidelines and terms and conditions. Each product you publish is uniquely identified by its product ID.

Note

When a product is initially created and published, all pre-existing finalized revisions within its data sets are published at the same time.

With AWS Data Exchange, providers publish data products and subscribers subscribe to those products.

Providers can publish and view their products using the AWS Data Exchange console. Providers can also list and view the details of their existing products using the AWS Marketplace Catalog API.

A product has the following parts:

- **Product details** – This information includes name, descriptions (both short and long), logo image, and support contact information. Providers complete the product details.
  - For more information as a subscriber, see Product subscriptions.
  - For more information as a provider, see Product details.
- **Product offers** – Offers define the terms that subscribers are agreeing to when they subscribe to a product. To make a product available on AWS Data Exchange, providers must define a public offer. This offer includes prices and durations, data subscription agreement, refund policy, and the option to create custom offers.
  - For more information as a subscriber, see Private products and offers and Bring Your Own Subscription (BYOS) offers
  - For more information as a provider, see Creating an offer for AWS Data Exchange products.
- **Data sets** – A product can contain one or more data sets. A data set in AWS Data Exchange is a dynamic set of data which is versioned through the use of revisions. Each revision can contain multiple assets. The provider can decide which revisions within a data set are published to a product. The provider creates owned data sets, and a subscriber can get access to entitled data sets through a product subscription. When a subscriber subscribes to a product, they get access to the product's data sets and some or all of the revisions that have been published to that product for the duration of their subscription.
  - For more information as a subscriber, see Data sets and revisions
  - For more information as a provider, see Data in AWS Data Exchange.

Malware prevention

Security and compliance is a shared responsibility between you and AWS. To promote a safe, secure, and trustworthy service for everyone, AWS Data Exchange scans all S3 object files published by providers before they are made available to subscribers. If AWS detects malware, the affected asset is removed.

Important

AWS Data Exchange does not guarantee that the data you consume as a subscriber is free of any potential malware. We encourage that you conduct your own additional due diligence to ensure compliance with your internal security controls. You can find anti-malware and security products in AWS Marketplace.

Supported data sets

AWS Data Exchange takes a responsible approach to facilitating data transactions by promoting transparency through use of the service. AWS Data Exchange reviews permitted data types, restricting products that are not permitted. Providers are limited to distributing data sets that meet the legal eligibility requirements set forth in the Terms and Conditions for AWS Marketplace Sellers.

For more information about permitted data types, see Publishing guidelines.

Important

As an AWS customer, you are encouraged to conduct your own additional due-diligence to ensure compliance with any applicable data privacy laws. If you suspect that a product or other resources on AWS Data Exchange are being used for abusive or illegal purposes, report it using the Report Amazon AWS abuse form.

Accessing AWS Data Exchange

Subscribers

As a subscriber, you can explore available AWS Data Exchange products through the following options:

- AWS Data Exchange console (**Browse catalog**)
- AWS Marketplace catalog

Providers

As an existing provider, you can access AWS Data Exchange through the following options:

- Directly through the AWS Data Exchange console (**Publish data**)
- Programmatically using the following APIs:
    - **AWS Data Exchange API** – Use the API operations to create, view, update, and delete data sets and revisions. You can also use these API operations to import and export assets to and from those

revisions. For more information, see the AWS Data Exchange API Reference.

- **AWS Marketplace Catalog API** – Use the API operations to view and update products on AWS Data Exchange and AWS Marketplace. For more information, see the AWS Marketplace Catalog API Reference.

Related services

The following services are related to AWS Data Exchange:

- **Amazon S3** – One supported asset type for data sets is Amazon S3 object snapshots. Subscribers can export data sets to Amazon S3 programmatically. For more information, see What Is Amazon S3? in the *Amazon Simple Storage Service User Guide*.
- **Amazon API Gateway** – Another supported asset type for data sets is APIs. Subscribers can call the API programmatically, call the API from the AWS Data Exchange console, or download the OpenAPI specification file. For more information, see What is Amazon API Gateway? in the *Amazon API Gateway* Developer Guide.
- **Amazon Redshift** – AWS Data Exchange supports Amazon Redshift data sets. Subscribers can get read-only access to query the data in Amazon Redshift without extracting, transforming, and loading data. For more information, see Getting started with Amazon Redshift in the *Amazon Redshift Getting Started Guide* and Amazon Redshift system overview in the *Amazon Redshift Database Developer Guide*.
- **AWS Marketplace** – AWS Data Exchange allows data sets to be published as products on AWS Marketplace. AWS Data Exchange providers must be registered as AWS Marketplace sellers, and can use the AWS Marketplace Management Portal or the AWS Marketplace Catalog API. For information about becoming an AWS Marketplace subscriber, see What Is AWS Marketplace? in the *AWS Marketplace Buyer Guide*. For information about becoming an AWS Marketplace seller, see What Is AWS Marketplace? in the *AWS Marketplace Seller Guide*.

AWS Data Exchange FAQs

General

Q: Who are the primary users of AWS Data Exchange?

AWS Data Exchange makes it easy for AWS customers to securely exchange and use third-party data on AWS. Data analysts, product managers, portfolio managers, data scientists, quants, clinical trial technicians, and developers in nearly every industry would like access to more data to drive analytics, train machine learning (ML) models, and make data-driven decisions. But there is no one place to find data from multiple providers and no consistency in how providers deliver data, leaving them to deal with a mix of shipped physical media, FTP credentials, and bespoke API calls. Conversely, many organizations would like to make their data available for either research or commercial purposes, but it's too hard and expensive to build and maintain data delivery, entitlement, and billing technology, which further depresses the supply of valuable data.

Q: What AWS Regions is AWS Data Exchange available in?

AWS Data Exchange has a single global product catalog offered by providers available from any supported AWS Region. You can see the same catalog regardless of which Region you are using. The resources underlying the product (data sets, revisions, assets) are regional resources that you manage programmatically or through the AWS Data Exchange console in specific AWS Regions. For a list of AWS Regions in which AWS Data Exchange is available today, see Service endpoints.

Q: What is the difference between AWS Data Exchange and the Registry of Open Data on AWS?

All publicly available data sets on the Registry of Open Data on AWS are also available through AWS Data Exchange. On AWS Data Exchange, customers can now discover and access more than 100 petabytes of high-value, cloud-optimized data sets available for public use from leading organizations such as NOAA, NASA, or the UK Met Office. These include open data sets hosted by the AWS Open Data Sponsorship Program and in the Amazon Sustainability Data Initiative (ASDI) catalog. Open data sets are different than other commercial or free data sets in four key ways. First, AWS Data Exchange requires customers to explicitly agree to the Data Subscription Agreement outlining the terms that the data provider set when publishing their product, whereas open data does not have terms of use and is only governed by the provider specific open data license. Second, customers must authenticate using an AWS account to subscribe to commercial or free data sets whereas open data sets can be accessed without any authentication via S3 APIs. Third, AWS Data Exchange gives data providers access to daily, weekly and monthly reports detailing subscription activity, whereas with Registry of Open Data on AWS, data providers must analyze their own logs to track usage of data. Finally, to become a data provider on AWS Data Exchange, qualified customers must register as a data provider on the AWS Marketplace Management Portal to be eligible to list both free and commercial products, whereas any customer can add free data to Registry of Open Data on AWS through GitHub and may apply to the Open Data Sponsorship Program for AWS to sponsor the costs of storage and bandwidth for select open data sets.

Q: What is AWS Data Exchange for APIs?

AWS Data Exchange for APIs is a feature that enables customers to find, subscribe to, and use third-party API products from providers on AWS Data Exchange. With AWS Data Exchange for APIs, customers can use AWS-native authentication and governance, explore consistent API documentation, and utilize supported AWS SDKs to make API calls. Now, by adding their APIs to the AWS Data Exchange catalog, data providers can reach millions of AWS customers that consume API-based data and more easily manage subscriber authentication, entitlement, and billing.

Subscriber

Q: What type of data can I subscribe to on AWS Data Exchange?

Today, AWS Data Exchange contains more than 3,500 data products from a broad range of domains, including financial services (for example, top US businesses by revenue); healthcare and life sciences (for example, population health management); geospatial (for example, satellite imagery); weather (for example, historical and future trajectories of temperature); and mapping (for example, street level imagery and foot traffic patterns). For a complete list of data providers, navigate to the AWS Data Exchange catalog. If customers need additional data sources not currently available on AWS Data Exchange, they can log these requests here.

Q: How can I see the catalog of AWS Data Exchange products?

Anyone can browse the AWS Data Exchange catalog in AWS Marketplace under the Data category, or by searching AWS Marketplace for keywords of interest. Authenticated AWS customers can also browse the AWS Data Exchange catalog alongside existing subscriptions on the AWS Data Exchange console. For more information about subscribing to data products, see Getting started as a subscriber.

Q: How will I know when new revisions to data sets I'm subscribed to become available?

As a subscriber with an active subscription to a product, you'll receive an Amazon CloudWatch Events event from AWS Data Exchange every time the provider publishes new revisions. You can use this CloudWatch event to automate consumption of new data. For more information about CloudWatch Events, see Logging and monitoring on AWS Data Exchange.

Q: Can I migrate preexisting data subscriptions to be delivered by AWS Data Exchange?

Yes. AWS Data Exchange allows qualified data providers to fulfill to existing subscribers using a Bring Your Own Subscription (BYOS) entitlement at no additional cost. With a BYOS offer, the existing billing relationship will continue between you and the data provider. Talk to your data provider about using this capability.

Q: What are the subscription durations that are available for AWS Data Exchange products?

Data providers list products subscriptions with 1- to 36-month duration terms. You can find subscription duration options on each product's detail page.

Q: Can I set my subscription to auto-renew?

Providers have the option to enable auto-renewal on individual offers. As a subscriber, you can choose to set your subscription to auto-renew on the offers that have the auto-renewal functionality enabled. Auto-renewal is available for public and private offers that do not have payment schedules.

Q: Can data providers change the terms of the offer that I am subscribed to? How would it affect my subscription and renewal?

Yes. Data providers can update the terms of the offer at any time, but this will not affect existing subscriptions. For subscriptions set to auto-renew, AWS Data Exchange will automatically renew the subscription at the latest terms that the provider specified on or by renewal date, which may be different from the original subscription terms. For more information, see Product subscriptions.

Q: How are refunds handled?

AWS Data Exchange requires data providers to specify their refund policy, which you can see on the subscription details page. For any refund requests, you'll need to contact the provider directly. After a provider approves a refund request, AWS will process and issue the approved refund.

Q: How do I know that data I subscribe to is free of any malware?

Security and Compliance is a shared responsibility (https://aws.amazon.com/compliance/shared-responsibility-model/) between AWS and the customer. To promote a safe, secure, and trustworthy service for everyone, AWS Data Exchange scans data files hosted in S3 buckets it manages before making them available to subscribers. If AWS detects malware, AWS will remove the affected file(s). AWS Data Exchange does not guarantee that the data you consume as a subscriber is free of any potential malware. Customers are encouraged to conduct their own additional due diligence to ensure compliance with their internal security controls. You can explore many third-party anti-malware and security products in AWS Marketplace.

Q: How do I remain compliant with applicable data privacy laws when subscribing to AWS Data Exchange products?

Security and Compliance is a shared responsibility among AWS, the data provider, and the subscriber. Detailed restrictions around eligible data sets and other related legal compliance matters are set forth in Terms and Conditions for AWS Marketplace Providers, which every data provider must agree to before listing any data products. If AWS learns that these terms are breached in any way, AWS will remove such content from AWS Data Exchange and the data provider may be suspended from the service.

Providers and subscribers are responsible for conducting their own additional due-diligence to ensure compliance with any data privacy laws. If you suspect that a data product or AWS Data Exchange resources are being used for abusive or illegal purposes, you can complete and submit the form found on Report Amazon AWS abuse.

Q: Are there are any restrictions for how AWS Data Exchange and any data obtained through AWS Data Exchange can be used?

Yes, AWS explicitly prohibits the use of AWS Data Exchange for any illegal or fraudulent activities. Data may not be used for any activities that result in the violation of an individual's rights or unlawfully discriminate against others based on race, ethnicity, sexual orientation, gender identity, or other related groups. Subscribers may not use any content obtained through AWS Data Exchange that was anonymized or aggregated (such that it is no longer associated with an identifiable individual) by the data provider to create, generate, or infer any information relating to a person's identity (for example, attempting to triangulate with other data sources).

Q: How do I report abusive content or request information be removed from a product suspected of abuse?

If you suspect that a data product or AWS Data Exchange resources are being used for abusive or illegal purposes, you can complete and submit the form found on Report Amazon AWS abuse. If AWS learns that our terms are breached in any way, AWS may remove the subscriber's access to the data product and the data provider or the subscriber may be suspended or terminated from future use of AWS Data Exchange

Q: I currently consume third-party data directly into my S3 bucket(s). Why should I consider using AWS Data Exchange to consume data from third parties instead?

AWS Data Exchange centralizes, simplifies, and accelerates your data acquisition process.

AWS Data Exchange allows you to consolidate your ingestion across data providers and receive your data using a single API. You can easily subscribe to new data products, and migrate any existing data feeds through the Bring Your Own Subscription feature.

AWS Data Exchange sends you a CloudWatch Events event as new data is published, allowing you to automate ingestion without having to manage any physical media, FTP credentials, or other legacy technology. It also stores copies of data you've licensed, allowing you to access point-in-time history all in one secure location.

AWS Data Exchange also centralizes your billing so you can receive one invoice from AWS instead of one for each data provider. With the AWS Data Exchange console, you can easily track your subscriptions in one place, and better manage your data pipeline.

Q: How do I pay for my data product subscriptions?

When you purchase a data product on AWS Data Exchange with upfront payments, you'll receive an invoice from AWS immediately. When you purchase a data product on AWS Data Exchange with multiple payments, you'll receive an invoice based on the dates specified in the payment schedule that will appear as part of your AWS Marketplace charges. You can see a breakout of charges for each data product by name in the Detail section of the invoice. You'll receive separate bills for usage of AWS infrastructure and analytics services such as Amazon Simple Storage Service (Amazon S3) or Amazon Athena. For more information related to AWS Billing and Cost Management, see Paying for products.

Q: How do I find API products in the AWS Data Exchange catalog?

You can find products containing APIs in the AWS Marketplace catalog, or you can navigate to the AWS Data Exchange catalog and select *API* under the *Data available through* filter.

Q: How do I subscribe to an API product?

After you've found the API that you want to subscribe to, select the product to learn more on the product detail page. Next, choose *Continue to subscribe*, review the subscription terms, and then choose the *Subscribe* button at the bottom of the page.

Note: You may be asked to submit information to the data provider before you can request to subscribe to their product. For more information on subscribing and using an API product, see the Subscribing to a product containing APIs topic in the *AWS Data Exchange User Guide*.

Q: How do I make an API call?

First, ensure you have successfully subscribed to a product containing an API data set. Then, navigate to the product's asset detail page to view API schemas and code snippets that will help you structure your API call. You can also utilize the AWS SDK to automatically sign your API requests with your AWS credentials. For more information

about how to structure API calls to AWS Data Exchange products containing API data sets, see the Making an API call (console) in the *AWS Data Exchange User Guide.*

Q: Does AWS Data Exchange for APIs have a Service Level Agreement (SLA)?

AWS Data Exchange for APIs does not currently offer an SLA.

Q: Are there any AWS Data Exchange for APIs-specific SDKs that I should be aware of?

We have added AWS Data Exchange for APIs-specific operations to the following SDKs:

- AWS SDK for Go
- AWS SDK for JavaScript
- AWS SDK for Node.js
- AWS SDK for PHP
- AWS SDK for Python
- AWS SDK for Ruby

For more information, see the AWS Data Exchange API Reference.

Provider

Q: How do I qualify to become a data provider on AWS Data Exchange?

To become a data provider on AWS Data Exchange, data providers must agree to the Terms and Conditions for AWS Marketplace Providers ("AWS Marketplace Terms & Conditions"). Data providers must use a valid legal entity domiciled in the United States or a member state of the EU, supply valid banking and taxation identification, and be qualified by the AWS Data Exchange business operations team. Each data provider will also undergo a detailed review by the AWS Data Exchange team prior to being granted permission to list data products on the catalog.

Q: How is data organized in AWS Data Exchange?

Data in AWS Data Exchange is organized using three building blocks: data sets, revisions, and assets. A data set is a container for data that belongs together (for example, end of data pricing for equities trading in the US). Data sets contain a series of revisions, which data providers publish as needed to make new assets available. Revisions can represent changes or new data (for example, today's end of day prices), corrections to previous revisions, or entirely new snapshots. Assets are any file that can be stored in Amazon Simple Storage Service (S3) (for example, CSV, Parquet, or image files). For more information about working with data, see Data in AWS Data Exchange. Using these building blocks, you can organize the data hierarchically to build complex data models or as single data files.

Q: After I create data sets, how do I publish and make them available to my subscribers?

Data sets are made available to subscribers as part of a product. A product is a collection of one or more data sets, metadata that makes the product discoverable in the AWS Data Exchange catalog, pricing, and a Data Subscription Agreement with terms for your customers. For more information, see Publishing a new product.

Q: Can I choose which customers can subscribe to my data?

Yes. You have an option to enable subscription verification on any public product, which will require prospective subscribers to fill out a subscription request form including their identity and intended use-case details before subscribing. For these products, you'll have up to 45 days to either approve or decline the subscription request. Note that public products that include personal data are required to have subscription verification enabled. For more information, see Subscription verification for providers.

You can also offer private products to specific subscribers by using their AWS Account ID. Private products can be custom-made from existing products on the public catalog or completely new products only offered to the specific customers of your choosing and not listed on the public catalog.

Q: Do I have to package files in a specific format?

AWS Data Exchange allows you to package files in any file format, but you'll want to consider what will allow subscribers to most conveniently gain insight from the data. Parquet formatted files, for example, will allow subscribers to instantly run queries using Amazon Athena in a cost-effective way. Binary or other proprietary file formats will require that subscribers understand how to parse the information, which AWS recommends explaining in each product description.

Q: Who owns the data I am distributing as a provider through AWS Data Exchange?

You retain ownership of the data you distribute as a data provider on AWS Data Exchange. The Terms and Conditions for AWS Marketplace Providers require each data provider to attest that they have the legal right to distribute the data they publish. Subscribers must legally agree to the Data Subscription Agreement specified by the data provider before gaining access to data sets contained in a product, which remains available for both data providers and subscribers. Consistent with the AWS acceptable use policy, AWS Data Exchange may suggest corrective action where there is evidence of abuse, but it is the data provider's responsibility to enforce and govern the terms of use.

Q: How do I specify the Data Subscription Agreement (DSA)?

AWS Data Exchange provides an optional Data Subscription Agreement (DSA) template that incorporates inputs from multiple AWS customers and data providers. You can choose to use this DSA template, copy and edit it with their own terms and conditions, or specify custom terms by uploading a DSA of your choice. AWS Data Exchange will associate the DSA specified for the product without any further modifications. For more information, see Publishing a new product.

Q: What ways can I price my data sets?

AWS Data Exchange currently supports subscription-based pricing from 1- to 36-month duration terms.

Q: Is AWS Data Exchange suitable for data providers who want to distribute their data for free?

Yes. Many data providers make their data products available for free for research, scientific, or other noncommercial use cases.

Q: Can I customize pricing or terms for select customers?

Yes. Private offers allow you to make public products available to select AWS customers and to set terms including price, duration, payment schedule, DSA, or refund policy. For more information, see Create private offers.

Q: Are there any restrictions on what data can be made available on AWS Data Exchange?

Yes. Publishing guidelines for listing products on AWS Data Exchange and Terms and Conditions for AWS Marketplace Providers restrict certain categories of data. Unless a provider is enrolled in the Extended Provider Program, data products listed on AWS Data Exchange may not include information that can be used to identify any person, except information that is already legally available to the public, such as newspaper articles, open court records, public company filings, or public online profiles. For more information about the Extended Provider Program, please contact AWS Support.

Q: Can I remove a product that I published from the catalog?

Yes. You can unpublish a product at any time to ensure that no new subscribers are able to view and subscribe to your product, including auto-renewal cancellation for existing subscribers. You'll need to keep data current for any existing subscribers until each subscription expires.

Q: What happens if I need to remove data from AWS Data Exchange?

You can remove or change the price or Data Subscription Agreement (DSA) of a product at any time, although existing subscriptions will remain in effect until their next renewal. If a data provider erroneously publishes data, you can open a support case open a support case to have the data unpublished.

Q: How do I know who is subscribing to the data I have listed on AWS Data Exchange?

AWS Data Exchange provides daily, weekly, and monthly reports detailing subscription activity.

Q: When and how often will I receive payments?

You will receive a disbursement for subscriptions less fulfillment fees once a month. AWS will disburse all funds that AWS has received from subscribers by that date to the bank account linked to the AWS account that you used at registration.

Q: How will AWS handle collection and remittances of US sales and use tax?

When listing your data sets, you can enable collection and remittance of US sales and use tax. You can also configure your tax nexus to account for places where you have a physical presence to direct AWS to collect appropriate taxes. It's helpful to review the AWS Marketplace U.S Tax Collection Support Terms and Conditions. For details on sales tax collection in other geographies, see AWS Marketplace Sellers & Tax Collection.

Q: Is Amazon.com or AWS providing customers' data on AWS Data Exchange?

No. Neither Amazon.com nor AWS are providing customers' data on AWS Data Exchange.

Q: Can AWS access data products listed on AWS Data Exchange?

AWS is vigilant about our customers' privacy. Companies listing data products on the AWS Data Exchange own them and are able to maintain control over who accesses their content. AWS does not access or use data products except as necessary to provide the AWS Data Exchange service.

Q: I currently publish data directly into my subscribers' S3 buckets. Why should I consider using AWS Data Exchange to publish data to third parties instead?

With AWS Data Exchange you can publish data simultaneously to all your customers and spend more time growing your business rather than managing logistics.

AWS Data Exchange allows you to publish data to your customers through an easy-to-use API and console. The data model for AWS Data Exchange simplifies managing and publishing data through three reusable constructs: data sets, revisions, and assets. AWS customers can subscribe to these published data products directly from AWS Marketplace. The service automatically grants entitlements to customers who have an active subscription to the product, which means that you don't have to manually configure and maintain custom permissions to an Amazon Simple Storage Service (S3) bucket for each subscription.

AWS Data Exchange automatically sends an Amazon CloudWatch Events event to all subscribers when new revisions are published, which allows subscribers to automate their consumption of new data. When a subscription is no-longer active, AWS Data Exchange revokes that subscriber's entitlement to your data.

If your product is already available on AWS Data Exchange, you can simply use Bring Your Own Subscription (BYOS) to configure an entitlement to your existing subscribers for no additional cost and without any additional programming work. For more information about using BYOS to migrate and fulfill existing subscriptions with AWS customers, see Create Bring Your Own Subscription offers.

Q: How do I publish a product containing APIs?

As a provider, you first need to set up an AWS account and register as an AWS Marketplace seller. You can then publish an API product by following the steps detailed in the Publishing a product containing APIs topic in the *AWS Data Exchange User Guide*.

Q: Do providers have to offer a Service Level Agreement (SLA) to AWS Data Exchange when offering an API product?

AWS Data Exchange for APIs does not require providers to offer an uptime or availability SLA. Providers and subscribers can negotiate custom terms as part of a DSA. See Publishing Products for further information.

Q: What guidelines do I need to follow as a provider on AWS Data Exchange for APIs?

In addition to following guidelines under the Terms and Conditions for AWS Marketplace Sellers and the AWS Customer Agreement, providers of products containing APIs must respond to subscriber support inquiries within 1 business day, as set forth in the *AWS Data Exchange User Guide*. Not following the guidelines may result in products being removed from AWS Data Exchange. For more information, see Publishing guidelines topic in the *AWS Data Exchange User Guide*.

Q: Can products contain APIs, Amazon S3 objects, and Amazon Redshift data shares?

Yes. As a data provider, you can publish products containing multiple data set types.

Amazon SageMaker

Build, train, and deploy machine learning (ML) models for any use case with fully managed infrastructure, tools, and workflows.

What Is Amazon SageMaker?

Amazon SageMaker is a fully managed machine learning service. With SageMaker, data scientists and developers can quickly and easily build and train machine learning models, and then directly deploy them into a production-ready hosted environment. It provides an integrated Jupyter authoring notebook instance for easy access to your data sources for exploration and analysis, so you don't have to manage servers. It also provides common machine learning algorithms that are optimized to run efficiently against extremely large data in a distributed environment. With native support for bring-your-own-algorithms and frameworks, SageMaker offers flexible distributed training options that adjust to your specific workflows. Deploy a model into a secure and scalable environment by launching it with a few clicks from SageMaker Studio or the SageMaker console. Training and hosting are billed by minutes of usage, with no minimum fees and no upfront commitments.

Amazon SageMaker Features

Amazon SageMaker includes the following features:

SageMaker Studio
An integrated machine learning environment where you can build, train, deploy, and analyze your models all in the same application.

SageMaker Canvas
An auto ML service that gives people with no coding experience the ability to build models and make predictions with them.

SageMaker Ground Truth Plus
A turnkey data labeling feature to create high-quality training datasets without having to build labeling applications and manage the labeling workforce on your own.

SageMaker Studio Lab
A free service that gives customers access to AWS compute resources in an environment based on open-source JupyterLab.

SageMaker Training Compiler
Train deep learning models faster on scalable GPU instances managed by SageMaker.

SageMaker Studio Universal Notebook
Easily discover, connect to, create, terminate and manage Amazon EMR clusters in single account and cross account configurations directly from SageMaker Studio.

SageMaker Serverless Endpoints
A serverless endpoint option for hosting your ML model. Automatically scales in capacity to serve your endpoint traffic. Removes the need to select instance types or manage scaling policies on an endpoint.

SageMaker Inference Recommender
Get recommendations on inference instance types and configurations (e.g. instance count, container parameters and model optimizations) to use your ML models and workloads.

SageMaker Model Registry
Versioning, artifact and lineage tracking, approval workflow, and cross account support for deployment of your machine learning models.

SageMaker Projects
Create end-to-end ML solutions with CI/CD by using SageMaker projects.

SageMaker Model Building Pipelines
Create and manage machine learning pipelines integrated directly with SageMaker jobs.

SageMaker ML Lineage Tracking
Track the lineage of machine learning workflows.

SageMaker Data Wrangler
Import, analyze, prepare, and featurize data in SageMaker Studio. You can integrate Data Wrangler into your machine learning workflows to simplify and streamline data pre-processing and feature engineering using little to no coding. You can also add your own Python scripts and transformations to customize your data prep workflow.

SageMaker Feature Store
A centralized store for features and associated metadata so features can be easily discovered and reused. You can create two types of stores, an Online or Offline store. The

Online Store can be used for low latency, real-time inference use cases and the Offline Store can be used for training and batch inference.

SageMaker JumpStart
Learn about SageMaker features and capabilities through curated 1-click solutions, example notebooks, and pretrained models that you can deploy. You can also fine-tune the models and deploy them.

SageMaker Clarify
Improve your machine learning models by detecting potential bias and help explain the predictions that models make.

SageMaker Edge Manager
Optimize custom models for edge devices, create and manage fleets and run models with an efficient runtime.

SageMaker Ground Truth
High-quality training datasets by using workers along with machine learning to create labeled datasets.

Amazon Augmented AI
Build the workflows required for human review of ML predictions. Amazon A2I brings human review to all developers, removing the undifferentiated heavy lifting associated with building human review systems or managing large numbers of human reviewers.

SageMaker Studio Notebooks
The next generation of SageMaker notebooks that include AWS IAM Identity Center (successor to AWS Single Sign-On) (IAM Identity Center) integration, fast start-up times, and single-click sharing.

SageMaker Experiments
Experiment management and tracking. You can use the tracked data to reconstruct an experiment, incrementally build on experiments conducted by peers, and trace model lineage for compliance and audit verifications.

SageMaker Debugger
Inspect training parameters and data throughout the training process. Automatically detect and alert users to commonly occurring errors such as parameter values getting too large or small.

SageMaker Autopilot
Users without machine learning knowledge can quickly build classification and regression models.

SageMaker Model Monitor
Monitor and analyze models in production (endpoints) to detect data drift and deviations in model quality.

SageMaker Neo
Train machine learning models once, then run anywhere in the cloud and at the edge.

SageMaker Elastic Inference

Speed up the throughput and decrease the latency of getting real-time inferences.

Reinforcement Learning
Maximize the long-term reward that an agent receives as a result of its actions.

Preprocessing
Analyze and preprocess data, tackle feature engineering, and evaluate models.

Batch Transform
Preprocess datasets, run inference when you don't need a persistent endpoint, and associate input records with inferences to assist the interpretation of results.

Amazon SageMaker FAQs

General

Q: What is Amazon SageMaker?

Amazon SageMaker is a fully managed service to build, train, and deploy machine learning (ML) models for any use case with fully managed infrastructure, tools, and workflows.

Q: In which Regions is Amazon SageMaker available?

For a list of the supported Amazon SageMaker AWS Regions, please visit the AWS Region table for all AWS global infrastructure. Also, for more information, see Regions and endpoints in the AWS general reference.

Q: What is the service availability of Amazon SageMaker?

Amazon SageMaker is designed for high availability. There are no maintenance windows or scheduled downtimes. SageMaker APIs run in Amazon's proven, high-availability data centers, with service stack replication configured across three facilities in each AWS Region to provide fault tolerance in the event of a server failure or Availability Zone outage.

Q: How does Amazon SageMaker secure my code?

Amazon SageMaker stores code in ML storage volumes, secured by security groups and optionally encrypted at rest.

Q: What security measures does Amazon SageMaker have?

Amazon SageMaker ensures that ML model artifacts and other system artifacts are encrypted in transit and at rest. Requests to the SageMaker API and console are made over a secure (SSL) connection. You pass AWS Identity and Access Management roles to SageMaker to provide permissions to access resources on your behalf for training and deployment. You can use encrypted Amazon S3 buckets for model artifacts and data, as well as pass an AWS Key Management Service (KMS) key to SageMaker notebooks, training jobs, and endpoints, to encrypt the attached ML storage volume. Amazon SageMaker also supports Amazon Virtual Private Cloud (VPC) and AWS PrivateLink support.

Q: Does Amazon SageMaker use or share models, training data, or algorithms?

Amazon SageMaker does not use or share customer models, training data, or algorithms. We know that customers care deeply about privacy and data security. That's why AWS gives you ownership and control over your content through simple, powerful tools that allow you to determine where your content will be stored, secure your content in transit and at rest, and manage your access to AWS services and resources for your users. We also implement responsible and sophisticated technical and physical controls that are designed to prevent unauthorized access to or disclosure of your content. As a customer, you maintain ownership of your content, and you select which AWS services can process, store, and host your content. We do not access your content for any purpose without your consent.

Q: How am I charged for Amazon SageMaker?

You pay for ML compute, storage, and data processing resources you use for hosting the notebook, training the model, performing predictions, and logging the outputs. Amazon SageMaker allows you to select the number and type of instance used for the hosted notebook, training, and model hosting. You pay only for what you use, as you use it; there are no minimum fees and no upfront commitments.

Q: How can I optimize my Amazon SageMaker costs, such as detecting and stopping idle resources in order to avoid unnecessary charges?

There are several best practices you can adopt to optimize your Amazon SageMaker resource utilization. Some approaches involve configuration optimizations; others involve programmatic solutions. A full guide on this concept, complete with visual tutorials and code samples, can be found in this blog post.

Q: What if I have my own notebook, training, or hosting environment?

Amazon SageMaker provides a full end-to-end workflow, but you can continue to use your existing tools with SageMaker. You can easily transfer the results of each stage in and out of SageMaker as your business requirements dictate.

Q: Is R supported with Amazon SageMaker?

Yes, R is supported with Amazon SageMaker. You can use R within SageMaker notebook instances, which include a preinstalled R kernel and the reticulate library. Reticulate offers an R interface for the Amazon SageMaker Python SDK, enabling ML practitioners to build, train, tune, and deploy R models.

Q: How can I check for imbalances in my model?

Amazon SageMaker Clarify helps improve model transparency by detecting statistical bias across the entire ML workflow. SageMaker Clarify checks for imbalances during data preparation, after training, and ongoing over time, and also includes tools to help explain ML models and their predictions. Findings can be shared through explainability reports.

Q: What kind of bias does Amazon SageMaker Clarify detect?

Measuring bias in ML models is a first step to mitigating bias. Bias may be measured before training and after training, as well as for inference for a deployed model. Each measure of bias corresponds to a different notion of fairness. Even considering simple notions of fairness leads to many different measures applicable in various contexts. You need to choose bias notions and metrics that are valid for the application and the situation under investigation. SageMaker currently supports the computation of different bias metrics for training data (as part of SageMaker data preparation), for the trained model (as part of SageMaker Experiments), and for inference for a deployed model (as part of SageMaker Model Monitor). For example, before training, we provide metrics for checking whether the training data is representative (that is, whether one group is underrepresented) and whether there are differences in the label distribution across groups. After training or during deployment, metrics can be helpful to measure whether (and by how much) the performance of the model differs across groups. For example, start by comparing the error rates (how likely a model's prediction is to differ from the true label) or break further down into precision (how likely a positive prediction is to be correct) and recall (how likely the model will correctly label a positive example).

Q: How does Amazon SageMaker Clarify improve model explainablity?

Amazon SageMaker Clarify is integrated with SageMaker Experiments to provide a feature importance graph detailing the importance of each input for your model's overall decision-making process after the model has been trained. These details can help determine if a particular model input has more influence than it should on overall model behavior. SageMaker Clarify also makes explanations for individual predictions available via an API.

Q: What is Amazon SageMaker Studio?

Amazon SageMaker Studio provides a single, web-based visual interface where you can perform all ML development steps. SageMaker Studio gives you complete access, control, and visibility into each step required to build, train, and deploy models. You can quickly upload data, create new notebooks, train and tune models, move back and forth between steps to adjust experiments, compare results, and deploy models to production all in one place, making you much more productive. All ML development activities including notebooks, experiment management, automatic model creation, debugging and profiling, and model drift detection can be performed within the unified SageMaker Studio visual interface.

Q: What is RStudio on SageMaker?

Amazon SageMaker Studio provides a web-based visual interface to easily build your models in the Python language. You can also analyze and visualize data for ML in the R language with elastic compute resources in RStudio on SageMaker, the first fully managed RStudio Workbench in the cloud.

Q: How does Amazon SageMaker Studio pricing work?

There is no additional charge for using Amazon SageMaker Studio. You pay only for the underlying compute and storage charges on the services you use within Amazon SageMaker Studio.
Q: In which Regions is Amazon SageMaker Studio supported?

You can find the Regions where Amazon SageMaker Studio is supported in the documentation here.

Low-code ML

Q: What is Amazon SageMaker Autopilot?

Amazon SageMaker Autopilot is the industry's first automated machine learning capability that gives you complete control and visibility into your ML models. SageMaker Autopilot automatically inspects raw data, applies feature processors, picks the best set of algorithms, trains and tunes multiple models, tracks their performance, and then ranks the models based on performance, all with just a few clicks. The result is the best-performing model that you can deploy at a fraction of the time normally required to train the model. You get full visibility into how the model was created and what's in it, and SageMaker Autopilot integrates with Amazon SageMaker Studio. You can explore up to 50 different models generated by SageMaker Autopilot inside SageMaker Studio so it's easy to pick the best model for your use case. SageMaker Autopilot can be used by people without ML experience to easily produce a model, or it can be used by experienced developers to quickly develop a baseline model on which teams can further iterate.
Q: What built-in algorithms are supported in Amazon SageMaker Autopilot?

Amazon SageMaker Autopilot supports 2 built-in algorithms: XGBoost and Linear Learner.
Q: Can I stop an Amazon SageMaker Autopilot job manually?

Yes. You can stop a job at any time. When an Amazon SageMaker Autopilot job is stopped, all ongoing trials will be stopped and no new trial will be started.
Q: How do I get started with Amazon SageMaker quickly?

Amazon SageMaker JumpStart helps you quickly and easily get started with ML. SageMaker JumpStart provides a set of solutions for the most common use cases that can be deployed readily with just a few clicks. The solutions are fully customizable and showcase the use of AWS CloudFormation templates and reference architectures so you can accelerate your ML journey. SageMaker JumpStart also supports one-click deployment and fine-tuning of more than 150 popular open-source models such as transformer, object detection, and image classification models.
Q: Which open-source models are supported with Amazon SageMaker JumpStart?

Amazon SageMaker JumpStart includes 150+ pre-trained open-source models from PyTorch Hub and TensorFlow Hub. For vision tasks such as image classification and object detection, you can use models such as ResNet, MobileNet, and Single-Shot Detector (SSD). For text tasks such as sentence classification, text classification, and question answering, you can use models such as BERT, RoBERTa, and DistilBERT.
Q: What solutions come pre-built with Amazon SageMaker JumpStart?

SageMaker JumpStart includes solutions that are preconfigured with all necessary AWS services to launch a solution into production. Solutions are fully customizable so you can easily modify them to fit your specific use case and dataset. You can use solutions for over 15 use cases including demand forecasting, fraud detection, and predictive maintenance, and readily deploy solutions with just a few clicks. For more information about all solutions available, visit the SageMaker getting started page.

Q: How does Amazon SageMaker JumpStart pricing work?

You are charged for the AWS services launched from SageMaker JumpStart, such as training jobs and endpoints, based on SageMaker pricing. There is no additional charge for using Amazon SageMaker JumpStart.

Q: What is Amazon SageMaker Canvas?

Amazon SageMaker Canvas is a no-code service with an intuitive, visual, point-and-click interface that enables you to easily create highly accurate ML-based predictions from your data. SageMaker Canvas enables you to access and combine data from a variety of sources using a drag-and-drop user interface, automatically cleans and prepares data to minimize manual cleanup, applies a variety of state-of-the-art ML algorithms to find highly accurate predictive models, and provides an intuitive interface to make predictions. You can use SageMaker Canvas to make much more precise predictions in a variety of business applications and easily collaborate with data scientists and analysts in your enterprise by sharing your models, data, and reports. To learn more about SageMaker Canvas, please visit the Canvas FAQ page.

Machine learning workflows

Q: How can I build a continuous integration and delivery (CI/CD) pipeline with Amazon SageMaker?

Amazon SageMaker Pipelines helps you create fully automated ML workflows from data preparation through model deployment so you can scale to thousands of ML models in production. SageMaker Pipelines comes with a Python SDK that connects to SageMaker Studio so you can take advantage of a visual interface to build each step of the workflow. Then using a single API, you can connect each step to create an end-to-end workflow. SageMaker Pipelines takes care of managing data between steps, packaging the code recipes, and orchestrating their execution, reducing months of coding to a few hours. Every time a workflow executes, a complete record of the data processed and actions taken is kept so data scientists and ML developers can quickly debug problems.

Q: How do I view all my trained models to choose the best model to move to production?

Amazon SageMaker Pipelines provides a central repository of trained models called a model registry. You can discover models and access the model registry visually through SageMaker Studio or programmatically through the Python SDK, making it easy to choose your desired model to deploy into production.

Q: What components of Amazon SageMaker can be added to Amazon SageMaker Pipelines?

The components available through Amazon SageMaker Studio, including SageMaker Clarify, SageMaker Data Wrangler, SageMaker Feature Store, SageMaker Experiments, SageMaker Debugger, and SageMaker Model Monitor, can be added to SageMaker Pipelines.

Q: How do I track my model components across the entire ML workflow?

Amazon SageMaker Pipelines automatically keeps track of all model constituents and keeps an audit trail of all changes, thereby eliminating manual tracking, and can help you achieve compliance goals. You can track data, code, trained models, and more with SageMaker Pipelines.

Q: How does the pricing for Amazon SageMaker Pipelines work?

There is no additional charge for Amazon SageMaker Pipelines. You pay only for the underlying compute or any separate AWS services you use within SageMaker Pipelines.

Q: Can I use Kubeflow with Amazon SageMaker?

Yes. Amazon SageMaker Components for Kubeflow Pipelines are open-source plugins that allow you to use Kubeflow Pipelines to define your ML workflows and use SageMaker for the

data labeling, training, and inference steps. Kubeflow Pipelines is an add-on to Kubeflow that lets you build and deploy portable and scalable end-to-end ML pipelines. However, when using Kubeflow Pipelines, ML ops teams need to manage a Kubernetes cluster with CPU and GPU instances and keep its utilization high at all times to reduce operational costs. Maximizing the utilization of a cluster across data science teams is challenging and adds additional operational overhead to the ML ops teams. As an alternative to an ML-optimized Kubernetes cluster, with Amazon SageMaker Components for Kubeflow Pipelines you can take advantage of powerful SageMaker features such as data labeling, fully managed large-scale hyperparameter tuning and distributed training jobs, one-click secure and scalable model deployment, and cost-effective training through Amazon EC2 Spot instances without needing to configure and manage Kubernetes clusters specifically to run the ML jobs.

Q: How does Amazon SageMaker Components for Kubeflow Pipelines pricing work?

There is no additional charge for using Amazon SageMaker Components for Kubeflow Pipelines.

Prepare data

Q: How can Amazon SageMaker prepare data for ML?

Amazon SageMaker Data Wrangler reduces the time it takes to aggregate and prepare data for ML. From a single interface in SageMaker Studio, you can import data from Amazon S3, Amazon Athena, Amazon Redshift, AWS Lake Formation, and Amazon SageMaker Feature Store, and in just a few clicks SageMaker Data Wrangler will automatically load, aggregate, and display the raw data. It will then make conversion recommendations based on the source data, transform the data into new features, validate the features, and provide visualizations with recommendations on how to remove common sources of error such as incorrect labels. Once your data is prepared, you can build fully automated ML workflows with Amazon SageMaker Pipelines or import that data into Amazon SageMaker Feature Store.

Q: How can I create model features with Amazon SageMaker Data Wrangler?

Without writing a single line of code, Amazon SageMaker Data Wrangler can automatically transform your data into new features. SageMaker Data Wrangler offers a selection of preconfigured data transforms, such as convert column type, one-hot encoding, impute missing data with mean or median, rescale columns, and data/time embeddings. For example, you can convert a text field column into a numerical column with a single click, or author custom transforms in PySpark, SQL, and Pandas.

Q: How can I visualize my data in Amazon SageMaker Data Wrangler?

Amazon SageMaker Data Wrangler helps you understand your data and identify potential errors and extreme values with a set of robust pre-configured visualization templates. Histograms, scatter plots, and ML-specific visualizations, such as target leakage detection, are all available without writing a single line of code. You can also create and edit your own visualizations.

Q: How does the pricing for Amazon SageMaker Data Wrangler work?

You pay for all ML compute, storage, and data processing resources you use for Amazon SageMaker Data Wrangler. You can review all the details of Amazon SageMaker Data Wrangler pricing here. As part of the AWS Free Tier, you can also get started with SageMaker Data Wrangler for free.

Q: How can I train machine learning models with data prepared in Amazon SageMaker Data Wrangler?

SageMaker Data Wrangler provides a unified experience enabling you to prepare data and seamlessly train a machine learning model in SageMaker Autopilot. Amazon SageMaker Autopilot automatically builds, trains, and tunes the best ML models based on your data. With SageMaker Autopilot, you still maintain full control and visibility of your data and model. SageMaker Data Wrangler also offers export options, including SageMaker Data Wrangler jobs, Amazon SageMaker Feature Store, and Amazon SageMaker Pipelines, providing you the ability to integrate your data preparation flow into your ML workflow.

Q: How do I store features for my ML models?

Amazon SageMaker Feature Store provides a central repository for data features with low latency (milliseconds) reads and writes. Features can be stored, retrieved, discovered, and shared through SageMaker Feature Store for easy reuse across models and teams with secure access and control. SageMaker Feature Store supports both online and offline features generated via batch or streaming pipelines. It supports backfilling the features and provides both online and offline stores to maintain parity between features used in model training and inference.

Q: How do I maintain consistency between online and offline features?

Amazon SageMaker Feature Store automatically maintains consistency between online and offline features without additional management or code. SageMaker Feature Store is fully managed and maintains consistency across training and inference environments.

Q: How can I reproduce a feature from a given moment in time?

Amazon SageMaker Feature Store maintains time stamps for all features at every instance of time. This helps you retrieve features at any period of time for business or compliance requirements. You can easily explain model features and their values from when they were first created to the present time by reproducing the model from a given moment in time.

Q: What are offline features?

Offline features are used for training because you need access to very large volumes over a long period of time. These features are served from a high-throughput, high-bandwidth repository.

Q: What are online features?

Online features are used in applications required to make real-time predictions. Online features are served from a high-throughput repository with single-digit millisecond latency for fast predictions.

Q: How does pricing work for Amazon SageMaker Feature Store?

You can get started with Amazon SageMaker Feature Store for free, as part of the AWS Free Tier. With SageMaker Feature Store, you pay for writing into the feature store, and reading and storage from the online feature store.

Q: What does Amazon SageMaker offer for data labeling?

Amazon SageMaker provides two data labeling offerings, Amazon SageMaker Ground Truth Plus and Amazon SageMaker Ground Truth. Both options allow you to identify raw data, such

as images, text files, and videos, and add informative labels to create high-quality training datasets for your ML models. To learn more, visit the SageMaker Data Labeling webpage.

Build models

Q: What are Amazon SageMaker Studio notebooks?

Amazon SageMaker Studio notebooks are collaborative, flexible, managed Jupyter notebooks that are part of Amazon SageMaker Studio, a fully integrated development environment for ML.

Q: How are SageMaker Studio notebooks different from the instance-based notebooks offering?

SageMaker Studio notebooks offer a few important features that differentiate them from the instance-based notebooks. With the Studio notebooks, you can quickly launch notebooks without needing to manually provision an instance and waiting for it to be operational. The startup time of launching the UI to read and execute a notebook is faster than the instance-based notebooks.

You also have the flexibility to choose from a large collection of instance types from within the UI at any time. You do not need to go to the AWS Management Console to start new instances and port over your notebooks.

Each user has an isolated home directory independent of a particular instance. This directory is automatically mounted into all notebook servers and kernels as they're started, so you can access your notebooks and other files even when you switch instances to view and run your notebooks.

SageMaker Studio notebooks are integrated with AWS IAM Identity Center (successor to AWS SSO), making it easy to use your organizational credentials to access the notebooks. Notebook sharing is an integrated feature in SageMaker Studio Notebooks. You can also share your notebooks with your peers using a single click.

Q: How do Amazon SageMaker Studio notebooks work?

Amazon SageMaker Studio notebooks are one-click Jupyter notebooks that can be spun quickly. The underlying compute resources are fully elastic, so you can easily dial up or down the available resources and the changes take place automatically in the background without interrupting your work. SageMaker also enables one-click sharing of notebooks. You can easily share notebooks with others and they'll get the exact same notebook, saved in the same place.

With SageMaker Studio notebooks you can sign in with your corporate credentials using AWS IAM Identity Center (successor to AWS SSO). Sharing notebooks within and across teams is easy, since the dependencies needed to run a notebook are automatically tracked in work images that are encapsulated with the notebook as it is shared.

Q: How do Amazon SageMaker Studio notebooks work with other AWS services?

Amazon SageMaker Studio Notebooks give you access to all SageMaker features, such as distributed training, batch transform, hosting, and experiment management. You can access other services such as datasets in Amazon S3, Amazon Redshift, AWS Glue, Amazon EMR, or AWS Lake Formation from SageMaker notebooks.

Q: How does SageMaker Studio notebooks pricing work?

You pay for both compute and storage when you use SageMaker Studio notebooks. See Amazon SageMaker Pricing for charges by compute instance type. Your notebooks and

associated artifacts such as data files and scripts are persisted on Amazon EFS. See Amazon EFS Pricing for storage charges. As part of the AWS Free Tier, you can get started with Amazon SageMaker Studio notebooks for free.

Q: Do I get charged separately for each notebook created and run in SageMaker Studio?

No. You can create and run multiple notebooks on the same compute instance. You pay only for the compute that you use, not for individual items. You can read more about this in our metering guide.

In addition to the notebooks, you can also start and run terminals and interactive shells in Studio, all on the same compute instance. Each application runs within a container or image. SageMaker Studio provides several built-in images purpose-built and preconfigured for data science and ML. You can read more about the Studio developer environment in the guide for using SageMaker Studio notebooks.

Q: How do I monitor and shut down the resources used by my notebooks?

You can monitor and shut down the resources used by your SageMaker Studio notebooks through both SageMaker Studio visual interface and the AWS Management Console. See the documentation for more details.

Q: I'm running a SageMaker Studio notebook. Will I still be charged if I close my browser, close the notebook tab, or just leave the browser open?

Yes, you will continue to be charged for the compute. This is similar to starting Amazon EC2 instances in the AWS Management Console and then closing the browser. The Amazon EC2 instances are still running and you still incur charges unless you explicitly shut down the instance.

Q: Do I get charged for creating and setting up a Studio domain?

No, you don't get charged for creating or configuring a Studio domain, including adding, updating, and deleting user profiles.

Q: How do I see the itemized charges for Studio notebooks or other SageMaker services?

As an admin, you can view the list of itemized charges for SageMaker, including Studio, in the AWS Billing console. From the AWS Management Console for SageMaker, choose Services on the top menu, type "billing" in the search box and select Billing from the dropdown, then select Bills on the left panel. In the Details section, you can click on SageMaker to expand the list of Regions and drill down to the itemized charges.

Q: What is Amazon SageMaker Studio Lab?

Amazon SageMaker Studio Lab is a free ML development environment that provides the compute, storage (up to 15 GB), and security—all at no cost—for anyone to learn and experiment with ML. All you need to get started is a valid email ID; you don't need to configure infrastructure or manage identity and access or even sign up for an AWS account. SageMaker Studio Lab accelerates model building through GitHub integration, and it comes preconfigured with the most popular ML tools, frameworks, and libraries to get you started immediately. SageMaker Studio Lab automatically saves your work so you don't need to restart between sessions. It's as easy as closing your laptop and coming back later.

Q: Why should I use Amazon SageMaker Studio Lab?

Amazon SageMaker Studio Lab is for students, researchers, and data scientists who need a free notebook development environment with no setup required for their ML classes and experiments. Amazon SageMaker Studio Lab is ideal for users who do not need a production environment but still want a subset of the SageMaker functionality to improve their ML skills. SageMaker sessions are automatically saved, enabling users to pick up where they left off for each user session.

Q: How does Amazon SageMaker Studio Lab work with other AWS services?

Amazon SageMaker Studio Lab is a service built on AWS and uses many of the same core services as Amazon SageMaker Studio, such as Amazon S3 and Amazon EC2. Unlike the other services, customers will not need an AWS account. Instead, they will create an Amazon SageMaker Studio Lab specific account with an email address. This will give the user access to a limited environment (15 GB of storage, and 12 hour sessions) for them to run ML notebooks.

Q: What is Amazon SageMaker Canvas?

Amazon SageMaker Canvas is a visual drag-and-drop service that allows business analysts to build ML models and generate accurate predictions without writing any code or requiring ML expertise. SageMaker Canvas makes it easy to access and combine data from a variety of sources, automatically clean data and apply a variety of data adjustments, and build ML models to generate accurate predictions with a single click. You can also easily publish results, explain and interpret models, and share models with others within your organization to review.

Q: What data sources does Amazon SageMaker Canvas support?

SageMaker Canvas enables you to seamlessly discover AWS data sources that your account has access to, including Amazon S3 and Amazon Redshift. You can browse and import data using the SageMaker Canvas visual drag-and-drop interface. Additionally, you can drag and drop files from your local disk, and use pre-built connectors to import data from third-party sources such as Snowflake.

Q: How do I build an ML model to generate accurate predictions?

Once you have connected sources, selected a dataset, and prepared your data, you can select the target column that you want to predict to initiate a model creation job. Amazon SageMaker Canvas will automatically identify the problem type, generate new relevant features, test a comprehensive set of prediction models using ML techniques such as linear regression, logistic regression, deep learning, time-series forecasting, and gradient boosting, and build the model that makes accurate predictions based on your dataset.

Q: How long does it take to build a model? How can I monitor progress during model creation?

The time it takes to build a model depends on the size of your dataset. Small datasets can take less than 30 minutes, and large datasets can take a few hours. As the model creation job progresses, Amazon SageMaker Canvas provides detailed visual updates, including percent job complete and the amount of time left for job completion.

Train models

Q: What is Amazon SageMaker Experiments?

Amazon SageMaker Experiments helps you organize and track iterations to ML models. SageMaker Experiments helps you manage iterations by automatically capturing the input parameters, configurations, and results, and storing them as "experiments". You can work within the visual interface of SageMaker Studio, where you can browse active experiments,

search for previous experiments by their characteristics, review previous experiments with their results, and compare experiment results visually.

Q: What is Amazon SageMaker Debugger?

Amazon SageMaker Debugger automatically captures real-time metrics during training, such as confusion matrices and learning gradients, to help improve model accuracy. The metrics from SageMaker Debugger can be visualized in SageMaker Studio for easy understanding. SageMaker Debugger can also generate warnings and remediation advice when common training problems are detected. SageMaker Debugger also automatically monitors and profiles system resources such as CPUs, GPUs, network, and memory in real time, and provides recommendations on re-allocation of these resources. This enables you to use your resources efficiently during training and helps reduce costs and resources.

Q: Does Amazon SageMaker support distributed training?

Yes. Amazon SageMaker can automatically distribute deep learning models and large training sets across AWS GPU instances in a fraction of the time it takes to build and optimize these distribution strategies manually. The two distributed training techniques that SageMaker applies are data parallelism and model parallelism. Data parallelism is applied to improve training speeds by dividing the data equally across multiple GPU instances, allowing each instance to train concurrently. Model parallelism is useful for models too large to be stored on a single GPU and require the model to be partitioned into smaller parts before distributing across multiple GPUs. With only a few lines of additional code in your PyTorch and TensorFlow training scripts, SageMaker will automatically apply data parallelism or model parallelism for you, allowing you to develop and deploy your models faster. SageMaker will determine the best approach to split your model by using graph partitioning algorithms to balance the computation of each GPU while minimizing the communication between GPU instances. SageMaker also optimizes your distributed training jobs through algorithms that fully utilize the AWS compute and network in order to achieve near-linear scaling efficiency, which allows you to complete training faster than manual open-source implementations.

Q: What is SageMaker Training Compiler?

Amazon SageMaker Training Compiler is a deep learning (DL) compiler that accelerates DL model training by up to 50 percent through graph- and kernel-level optimizations to use GPUs more efficiently. Training Compiler is integrated with versions of TensorFlow and PyTorch in SageMaker, so you can speed up training in these popular frameworks with minimal code changes.

Q: How does SageMaker Training Compiler work?

SageMaker Training Compiler accelerates training jobs by converting DL models from their high-level language representation to hardware-optimized instructions that train faster than jobs with the native frameworks. More specifically, SageMaker Training Compiler uses graph-level optimization (operator fusion, memory planning, and algebraic simplification), data flow-level optimizations (layout transformation, common sub-expression elimination), and backend optimizations (memory latency hiding, loop oriented optimizations) to produce an optimized model training job that more efficiently uses hardware resources and, as a result, trains faster.

Q: How can I use SageMaker Training Compiler?

Amazon SageMaker Training Compiler is built into the SageMaker Python SDK and SageMaker Hugging Face Deep Learning Containers. You don’t need to change your workflows to access its speedup benefits. You can run training jobs in the same way as you already do, using any of the SageMaker interfaces: Amazon SageMaker notebook instances, SageMaker Studio, AWS

SDK for Python (Boto3), and AWS Command Line Interface. You can enable SageMaker Training Compiler by adding a TrainingCompilerConfig class as a parameter when you create a framework estimator object. Practically, this means a couple of lines of code added to your existing training job script for a single GPU instance. Most up-to-date detailed documentation, sample notebooks, and examples are available in the documentation.

Q: What is the pricing of SageMaker Training Compiler?

Training Compiler is a SageMaker Training feature and is provided at no additional charge exclusively to SageMaker customers. Customers can actually reduce their costs with Training Compiler as training times are reduced.

Q: What is Managed Spot Training?

Managed Spot Training with Amazon SageMaker lets you train your ML models using Amazon EC2 Spot instances, while reducing the cost of training your models by up to 90%.

Q: How do I use Managed Spot Training?

You enable the Managed Spot Training option when submitting your training jobs and you also specify how long you want to wait for Spot capacity. Amazon SageMaker will then use Amazon EC2 Spot instances to run your job and manages the Spot capacity. You have full visibility into the status of your training jobs, both while they are running and while they are waiting for capacity.

Q: When should I use Managed Spot Training?

Managed Spot Training is ideal when you have flexibility with your training runs and when you want to minimize the cost of your training jobs. With Managed Spot Training, you can reduce the cost of training your ML models by up to 90%.

Q: How does Managed Spot Training work?

Managed Spot Training uses Amazon EC2 Spot instances for training, and these instances can be pre-empted when AWS needs capacity. As a result, Managed Spot Training jobs can run in small increments as and when capacity becomes available. The training jobs need not be restarted from scratch when there is an interruption, as Amazon SageMaker can resume the training jobs using the latest model checkpoint. The built-in frameworks and the built-in computer vision algorithms with SageMaker enable periodic checkpoints, and you can enable checkpoints with custom models.

Q: Do I need to periodically checkpoint with Managed Spot Training?

We recommend periodic checkpoints as a general best practice for long-running training jobs. This prevents your Managed Spot Training jobs from restarting if capacity is pre-empted. When you enable checkpoints, Amazon SageMaker resumes your Managed Spot Training jobs from the last checkpoint.

Q: How do you calculate the cost savings with Managed Spot Training jobs?

Once a Managed Spot Training job is completed, you can see the savings in the AWS Management Console and also calculate the cost savings as the percentage difference between the duration for which the training job ran and the duration for which you were billed.

Regardless of how many times your Managed Spot Training jobs are interrupted, you are charged only once for the duration for which the data was downloaded.

Q: Which instances can I use with Managed Spot Training?

Managed Spot Training can be used with all instances supported in Amazon SageMaker.

Q: Which AWS Regions are supported with Managed Spot Training?

Managed Spot Training is supported in all AWS Regions where Amazon SageMaker is currently available.

Q: Are there limits to the size of the dataset I can use for training?

There are no fixed limits to the size of the dataset you can use for training models with Amazon SageMaker.

Q: What algorithms does Amazon SageMaker use to generate models?

Amazon SageMaker includes built-in algorithms for linear regression, logistic regression, k-means clustering, principal component analysis, factorization machines, neural topic modeling, latent dirichlet allocation, gradient boosted trees, sequence2sequence, time-series forecasting, word2vec, and image classification. SageMaker also provides optimized Apache MXNet, Tensorflow, Chainer, PyTorch, Gluon, Keras, Horovod, Scikit-learn, and Deep Graph Library containers. In addition, Amazon SageMaker supports your custom training algorithms provided through a Docker image adhering to the documented specification.

Q: What is Automatic Model Tuning?

Most ML algorithms expose a variety of parameters that control how the underlying algorithm operates. Those parameters are generally referred to as hyperparameters and their values affect the quality of the trained models. Automatic model tuning is the process of finding a set of hyperparameters for an algorithm that can yield an optimal model.

Q: What models can be tuned with Automatic Model Tuning?

You can run automatic model tuning in Amazon SageMaker on top of any algorithm as long as it's scientifically feasible, including built-in SageMaker algorithms, deep neural networks, or arbitrary algorithms you bring to SageMaker in the form of Docker images.

Q: Can I use Automatic Model Tuning outside of Amazon SageMaker?

Not at this time. The best model tuning performance and experience is within Amazon SageMaker.

Q: What is the underlying tuning algorithm?

Currently, our algorithm for tuning hyperparameters is a customized implementation of Bayesian Optimization. It aims to optimize a customer-specified objective metric throughout the tuning process. Specifically, it checks the object metric of completed training jobs, and uses the knowledge to infer the hyperparameter combination for the next training job.

Q: Will you recommend specific hyperparameters for tuning?

No. How certain hyperparameters impact the model performance depends on various factors, and it is hard to definitively say one hyperparameter is more important than the others and thus needs to be tuned. For built-in algorithms within Amazon SageMaker, we do call out whether or not a hyperparameter is tunable.

Q: How long does a hyperparameter tuning job take?

The length of time for a hyperparameter tuning job depends on multiple factors, including the size of the data, the underlying algorithm, and the values of the hyperparameters. Additionally, customers can choose the number of simultaneous training jobs and total number of training jobs. All these choices affect how long a hyperparameter tuning job can last.

Q: Can I optimize multiple objectives simultaneously, such as optimizing a model to be both fast and accurate?

Not at this time. Currently, you need to specify a single objective metric to optimize or change your algorithm code to emit a new metric, which is a weighted average between two or more useful metrics, and have the tuning process optimize towards that objective metric.

Q: How much does Automatic Model Tuning cost?

There is no charge for a hyperparameter tuning job itself. You will be charged by the training jobs that are launched by the hyperparameter tuning job, based on model training pricing.

Q: How do I decide to use Amazon SageMaker Autopilot or Automatic Model Tuning?

Amazon SageMaker Autopilot automates everything in a typical ML workflow, including feature preprocessing, algorithm selection, and hyperparameter tuning, while specifically focusing on classification and regression use cases. Automatic Model Tuning, on the other hand, is designed to tune any model, no matter whether it is based on built-in algorithms, deep learning frameworks, or custom containers. In exchange for the flexibility, you have to manually pick the specific algorithm, hyperparameters to tune, and corresponding search ranges.

Q: What is reinforcement learning?

Reinforcement learning is a ML technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.

Q: Can I train reinforcement learning models in Amazon SageMaker?

Yes, you can train reinforcement learning models in Amazon SageMaker in addition to supervised and unsupervised learning models.

Q: How is reinforcement learning different from supervised learning?

Though both supervised and reinforcement learning use mapping between input and output, unlike supervised learning where the feedback provided to the agent is the correct set of actions for performing a task, reinforcement learning uses a delayed feedback where reward signals are optimized to ensure a long-term goal through a sequence of actions.

Q: When should I use reinforcement learning?

While the goal of supervised learning techniques is to find the right answer based on the patterns in the training data, the goal of unsupervised learning techniques is to find similarities and differences between data points. In contrast, the goal of reinforcement learning (RL) techniques is to learn how to achieve a desired outcome even when it is not clear how to accomplish that outcome. As a result, RL is more suited to enabling intelligent applications where an agent can make autonomous decisions such as robotics, autonomous vehicles, HVAC, industrial control, and more.

Q: What type of environments can I use for training RL models?

Amazon SageMaker RL supports a number of different environments for training RL models. You can use AWS services such as AWS RoboMaker, open-source environments or custom

environments developed using Open AI Gym interfaces, or commercial simulation environments such as MATLAB and SimuLink.

Q: Do I need to write my own RL agent algorithms to train RL models?

No, Amazon SageMaker RL includes RL toolkits such as Coach and Ray RLLib that offer implementations of RL agent algorithms such as DQN, PPO, A3C, and many more.

Q: Can I bring my own RL libraries and algorithm implementation and run them in Amazon SageMaker RL?

Yes, you can bring your own RL libraries and algorithm implementations in Docker Containers and run those in Amazon SageMaker RL.

Q: Can I do distributed rollouts using Amazon SageMaker RL?

Yes. You can even select a heterogeneous cluster where the training can run on a GPU instance and the simulations can run on multiple CPU instances.

Deploy models

Q: What deployment options does Amazon SageMaker provide?

After you build and train models, Amazon SageMaker provides three options to deploy them so you can start making predictions. Real-time inference is suitable for workloads with millisecond latency requirements, payload sizes up to 6 MB, and processing times of up to 60 seconds. Batch transform is ideal for offline predictions on large batches of data that are available up front. Asynchronous inference is designed for workloads that do not have sub-second latency requirements, payload sizes up to 1 GB, and processing times of up to 15 minutes.

Q: What is SageMaker Asynchronous Inference?

Amazon SageMaker Asynchronous Inference queues incoming requests and processes them asynchronously. This option is ideal for requests with large payload sizes and/or long processing times that need to be processed as they arrive. Optionally, you can configure auto-scaling settings to scale down the instance count to zero when not actively processing requests to save on costs.

Q: How do I configure auto-scaling settings to scale down the instance count to zero when not actively processing requests?

You can scale down the Amazon SageMaker Asynchronous Inference endpoint instance count to zero in order to save on costs when you are not actively processing requests. You need to define a scaling policy that scales on the "ApproximateBacklogPerInstance" custom metric and set the "MinCapacity" value to zero. For step-by-step instructions, please visit the autoscale an asynchronous endpoint section of the developer guide.

Q: What is Amazon SageMaker Serverless Inference?

Amazon SageMaker Serverless Inference is a purpose-built serverless model serving option that makes it easy to deploy and scale ML models. SageMaker Serverless Inference endpoints automatically start the compute resources and scale them in and out depending on traffic, eliminating the need for you to choose instance type, run provisioned capacity, or manage scaling. You can optionally specify the memory requirements for your serverless inference endpoint. You pay only for the duration of running the inference code and the amount of data processed, not for idle periods.

Q: Why should I use Amazon SageMaker Serverless Inference?

Amazon SageMaker Serverless Inference simplifies the developer experience by eliminating the need to provision capacity up front and manage scaling policies. SageMaker Serverless Inference can scale instantly from tens to thousands of inferences within seconds based on the usage patterns, making it ideal for ML applications with intermittent or unpredictable traffic. For example, a chatbot service used by a payroll processing company experiences an increase in inquiries at the end of the month while for rest of the month traffic is intermittent. Provisioning instances for the entire month in such scenarios is not cost-effective, as you end up paying for idle periods. SageMaker Serverless Inference helps address these types of use cases by providing you automatic and fast scaling out of the box without the need for you to forecast traffic up front or manage scaling policies. Additionally, you pay only for the compute time to run your inference code (billed in milliseconds) and for data processing, making it a cost-effective option for workloads with intermittent traffic.

Q: What is Amazon SageMaker Inference Recommender?

Amazon SageMaker Inference Recommender is a new capability of Amazon SageMaker that reduces the time required to get ML models in production by automating performance benchmarking and tuning model performance across SageMaker ML instances. You can now use SageMaker Inference Recommender to deploy your model to an endpoint that delivers the best performance and minimizes cost. You can get started with SageMaker Inference Recommender in minutes while selecting an instance type and get recommendations for optimal endpoint configurations within hours, eliminating weeks of manual testing and tuning time. With SageMaker Inference Recommender, you pay only for the SageMaker ML instances used during load testing, and there are no additional charges.

Q: Why should I use SageMaker Inference Recommender?

You should use SageMaker Inference Recommender if you need recommendations for the right endpoint configuration to improve performance and reduce costs. Previously, data scientists who wanted to deploy their models had to run manual benchmarks to select the right endpoint configuration. They had to first select the right ML instance type out of the 70+ available instance types based on the resource requirements of their models and sample payloads, and then optimize the model to account for differing hardware. Then, they had to conduct extensive load tests to validate that latency and throughput requirements are met and that the costs are low. SageMaker Inference Recommender eliminates this complexity by making it easy for you to: 1) get started in minutes with an instance recommendation; 2) conduct load tests across instance types to get recommendations on your endpoint configuration within hours; and 3) automatically tune container and model server parameters as well as perform model optimizations for a given instance type.

Q: How does SageMaker Inference Recommender work with other AWS services?

Data scientists can access SageMaker Inference Recommender from SageMaker Studio, AWS SDK for Python (Boto3), or AWS CLI. They can get deployment recommendations within SageMaker Studio in the SageMaker model registry for registered model versions. Data scientists can search and filter the recommendations through SageMaker Studio, AWS SDK, or AWS CLI.

Q: Can SageMaker Inference Recommender support multi-model endpoints or multi-container endpoints?

No, we currently support only a single model per endpoint.

Q: What type of endpoints does SageMaker Inference Recommender support?

Currently we support only real-time endpoints.

Q: Can I use SageMaker Inference Recommender in one Region and benchmark in different Regions?

At launch, we will support all Regions supported by SageMaker, except the AWS China Regions.

Q: Does SageMaker Inference Recommender support Amazon EC2 Inf1 instances?

Yes, we support all types of containers. Amazon EC2 Inf1, based on the AWS Inferentia chip, requires a compiled model artifact using either the Neuron compiler or SageMaker Neo. Once you have a compiled model for an Inferentia target and the associated container image URI, you can use SageMaker Inference Recommender to benchmark different Inferentia instance types.

Q: What is Amazon SageMaker Model Monitor?

Amazon SageMaker Model Monitor allows developers to detect and remediate concept drift. SageMaker Model Monitor automatically detects concept drift in deployed models and provides detailed alerts that help identify the source of the problem. All models trained in SageMaker automatically emit key metrics that can be collected and viewed in SageMaker Studio. From inside SageMaker Studio, you can configure data to be collected, how to view it, and when to receive alerts.

Q: Can I access the infrastructure that Amazon SageMaker runs on?

No. Amazon SageMaker operates the compute infrastructure on your behalf, allowing it to perform health checks, apply security patches, and do other routine maintenance. You can also deploy the model artifacts from training with custom inference code in your own hosting environment.

Q: How do I scale the size and performance of an Amazon SageMaker model once in production?

Amazon SageMaker hosting automatically scales to the performance needed for your application using Application Auto Scaling. In addition, you can manually change the instance number and type without incurring downtime by modifying the endpoint configuration.

Q: How do I monitor my Amazon SageMaker production environment?

Amazon SageMaker emits performance metrics to Amazon CloudWatch Metrics so you can track metrics, set alarms, and automatically react to changes in production traffic. In addition, Amazon SageMaker writes logs to Amazon CloudWatch Logs to let you monitor and troubleshoot your production environment.

Q: What kinds of models can be hosted with Amazon SageMaker?

Amazon SageMaker can host any model that adheres to the documented specification for inference Docker images. This includes models created from Amazon SageMaker model artifacts and inference code.

Q: How many concurrent real-time API requests does Amazon SageMaker support?

Amazon SageMaker is designed to scale to a large number of transactions per second. The precise number varies based on the deployed model and the number and type of instances to which the model is deployed.

Q: What is Batch Transform?

Batch Transform enables you to run predictions on large or small batch data. There is no need to break down the dataset into multiple chunks or manage real-time endpoints. With a simple API, you can request predictions for a large number of data records and transform the data quickly and easily.

Q: What is Amazon SageMaker Edge Manager?

Amazon SageMaker Edge Manager is a capability in Amazon SageMaker that makes it easier to optimize, secure, monitor, and maintain ML models on fleets of edge devices such as smart cameras, robots, personal computers, and mobile devices. SageMaker Edge Manager helps ML developers operate ML models on a variety of edge devices at scale.

Q: How do I get started with SageMaker Edge Manager?

To get started with SageMaker Edge Manager, you need to compile and package your trained ML models in the cloud, register your devices, and prepare your devices with the SageMaker Edge Manager SDK. To prepare your model for deployment, SageMaker Edge Manager uses SageMaker Neo to compile your model for your target edge hardware. Once a model is compiled, SageMaker Edge Manager signs the model with an AWS generated key, then packages the model with its runtime and your necessary credentials to get it ready for deployment. On the device side, you register your device with SageMaker Edge Manager, download the SageMaker Edge Manager SDK, and then follow the instructions to install the SageMaker Edge Manager agent on your devices. The tutorial notebook provides a step-by-step example of how you can prepare the models and connect your models on edge devices with SageMaker Edge Manager.

Q: What devices are supported by SageMaker Edge Manager?

Amazon SageMaker Edge Manager supports common CPU (ARM, x86) and GPU (ARM, Nvidia) based devices with Linux and Windows operating systems. Over time, SageMaker Edge Manager will expand to support more embedded processors and mobile platforms that are also supported by SageMaker Neo.

Q: Do I need to use Amazon SageMaker to train my model in order to use Amazon SageMaker Edge Manager?

No, you do not. You can train your models elsewhere or use a pre-trained model from open source or from your model vendor.

Q: Do I need to use Amazon SageMaker Neo to compile my model in order to use Amazon SageMaker Edge Manager?

Yes, you do. Amazon SageMaker Neo converts and compiles your models into an executable that you can then package and deploy on your edge devices. Once the model package is deployed, the Amazon SageMaker Edge Manager agent will unpack the model package and run the model on the device.

Q: How do I deploy models to the edge devices?

Amazon SageMaker Edge Manager stores the model package in your specified Amazon S3 bucket. You can use the over-the-air (OTA) deployment feature provided by AWS IoT Greengrass or any other deployment mechanism of your choice to deploy the model package from your S3 bucket to the devices.

Q: How is Amazon SageMaker Edge Manager SDK different from the SageMaker Neo runtime (dlr)?

Neo dlr is an open-source runtime that only runs models compiled by the Amazon SageMaker Neo service. Compared to the open source dlr, the SageMaker Edge Manager SDK includes an enterprise grade on-device agent with additional security, model management, and model serving features. The SageMaker Edge Manager SDK is suitable for production deployment at scale.

Q: How is Amazon SageMaker Edge Manager related to AWS IoT Greengrass?

Amazon SageMaker Edge Manager and AWS IoT Greengrass can work together in your IoT solution. Once your ML model is packaged with SageMaker Edge Manager, you can use AWS IoT Greengrass's OTA update feature to deploy the model package to your device. AWS IoT Greengrass allows you to monitor your IoT devices remotely, while SageMaker Edge Manager helps you monitor and maintain the ML models on the devices.

Q: How is Amazon SageMaker Edge Manager related to AWS Panorama? When should I use Amazon SageMaker Edge Manager versus AWS Panorama?

AWS offers the most breadth and depth of capabilities for running models on edge devices. We have services to support a wide range of use cases, including computer vision, voice recognition, and predictive maintenance.

For companies looking to run computer vision on edge devices such as cameras and appliances, you can use AWS Panorama. Panorama offers ready-to-deploy computer vision applications for edge devices. It's easy to get started with AWS Panorama by logging into the cloud console, specifying the model you would like to use in Amazon S3 or in SageMaker, and then writing business logic as a Python script. AWS Panorama compiles the model for the target device and creates an application package so it can be deployed to your devices with just a few clicks. In addition, independent software vendors who want to build their own custom applications can use the AWS Panorama SDK, and device manufacturers can use the Device SDK to certify their devices for AWS Panorama.

Customers who want to build their own models and have more granular control over model features can use Amazon SageMaker Edge Manager. SageMaker Edge Manager is a managed service to prepare, run, monitor, and update ML models across fleets of edge devices such as smart cameras, smart speakers, and robots for any use case such as natural langue processing, fraud detection, and predictive maintenance. SageMaker Edge Manager is for ML edge developers who want control over their model, including engineering different model features and monitoring models for drift. Any ML edge developer can use SageMaker Edge Manager through the SageMaker console and the SageMaker APIs. SageMaker Edge Manager brings the capabilities of SageMaker to build, train, and deploy models in the cloud to edge devices.

Q: In which AWS Regions is Amazon SageMaker Edge Manager available?

Amazon SageMaker Edge Manager is available in six AWS Regions: US East (N. Virginia), US East (Ohio), US West (Oregon), EU (Ireland), EU (Frankfurt), and Asia Pacific (Tokyo). For details, see the AWS Regions Table.

Q: What is Amazon SageMaker Neo?

Amazon SageMaker Neo enables ML models to train once and run anywhere in the cloud and at the edge. SageMaker Neo automatically optimizes models built with popular deep learning frameworks that can be used to deploy on multiple hardware platforms. Optimized models run up to 25 times faster and consume less than a tenth of the resources of typical ML models.

Q: How do I get started with Amazon SageMaker Neo?

To get started with Amazon SageMaker Neo, log into the Amazon SageMaker console, choose a trained model, follow the example to compile models, and deploy the resulting model onto your target hardware platform.

Q: What are the major components of Amazon SageMaker Neo?

Amazon SageMaker Neo contains two major components: a compiler and a runtime. First, the Neo compiler reads models exported by different frameworks. It then converts the framework-specific functions and operations into a framework-agnostic intermediate representation. Next, it performs a series of optimizations. Then, the compiler generates binary code for the optimized operations and writes them to a shared object library. The compiler also saves the model definition and parameters into separate files. During execution, the Neo runtime loads the artifacts generated by the compiler—model definition, parameters, and the shared object library to run the model.

Q: Do I need to use Amazon SageMaker to train my model in order to use Amazon SageMaker Neo to convert the model?

No. You can train models elsewhere and use Neo to optimize them for Amazon SageMaker ML instances or AWS IoT Greengrass supported devices.

Q: Which models does Amazon SageMaker Neo support?

Currently, Amazon SageMaker Neo supports the most popular deep learning models that power computer vision applications and the most popular decision tree models used in Amazon SageMaker today. Neo optimizes the performance of AlexNet, ResNet, VGG, Inception, MobileNet, SqueezeNet, and DenseNet models trained in MXNet and TensorFlow, and classification and random cut forest models trained in XGBoost.

Q: Which hardware platforms does Amazon SageMaker Neo support?

You can find the lists of supported cloud instances, edge devices, and framework versions in the Amazon SageMaker Neo documentation.

Q: In which AWS Regions is Amazon SageMaker Neo available?

To see a list of supported Regions, view the AWS Region table.

Amazon SageMaker Savings Plans

Q: What are Amazon SageMaker Savings Plans?

Amazon SageMaker Savings Plans offer a flexible usage-based pricing model for Amazon SageMaker in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a one- or three-year term. Amazon SageMaker Savings Plans provide the most flexibility and help to reduce your costs by up to 64%. These plans automatically apply to eligible SageMaker ML instance usages, including SageMaker Studio notebooks, SageMaker On-Demand notebooks, SageMaker Processing, SageMaker Data Wrangler, SageMaker Training, SageMaker Real-Time Inference, and SageMaker Batch Transform regardless of instance family, size, or Region. For example, you can change usage from a CPU instance ml.c5.xlarge running in US East (Ohio) to an ml.Inf1 instance in US West (Oregon) for inference workloads at any time and automatically continue to pay the Savings Plans price.

Q: Why should I use Amazon SageMaker Savings Plans?

If you have a consistent amount of Amazon SageMaker instance usage (measured in $/hour) and use multiple SageMaker components or expect your technology configuration (such as instance family, or Region) to change over time, SageMaker Savings Plans make it simpler to maximize your savings while providing flexibility to change the underlying technology configuration based on application needs or new innovation. The Savings Plans rate applies automatically to all eligible ML instance usage with no manual modifications required.

Q: How can I get started with Amazon SageMaker Savings Plans?

You can get started with Savings Plans from AWS Cost Explorer in the AWS Management Console or by using the API/CLI. You can easily make a commitment to Savings Plans by using the recommendations provided in AWS Cost Explorer to realize the biggest savings. The recommended hourly commitment is based on your historical On-Demand usage and your choice of plan type, term length, and payment option. Once you sign up for a Savings Plan, your compute usage will automatically be charged at the discounted Savings Plans prices and any usage beyond your commitment will be charged at regular On-Demand rates.

Q: How are Savings Plans for Amazon SageMaker different from Compute Savings Plans for Amazon EC2?

The difference between Savings Plans for Amazon SageMaker and Savings Plans for EC2 is in the services they include. SageMaker Savings Plans apply only to SageMaker ML Instance usage.

Q: How do Savings Plans work with AWS Organizations/Consolidated Billing?

Savings Plans can be purchased in any account within an AWS Organization/Consolidated Billing family. By default, the benefit provided by Savings Plans is applicable to usage across all accounts within an AWS Organization/Consolidated Billing family. However, you can also choose to restrict the benefit of Savings Plans to only the account that purchased them.

Amazon SageMaker (AMS SSPS)

SageMaker provides every developer and data scientist with the ability to build, train, and deploy machine learning models quickly. Amazon SageMaker is a fully-managed service that covers the entire machine learning workflow to label and prepare your data, choose an algorithm, train the model, tune and optimize it for deployment, make predictions, and take action. Your models get to production faster with much less effort and lower cost.

SageMaker in AWS Managed Services FAQs

Common questions and answers:

Q: How do I request access to SageMaker in my AMS account?

Request access by submitting a Management | AWS service | Self-provisioned service | Add (ct-1w8z66n899dct) change type. This RFC provisions the following IAM roles to your account: customer_sagemaker_admin_role and service

role AmazonSageMaker-ExecutionRole-Admin. After SageMaker is provisioned in your account, you must onboard the customer_sagemaker_admin_role role in your federation solution. The service role cannot be accessed by you directly; the SageMaker service uses it while doing various actions as described here: Passing Roles.

Q: What are the restrictions to using SageMaker in my AMS account?

- The following use cases are not supported by the AMS Amazon SageMaker IAM role:
  - SageMaker Studio is not supported at this time.
  - SageMaker Ground Truth to manage private workforces is not supported since this feature requires overly permissive access to Amazon Cognito resources. If managing a private workforce is required, you can request a custom IAM role with combined SageMaker and Amazon Cognito permissions. Otherwise, we recommend using public workforce (backed by Amazon Mechanical Turk), or AWS Marketplace service providers, for data labeling.
- Creating VPC Endpoints to support API calls to SageMaker services (aws.sagemaker.{region}.notebook, com.amazonaws.{region}.sagemaker.api & com.amazonaws.{region}.sagemaker.runtime) is not supported as permissions can't be scoped down to SageMaker related services only. To support this use case, submit a Management | Other | Other RFC to create related VPC endpoints.
- SageMaker endpoint auto scaling is not supported as SageMaker requires DeleteAlarm permissions on any ("*") resource. To support endpoint auto scaling, submit a Management | Other | Other RFC to setup auto scaling for a SageMaker endpoint.

Q: What are the prerequisites or dependencies to using SageMaker in my AMS account?

- The following use cases require special configuration prior to use:
  - If an S3 bucket will be used to store model artifacts and data, then you must request an S3 bucket named with the required keywords ("SageMaker", "Sagemaker", "sagemaker" or "aws-glue") with a Deployment | Advanced stack components | S3 storage | Create RFC.
  - If Elastic File Store (EFS) will be used, then EFS storage must be configured in the same subnet, and allowed by security groups.

    - If other resources require direct access to SageMaker services (notebooks, API, runtime, and so on), then configuration must be requested by:
        - Submitting an RFC to create a security group for the endpoint (Deployment | Advanced stack components | Security group | Create (auto)).
        - Submitting a Management | Other | Other | Create RFC to set up related VPC endpoints.

Q: What are the supported naming conventions for resources that the customer_sagemaker_admin_role can access directly? (The following are for update and delete permissions; if you require additional supported naming conventions for your resources, reach out to an AMS Cloud Architect for consultation.)

- Resource: Passing AmazonSageMaker-ExecutionRole-* role
    - Permissions: The SageMaker self-provisioned service role supports your use of the SageMaker service role (AmazonSageMaker-ExecutionRole-*) with AWS Glue, AWS RoboMaker, and AWS Step Functions.
- Resource: Secrets on AWS Secrets Manager
    - Permissions: Describe, Create, Get, Update secrets with a AmazonSageMaker-* prefix.
    - Permissions: Describe, Get secrets when the SageMaker resource tag is set to true.
- Resource: Repositories on AWS CodeCommit
    - Permissions: Create/ delete repositories with a AmazonSageMaker-* prefix.
    - Permissions: Git Pull/Push on repositories with following prefixes, *sagemaker*, *SageMaker*, and *Sagemaker*.
- Resource: Amazon ECR (Amazon Elastic Container Registry) Repositories
    - Permissions: Permissions: Set, delete repository policies, and upload container images, when the following resource naming convention is used, *sagemaker*.
- Resource: Amazon S3 buckets
    - Permissions: Get, Put, Delete object, abort multipart upload S3 objects when resources have the following prefixes: *SageMaker*, *Sagemaker*, *sagemaker* and aws-glue.
    - Permissions: Get S3 objects when the SageMaker tag is set to true.
- Resource: Amazon CloudWatch Log Group

    - Permissions: Create Log Group or Stream, Put Log Event, List, Update, Create , Delete log delivery with following prefix: /aws/sagemaker/*.
- Resource: Amazon CloudWatch Metric
    - Permissions: Put metric data when the following prefixes are used: AWS/SageMaker, AWS/SageMaker/, aws/SageMaker, aws/SageMaker/, aws/sagemaker, aws/sagemaker/, and /aws/sagemaker/..
- Resource: Amazon CloudWatch Dashboard
    - Permissions: Create/Delete dashboards when the following prefixes are used: customer_*.
- Resource: Amazon SNS (Simple Notification Service) topic
    - Permissions: Subscribe/Create topic when following prefixes are used: *sagemaker*, *SageMaker*, and *Sagemaker*.

Q: What's the difference between AmazonSageMakerFullAccess and customer_sagemaker_admin_role?

The customer_sagemaker_admin_role with the customer_sagemaker_admin_policy provides almost the same permissions as AmazonSageMakerFullAccess except:

- Permission to connect with RoboMaker, Cognito, and Glue resources.
- Sagemaker endpoint autoscaling. You must submit a Management | Other | Other | Update RFC to elevate to autoscaling permissions temporarily, or permanently, as autoscaling requires permissive access on CloudWatch service.

Q: How do I adopt KMS CMKs in data encryption at rest?

You must ensure that the key policy has been set up properly on the CMKs so that related IAM users/roles can use the keys. For more information, see the AWS KMS Key Policy document.

Amazon Rekognition

Automate your image and video analysis with machine learning.

What can Amazon Rekognition do?

Amazon Rekognition can enable the business and development teams to solve the most pressing computer vision needs – no ML skills required.

Features

Amazon Rekognition offers pre-trained and customizable computer vision (CV) capabilities to extract information and insights from your images and videos.

Content moderation
Detect potentially unsafe, inappropriate, or unwanted content across images and videos.

Labels
Detect objects, scenes, activities, landmarks, dominant colors, and image quality.

Celebrity recognition
Identify well-known people to catalog photos and footage for media, marketing, and advertising.

Face compare and search
Determine the similarity of a face against another picture or from your private image repository.

Custom labels
Detect custom objects such as brand logos using automated machine learning (AutoML) to train your models with as few as 10 images.

Video segment detection
Detect key segments in videos, such as black frames, start or end credits, slates, color bars, and shots.

Face detection and analysis
Detect faces appearing in images and videos and recognize attributes such as open eyes, glasses, and facial hair for each.

Text detection
Extract skewed and distorted text from images and videos of street signs, social media posts, and product packaging.

Streaming Video Events detection
Detect objects such as packages, pets, or people in real-time from live video streams.

Use cases

Detect inappropriate content
Quickly and accurately identify unsafe or inappropriate content across image and video assets based on general or business-specific standards and practices.

Verify identity online
Use facial comparison and analysis in your user onboarding and authentication workflows to remotely verify the identity of opted-in users.

Streamline media analysis
Automatically detect key video segments to reduce the time, effort, and costs of video ad insertion, content operations, and content production.

Send connected home smart alerts
Deliver timely and actionable alerts when a desired object is detected in your live video streams. Create home automation experiences such as automatically turning on the light when a person is detected.

What is Amazon Rekognition?

Amazon Rekognition makes it easy to add image and video analysis to your applications. You just provide an image or video to the Amazon Rekognition API, and the service can identify objects, people, text, scenes, and activities. It can detect any inappropriate content as well. Amazon Rekognition also provides highly accurate facial analysis, face comparison, and face search capabilities. You can detect, analyze, and compare faces for a wide variety of use cases, including user verification, cataloging, people counting, and public safety.

Amazon Rekognition is based on the same proven, highly scalable, deep learning technology developed by Amazon's computer vision scientists to analyze billions of images and videos daily. It requires no machine learning expertise to use. Amazon Rekognition includes a simple, easy-to-use API that can quickly analyze any image or video file that's stored in Amazon S3. Amazon Rekognition is always learning from new data, and we're continually adding new labels and facial comparison features to the service. For more information, see the Amazon Rekognition FAQs.

Common use cases for using Amazon Rekognition include the following:

- **Searchable image and video libraries** – Amazon Rekognition makes images and stored videos searchable so you can discover objects and scenes that appear within them.

- **Face-based user verification** – Amazon Rekognition enables your applications to confirm user identities by comparing their live image with a reference image.

- **Detection of Personal Protective Equipment**

  Amazon Rekognition detects Personal Protective Equipment (PPE) such as face covers, head covers, and hand covers on persons in images. You can use PPE detection where safety is the highest priority. For example, industries such as construction, manufacturing, healthcare, food processing, logistics, and retail. With PPE detection, you can automatically detect if a person is wearing a specific type of PPE. You can use the detection results to send a notification or to identify places where safety warnings or training practices can be improved.

- **Sentiment and demographic analysis** – Amazon Rekognition interprets emotional expressions such as happy, sad, or surprise, and demographic information such as gender from facial images. Amazon Rekognition can analyze images, and send the emotion and demographic attributes to Amazon Redshift for periodic reporting on trends such as in store locations and similar scenarios. Note that a prediction of an emotional expression is based on the physical appearance of a person's face only. It is not indicative of a person’s internal emotional state, and Rekognition should not be used to make such a determination.

- **Facial Search** – With Amazon Rekognition, you can search images, stored videos, and streaming videos for faces that match those stored in a container known as a face collection. A face collection is an index of faces that you own and manage. Searching for people based on their faces requires two major steps in Amazon Rekognition:
  1. Index the faces.
  2. Search the faces.

- **Unsafe content detection** – Amazon Rekognition can detect adult and violent content in images and in stored videos. Developers can use the returned metadata to filter inappropriate content based on their business needs. Beyond flagging an image based on the presence of unsafe content, the API also returns a hierarchical list of labels with confidence scores. These labels indicate specific categories of unsafe content, which enables granular filtering and management of large volumes of user-generated content (UGC). Examples include social and dating sites, photo sharing platforms, blogs and forums, apps for children, ecommerce sites, entertainment, and online advertising services.

- **Celebrity recognition** – Amazon Rekognition can recognize celebrities within supplied images and in videos. Amazon Rekognition can recognize thousands of celebrities across a number of categories, such as politics, sports, business, entertainment, and media.

- **Text detection** – Amazon Rekognition Text in Image enables you to recognize and extract textual content from images. Text in Image supports most fonts, including highly stylized ones. It detects text and numbers in different orientations, such as those commonly found in banners and posters. In image sharing and social media applications, you can use it to enable visual search based on an index of images that contain the same keywords. In media and entertainment applications, you can catalog videos based on relevant text on screen, such as ads, news, sport scores, and captions. Finally, in public safety applications, you can identify vehicles based on license plate numbers from images taken by street cameras.

- **Custom labels**– With Amazon Rekognition Custom Labels, you can identify the objects and scenes in images that are specific to your business needs. For example, you can find your logo in social media posts, identify your products on store shelves, classify machine parts in an assembly line, distinguish healthy and infected plants, or detect animated characters in videos. For more information, see What is Amazon Rekognition Custom Labels? in the *Amazon Rekognition Custom Labels Developer Guide*.

Some of the benefits of using Amazon Rekognition include:

- **Integrating powerful image and video analysis into your apps** – You don't need computer vision or deep learning expertise to take advantage of the reliable image and video analysis in Amazon Rekognition. With the API, you can build image and video analysis into any web, mobile, or connected device application.

- **Deep learning-based image and video analysis** – Amazon Rekognition uses deep-learning technology to accurately analyze images, find and compare faces in images, and detect objects and scenes within your images and videos. You can analyze images for the presence of many different labels and then filter the results to include and/or exclude sets of labels or label categories.

- **Scalable image analysis** – Amazon Rekognition enables you to analyze millions of images so you can curate and organize massive amounts of visual data.

- **Analyze and filter images based on image properties** – Amazon Rekognition lets you analyze image properties like quality or colors. You can determine the sharpness, brightness, and contrast of images. You can also detect dominant colors in the entire image, foreground, background, and objects with bounding boxes.

- **Integration with other AWS services** – Amazon Rekognition is designed to work seamlessly with other AWS services like Amazon S3 and AWS Lambda. You can call the Amazon Rekognition API directly from Lambda in response to Amazon S3 events. Because Amazon S3 and Lambda scale automatically in response to your application's demand, you can build scalable, affordable, and reliable image analysis applications. For example, each time a person arrives at your residence, your door camera can upload a photo of the visitor to Amazon S3. This triggers a Lambda function that uses Amazon Rekognition API operations to identify your guest. You can run analysis directly on images that are stored in Amazon S3 without having to load or move the data. Support for AWS Identity and Access Management (IAM) makes it easy to securely control access to Amazon Rekognition API operations. Using IAM, you can create and manage AWS users and groups to grant the appropriate access to your developers and end users.

- **Low cost** – With Amazon Rekognition, you pay for the images and videos that you analyze, and the face metadata that you store. There are no minimum fees or upfront commitments. You can get started for free, and save more as you grow with the Amazon Rekognition tiered pricing model.

Amazon Rekognition and HIPAA eligibility

This is a HIPAA Eligible Service. For more information about AWS, U.S. Health Insurance Portability and Accountability Act of 1996 (HIPAA), and using AWS services to process, store, and transmit protected health information (PHI), see HIPAA Overview.

Amazon Rekognition (AMS SSPS)

Amazon Rekognition makes it easy to add image and video analysis to your applications using proven, highly scalable, deep learning technology that requires no machine learning expertise to use. With Amazon Rekognition, you can identify objects, people, text, scenes, and activities in images and videos, as well as detect any inappropriate content. Amazon Rekognition also provides highly accurate facial analysis and facial search capabilities that you can use to detect, analyze, and compare faces for a wide variety of user verification, people counting, and public safety use cases.

With Amazon Rekognition Custom Labels, you can identify objects and scenes in images that are specific to your business needs. For example, you can build a model to classify specific machine parts on your assembly line or to detect unhealthy plants. Amazon Rekognition Custom Labels takes care of the model development heavy lifting for you, so no machine learning experience is required. You simply need to supply images of objects or scenes you want to identify, and the service handles the rest.

Amazon Rekognition in AWS Managed Services FAQs

Common questions and answers:

Q: How do I request access to Amazon Rekognition in my AMS account?

Request access by submitting a Management | AWS service | Self-provisioned service | Add (review required) (ct-3qe6io8t6jtny) change type. This RFC provisions the following IAM role to your account: customer_rekognition_console_role & customer_rekognition_service_role. Once provisioned in your account, you must onboard the role in your federation solution.

Q: What are the restrictions to using Amazon Rekognition in my AMS account?

Full functionality of Amazon Rekognition is available with the Amazon Rekognition self-provisioned service role.

Q: What are the prerequisites or dependencies to using Amazon Rekognition in my AMS account?

If you use Kinesis Video Streams that provide the source streaming video for an Amazon Rekognition Video stream processor or a data stream as a destination to write data to Kinesis Data Streams, kindly provide AMS with a kinesisStreamName when creating the RFC.

Amazon Rekognition FAQs

General

Q: What is Amazon Rekognition?

Amazon Rekognition is a service that makes it easy to add powerful visual analysis to your applications. Rekognition Image lets you easily build powerful applications to search, verify, and organize millions of images. Rekognition Video lets you extract motion-based context from stored or live stream videos and helps you analyze them.

Rekognition Image is an image recognition service that detects objects, scenes, activities, landmarks, faces, dominant colors, and image quality. Rekognition Image also extracts text, recognizes celebrities, and identifies inappropriate content in images. It also allows you to search and compare faces.

Rekognition Video is a video recognition service that detects activities; understands the movement of people in frame; and recognizes objects, celebrities, and inappropriate content in videos stored in Amazon S3 and live video streams from Acuity. Rekognition Video detects persons and tracks them through the video even when their faces are not visible, or as the whole person might go in and out of the scene. For example, this could be used in an application that sends a real-time notification when someone delivers a package to your door. Rekognition Video allows you also to index metadata like objects, activities, scene, celebrities, and faces that make video search easy.

Q: What is deep learning?

Deep learning is a sub-field of Machine Learning and a significant branch of Artificial Intelligence. It aims to infer high-level abstractions from raw data by using a deep graph with

multiple processing layers composed of multiple linear and non-linear transformations. Deep learning is loosely based on models of information processing and communication in the brain. Deep learning replaces handcrafted features with ones learned from very large amounts of annotated data. Learning occurs by iteratively estimating hundreds of thousands of parameters in the deep graph with efficient algorithms.

Several deep learning architectures such as convolutional deep neural networks (CNNs), and recurrent neural networks have been applied to computer vision, speech recognition, natural language processing, and audio recognition to produce state-of-the-art results on various tasks.

Amazon Rekognition is a part of the Amazon AI family of services. Amazon AI services use deep learning to understand images, turn text into lifelike speech, and build intuitive conversational text and speech interfaces.

Q: Do I need any deep learning expertise to use Amazon Rekognition?

No. With Amazon Rekognition, you don't have to build, maintain or upgrade deep learning pipelines.

To achieve accurate results on complex computer vision tasks such as object and scene detection, face analysis, and face recognition, deep learning systems need to be tuned properly and trained with massive amounts of labeled ground truth data. Sourcing, cleaning, and labeling data accurately is a time-consuming and expensive task. Moreover, training a deep neural network is computationally expensive and often requires custom hardware built using Graphics Processing Units (GPU).

Amazon Rekognition is fully managed and comes pre-trained for image and video recognition tasks, so that you don't have invest your time and resources on creating a deep learning pipeline. Amazon Rekognition continues to improve the accuracy of its models by building upon the latest research and sourcing new training data. This allows you to focus on high-value application design and development.

Q: What are the most common use cases for Amazon Rekognition?

The most common use-cases for Rekognition Image include:

- Searchable Image Library
- Face-Based User Verification
- Sentiment Analysis
- Facial Recognition
- Image Moderation

The most common use-cases for Rekognition Video include:

- Search Index for video archives
- Easy filtering of video for explicit and suggestive content

Q: How do I get started with Amazon Rekognition?

If you are not already signed up for Amazon Rekognition, you can click the "Try Amazon Rekognition" button on the Amazon Rekognition page and complete the sign-up process. You must have an Amazon Web Services account; if you do not already have one, you will be prompted to create one during the sign-up process. Once you are signed up, try out Amazon Rekognition with your own images and videos using the Amazon Rekognition Management Console or download the Amazon Rekognition SDKs to start creating your own applications. Please refer to our step-by-step Getting Started Guide for more information.

Q: What image and video formats does Amazon Rekognition support?

Amazon Rekognition Image currently supports the JPEG and PNG image formats. You can submit images either as an S3 object or as a byte array. Amazon Rekognition Video operations can analyze videos stored in Amazon S3 buckets. The video must be encoded using the H.264 codec. The supported file formats are MPEG-4 and MOV. A codec is software or hardware that compresses data for faster delivery and decompresses received data into its original form. The H.264 codec is commonly used for the recording, compression and distribution of video content. A video file format may contain one or more codecs. If your MOV or MPEG-4 format video file does not work with Rekognition Video, check that the codec used to encode the video is H.264.

Q: What file sizes can I use with Amazon Rekognition?

Amazon Rekognition Image supports image file sizes up to 15MB when passed as an S3 object, and up to 5MB when submitted as an image byte array. Amazon Rekognition Video supports up to 10 GB files and up to 6 hour videos when passed through as an S3 file.

Q: How does image resolution affect the quality of Rekognition Image API results ?

Amazon Rekognition works across a wide range of image resolutions. For best results we recommend using VGA (640x480) resolution or higher. Going below QVGA (320x240) may increase the chances of missing faces, objects, or inappropriate content; although Amazon Rekognition accepts images that are at least 80 pixels in both dimensions.

Q: How small can an object be for Amazon Rekognition Image to detect and analyze it?

As a rule of thumb, please ensure that the smallest object or face present in the image is at least 5% of the size (in pixels) of the shorter image dimension. For example, if you are working with a 1600x900 image, the smallest face or object should be at least 45 pixels in either dimension.

Q: How can I get Amazon Rekognition predictions reviewed by humans?

Amazon Rekognition is directly integrated with Amazon Augmented AI (Amazon A2I) so you can easily route low confidence predictions from Amazon Rekognition Image to human reviewers. Using the Amazon Rekognition API for content moderation or the Amazon A2I console, you can specify the conditions under which Amazon A2I routes predictions to reviewers, which can be either a confidence threshold or a random sampling percentage. If you specify a confidence threshold, Amazon A2I routes only those predictions that fall below the threshold for human review. You can adjust these thresholds at any time to achieve the right balance between accuracy and cost-effectiveness. Alternatively, if you specify a sampling percentage, Amazon A2I routes a random sample of the predictions for human review. This can help you implement audits to monitor the prediction accuracy regularly. Amazon A2I also provides reviewers with a

web interface consisting of all the instructions and tools they need to complete their review tasks. For more information about implementing human review with Amazon Rekognition, see the Amazon A2I webpage.

Q: How does video resolution affect the quality of Rekognition Video API results?

The system is trained to recognize faces larger than 32 pixels (on the shortest dimension), which translate into a minimum size for a face to be recognized that varies from approximately 1/7 of the screen smaller dimension at QVGA resolution to 1/30 at HD 1080p resolution. For example, at VGA resolution, users should expect lower performances for faces smaller than 1/10 of the screen smaller dimension.

Q: What else can affect the quality of the Rekognition Video APIs ?

Besides video resolution, heavy blur, fast moving persons, lighting conditions, pose may affect the quality of the APIs.

Q: What is the preferred user video content that is suitable for Rekognition Video APIs?

This API works best with consumer and professional videos taken with frontal field of view in normal color and lighting conditions. This API is not tested for black and white, IR or extreme lighting condition. Applications that are sensitive to false alarms are advised to discard outputs with confidence score below a selected (application-specific) confidence score.

Q: In which AWS regions is Amazon Rekognition available?

For a list of all regions where Amazon Rekognition is available, see the AWS Region table.

Label Detection

Q: What is a label?

A label is an object, scene, or concept found in an image based on its contents. For example, a photo of people on a tropical beach may contain labels such as ‘Person’, ‘Water’, ‘Sand’, ‘Palm Tree’, and ‘Swimwear’ (objects), ‘Beach’ (scene), and ‘Outdoors’ (concept).

Q: What is a confidence score and how do I use it?

A confidence score is a number between 0 and 100 that indicates the probability that a given prediction is correct. In the tropical beach example, if the object and scene detection process returns a confidence score of 99 for the label ‘Water’ and 35 for the label ‘Palm Tree’, then it is more likely that the image contains water but not a palm tree.

Applications that are very sensitive to detection errors (false positives) should discard results associated with confidence scores below a certain threshold. The optimum threshold depends on the application. In many cases, you will get the best user experience by setting minimum confidence values higher than the default value.

Q: What is Object and Scene Detection?

Object and Scene Detection refers to the process of analyzing an image or video to assign labels based on its visual content. Amazon Rekognition Image does this through

the DetectLabels API. This API lets you automatically identify thousands of objects, scenes, and concepts and returns a confidence score for each label. DetectLabels uses a default confidence threshold of 50. Object and Scene detection is ideal for customers who want to search and organize large image libraries, including consumer and lifestyle applications that depend on user-generated content and ad tech companies looking to improve their targeting algorithms.

Q: Can Amazon Rekognition detect object locations and return bounding boxes?

Yes, Amazon Rekognition can detect the location of many common objects such as ‘Person’, ‘Car’, ‘Gun’, or ‘Dog’ in both images and videos. You get the coordinates of the bounding rectangle for each instance of the object found, as well as a confidence score. For more details on the API response structure for object bounding boxes, please refer to the documentation.

Q: Does Amazon Rekognition provide information on the relationship between detected labels?

Yes, for every label found, Amazon Rekognition returns its parent, alias, and category if they exist. Parents are returned in the "parents" field in hierarchical order. The first parent label is the immediate parent, while the following labels are parents of parents. For example, when a 'Car' is identified, Amazon Rekognition returns two parent labels 'Vehicle' (parent), and 'Transportation' (parent's parent). Aliases are labels with the same meaning as the primary labels and returned in the "aliases" field. For example, since 'Cell Phone' is an alias of 'Mobile Phone’, Amazon Rekognition returns 'Cell Phone' in the "aliases" field of a 'Mobile Phone' label. Categories group labels based on common themes and returned in the "categories" field. For example, since ’Dog’ is a label under the 'Animals and Pets' category, Amazon Rekognition returns 'Animal and Pets' in the "categories" field of a 'Dog' label. For more details on the full list of supported labels and their taxonomy, please visit Amazon Rekognition Label Detection documentation.

Q: What types of labels does Amazon Rekognition support?

Rekognition supports thousands of labels belonging to common categories including, but not limited to:

- People and Events: ‘Wedding’, ‘Bride’, ‘Baby’, ‘Birthday Cake’, ‘Guitarist’, etc.
- Food and Drink: ‘Apple’, ‘Sandwich’, ‘Wine’, ‘Cake’, ‘Pizza’, etc.
- Nature and Outdoors: ‘Beach’, ‘Mountains’, ‘Lake’, ‘Sunset’, ‘Rainbow’, etc.
- Animals and Pets: ‘Dog’, ‘Cat’, ‘Horse’, ‘Tiger’, ‘Turtle’, etc.
- Home and Garden: ‘Bed’, ‘Table’, ‘Backyard’, ‘Chandelier’, ‘Bedroom’, etc.
- Sports and Leisure: ‘Golf’, ‘Basketball’, ‘Hockey’, ‘Tennis’, ‘Hiking’, etc.
- Plants and Flowers: ‘Rose’, ‘Tulip’, ‘Palm Tree’, ‘Forest’, ‘Bamboo’, etc.
- Art and Entertainment: ‘Sculpture’, ‘Painting’, ‘Guitar’, ‘Ballet’, ‘Mosaic’, etc.
- Transportation and Vehicles: ‘Airplane’, ‘Car’, ‘Bicycle’, ‘Motorcycle’, ‘Truck’, etc.
- Electronics: ‘Computer’, ‘Mobile Phone’, ‘Video Camera’, ‘TV’, ‘Headphones’, etc.
- Landmarks: 'Brooklyn Bridge', 'Colosseum', 'Eiffel Tower', 'Machu Picchu', 'Taj Mahal', etc.

Q: How is Object and Scene Detection different for video analysis?

Rekognition Video enables you to automatically identify thousands of objects - such as vehicles or pets - and activities - such as celebrating or dancing - and provides you with timestamps and a confidence score for each label. It also relies on motion and time context in the video to accurately identify complex activities, such as “blowing a candle” or “extinguishing fire”.

Q: I can’t find the label I need. How do I request a new label?

Please send us your label requests through the Amazon Rekognition Console by typing the label name in the input field of the 'Search all labels' section and click 'Request Rekognition to detect' the requested label. Amazon Rekognition continuously expands its catalog of labels based on customer feedback.

Q: What is Image Properties?

Image Properties is a feature of Amazon Rekognition Image to detect dominant colors and image quality. Image Properties detects dominant colors of the entire image, image foreground, image background, and objects with localized bounding boxes. Image Properties also measure image quality through brightness, sharpness, and contrast scores. Image Properties can be called through the DetectLabels API using IMAGE_PROPERTIES as input parameter, with or without GENERAL_LABEL input parameter for label detection. Visit the Amazon Rekognition Label Detection documentation to learn more.

Q: How are dominant colors determined?

Image Properties returns dominant colors in four formats: RGB, hexcode, CSS color, and simplified colors. Amazon Rekognition first identifies the dominant colors of by pixel percentage, and then maps these colors to the 140 CSS color palette, RGB, hex code, and 12 simplified colors (i.e., 'green', 'pink', 'black', 'red', 'yellow', 'cyan', 'brown', 'orange', 'white', 'purple', 'blue', 'grey'). By default, Image Properties returns ten (10) dominant colors unless customers specify the number of colors to return. The maximum number of dominant colors the API can return is 12.

Q: How do I interpret the brightness, sharpness, and contrast scores?

Image Properties provides a value ranging from 0 to 100 for each brightness, sharpness, and contrast score. For example, an underexposed image will return a low brightness score, while a brightly lit image will return a high brightness score.

Q: How can you check if Amazon Rekognition has updated its models?

Amazon Rekognition returns a LabelModelVersion parameter that lets you know whether the model has been updated. Object and Scene detection models are updated frequently based on customer feedback.

Amazon Rekognition Custom Labels

Q: Can I use Custom Labels for analyzing faces, customized text detection?

No. Custom Labels is meant for finding objects and scenes in images. Custom Labels is not designed for analyzing faces, customized text detection. You should use other Rekognition APIs for these tasks. Please refer to the documentation for face analysis, Text detection.

Q: Can I use Custom Labels for finding unsafe image content?

Yes. Custom Labels is meant for finding objects and scenes in images. Custom Labels, when trained to detect your use case specific unsafe image content, can detect unsafe image content specific to your use case. Please also refer to the documentation for Moderation API to detect generic unsafe image content.

Q: How many images are needed to train a custom model?

The number of images required to train a custom model depends on the variability of the custom labels you want the model to predict and the quality of the training data. For example, a distinct logo overlaid on an image can be detected with 1-2 training images, while a more subtle logo required to be detected under many variations (scale, viewpoint, deformations) may need in the order of tens to hundreds of training examples with high quality annotations. If you already have a high number of labeled images, we recommend training a model with as many images as you have available. Please refer to the documentation for limits on maximum training dataset size.

Although hundreds of images may sometimes be required to train a custom model with high accuracy, with Custom Labels you can first train a model with tens of images per label, review your test results to understand where it does not work, and accordingly add new training images and train again to iteratively improve your model.

Q: How many inference compute resources should I provision for my custom model?

The number of parallel inference compute resources needed depends on how many images you need to process at a given point in time. The throughput of a single resource will depend factors including the size of the images, the complexity of those images (how many detected objects are visible), and the complexity of your custom model. We recommend that you monitor the frequency at which you need provision your custom model, and the number of images that need to be processed at a single time, in order to schedule provisioning of your custom model most efficiently.
If you expect to process images periodically (e.g. once a day or week, or scheduled times during the day), you should Start provisioning your custom model at a scheduled time, process all your images, and then Stop provisioning. If you don't stop provisioning, you will be charged even if no images are processed.

Q: My training has failed. Will I be charged?

No. You will not be charged for the compute resources if your training fails.

Content Moderation

Q: What is Content Moderation?

Amazon Rekognition's Content Moderation API uses deep learning to detect explicit or suggestive adult content, violent content, weapons, visually disturbing content, drugs, alcohol, tobacco, hate symbols, gambling, and rude gestures in image and videos. Beyond flagging an image or video based on presence of inappropriate or offensive content, Amazon Rekognition also returns a hierarchical list of labels with confidence scores. These labels indicate specific sub-categories of the type of content detected, thus providing more granular control to developers to filter and manage large volumes of user generated content (UGC). This API can be used in moderation workflows for applications such as social and dating sites, photo sharing platforms, blogs and forums, apps for children, e-commerce site, entertainment and online advertising services.

Q: What types of inappropriate, offensive, and unwanted content does Amazon Rekognition detect?

You can find a full list of content categories detected by Amazon Rekognition here.

Amazon Rekognition returns a hierarchy of labels, as well as a confidence score for each detected label. For instance, given an inappropriate image, Rekognition may return “Explicit Nudity” with a confidence score as a top-level label. Developers can use this metadata to flag content at a high level, for example, when all types of explicit adult content is to be flagged. In the same response, Rekognition also returns second level of granularity by providing additional context like “Graphic Male Nudity” with its own confidence score. Developers can use this information to build more complex filtering logic to serve different geographies and demographics.

Please note that the Content Moderation API is not an authority on, or in any way purports to be an exhaustive filter of inappropriate and offensive content. Furthermore, this API does not detect whether an image includes illegal content (such as child pornography) or unnatural adult content.

If you require other types of inappropriate content to be detected in images, please reach out to us using the feedback process outlined later in this section.

Q: How can I know which model version I am currently using?

Amazon Rekognition makes regular improvement to its models. To keep track of model version, you can use the 'ModerationModelVersion' field in the API response.

Q: How can I ensure that Amazon Rekognition meets accuracy goals for my image or video moderation use case?

Amazon Rekognition’s Content Moderation models have been and tuned and tested extensively, but we recommend that you measure the accuracy on your own data sets to gauge performance.

You can use the ‘MinConfidence’ parameter in your API requests to balance detection of content (recall) vs the accuracy of detection (precision). If you reduce ‘MinConfidence’, you are likely to detect most of the inappropriate content, but are also likely to pick up content that is not actually inappropriate. If you increase ‘MinConfidence’ you are likely to ensure that all your detected content is truly inappropriate but some content may not be tagged.

Q: How can I give feedback to Rekognition to improve its Content Moderation APIs?

Please send us your requests through AWS Customer Support. Amazon Rekognition continuously expands the types of inappropriate content detected based on customer feedback. Please note that illegal content (such as child pornography) will not be accepted through this process.

Facial Analysis

Q: What is Facial Analysis?

Facial analysis is the process of detecting a face within an image and extracting relevant face attributes from it. Amazon Rekognition Image takes returns the bounding box for each face detected in an image along with attributes such as gender, presence of sunglasses, and face landmark points. Rekognition Video will return the faces detected in a video with timestamps and, for each detected face, the position and a bounding box along with face landmark points.

Q: What face attributes can I get from Amazon Rekognition?

Amazon Rekognition returns the following facial attributes for each face detected, along with a bounding box and confidence score for each attribute:

- Gender
- Smile
- Emotions
- Eyeglasses
- Sunglasses
- Eyes open
- Mouth open
- Mustache
- Beard
- Pose
- Quality
- Face landmarks

Q: What is face pose?

Face pose refers to the rotation of a detected face on the pitch, roll, and yaw axes. Each of these parameters is returned as an angle between -180 and +180 degrees. Face pose can be used to find the orientation of the face bounding polygon (as opposed to a rectangular bounding box), to measure deformation, to track faces accurately, and more.

Q: What is face quality?

Face quality describes the quality of the detected face image using two parameters: sharpness and brightness. Both parameters are returned as values between 0 and 1. You can apply a threshold to these parameters to filter well-lit and sharp faces. This is useful for applications that benefit from high-quality face images, such as face comparison and face recognition.

Q: What are face landmarks?

Face landmarks are a set of salient points, usually located on the corners, tips or mid points of key facial components such as the eyes, nose, and mouth. Amazon Rekognition DetectFaces API returns a set of face landmarks that can be used to crop faces, morph one face into another, overlay custom masks to create custom filters, and more.

Q: How many faces can I detect in an image?

You can detect up to 100 faces in an image using Amazon Rekognition.

Q: How is Facial Analysis different for video analysis?

With Rekognition Video, you can locate faces across a video and analyze face attributes, such as whether the face is smiling, eyes are open, or showing emotions. Rekognition Video will return the detected faces with timestamps and, for each detected face, the position and a bounding box along with landmark points such as left eye, right eye, nose, left corner of the mouth, and right corner of the mouth. This position and time information can be used to easily track user sentiment over time and deliver additional functionality such as automatic face frames, highlights, or crops.

Q: In addition to Video resolution, what else can affect the quality of the Rekognition Video APIs?

Besides video resolution, the quality and representative faces, part of the face collections to search, has major impact. Using multiple face instances per person with variations like beard, glasses, poses (profile and frontal) will significantly improve the performance. Typically very fast moving people and blurred videos may experience lower quality.

Face Comparison

Q: What is Face Comparison?

Face Comparison is the process of comparing one face to one or more faces to measure similarity. Using the CompareFaces API, Amazon Rekognition Image lets you measure the likelihood that faces in two images are of the same person. The API compares a face in the source input image with each face detected in the target input image and returns a similarity score for each comparison. You also get a bounding box and confidence score for each face detected. You can use face comparison to verify a person's identity against their personnel photo on file in near real-time.

Q: Can I use a source image with more than one face?

Yes. If the source image contains multiple faces, CompareFaces detects the largest face and uses it to compare with each face detected in the target image.

Q: How many faces can I compare against?

You can compare one face in the source image with up to 15 detected faces in the target image.

Facial Recognition

Q: What is Facial Recognition?

Facial recognition is the process of identifying or verifying a person's identity by searching for their face in a collection of faces. Using facial recognition, you can easily build applications such as multi-factor authentication for bank payments, automated building entry for employees, and more.

Q: What is a face collection and how do I create one?

A face collection is a searchable index of face feature vectors, owned and managed by you. Using the CreateCollection API, you can easily create a collection in a supported AWS region

and get back an Amazon Resource Name (ARN). Each face collection has a unique CollectionId associated with it.

Q: How do I add faces to or delete faces from a face collection?

To add a face to an existing face collection, use the IndexFaces API. This API accepts an image in the form of an S3 object or image byte array and adds a vector representation of the faces detected to the face collection. IndexFaces also returns a unique FaceId and face bounding box for each of the faces added.

To delete a face from an existing face collection, use the DeleteFaces API. This API operates on the face collection supplied (using a CollectionId) and removes the entries corresponding to the list of FaceIds. For more information on adding and deleting faces, please refer to our Managing Collections example.
Q: How do I search for a face within a face collection?

Once you have created an indexed collection of faces, you can search for a face within it using either an image (SearchFaceByImage) or a FaceId (SearchFaces). These APIs take in an input face and return a set of faces that match, ordered by similarity score with the highest similarity first. For more details, please refer to our Searching Faces example.

Q: How is Facial Recognition different for video analysis?

Rekognition Video allows you to perform real time face searches against collections with tens of millions of faces. First, you create a face collection, where you can store faces, which are vector representations of facial features. Rekognition then searches the face collection for visually similar faces throughout your video. Rekognition will return a confidence score for each of the faces in your video, so you can display likely matches in your application.

Q: In addition to Video resolution what else can affect the quality of the Video APIs ?

Besides video resolution, the quality and representative faces part of the face collections to search has major impact. Using multiple face instances per person with variations like beard, glasses, poses (profile and frontal) will significantly improve the performance. Typically very fast moving people may experience low recall. In addition, blurred videos may also experience lower quality.

Celebrity Recognition

Q: What is Celebrity Recognition?

Amazon Rekognition's Celebrity Recognition is a deep learning based easy-to-use API for detection and recognition of individuals who are famous, noteworthy, or prominent in their field. The RecognizeCelebrities API has been built to operate at scale and recognize celebrities across a number of categories, such as politics, sports, business, entertainment, and media. Our Celebrity Recognition feature is ideal for customers who need to index and search their digital image libraries for celebrities based on their particular interest.

Q: Who can be identified by the Celebrity Recognition API?

Amazon Rekognition can only identify celebrities that the deep learning models have been trained to recognize. Please note that the RecognizeCelebrities API is not an authority on, and in

no way purports to be, an exhaustive list of celebrities. The feature has been designed to include as many celebrities as possible, based on the needs and feedback of our customers. We are constantly adding new names, but the fact that Celebrity Recognition does not recognize individuals that may be deemed prominent by any other groups or by our customers is not a reflection of our opinion of their celebrity status. If you would like to see additional celebrities identified by Celebrity Recognition, please submit feedback.

Q: Can a celebrity identified through the Amazon Rekognition API request to be removed from the feature?

Yes. If a celebrity wishes to be removed from the feature, he or she can send an email to AWS Customer Support and we will process the removal request.

Q: What sources are supported to provide additional information about a Celebrity ?

The API supports an optional list of sources to provide additional information about the celebrity as a part of the API response. We currently provide the IMDB URL, when it is available. We may add other sources at a later date.

Q: How is Celebrity Recognition different for video analysis?

With Rekognition Video, you can detect and recognize when and where well known persons appear in a video. The time-coded output includes the name and unique id of the celebrity, bounding box coordinates, confidence score, and URLs pointing to related content for the celebrity, for example, the celebrity's IMDB link. The celebrity is also detected even if sometimes the face becomes occluded in the video. This feature allows you to index and search digital video libraries for use cases related to your specific marketing and media needs.

Q: In addition to Video resolution, what else can affect the quality of the Rekognition Video APIs?

Very fast moving celebrities and blurred videos can affect the quality of the Rekognition Video APIs. In addition, heavy makeup and camouflage common for actors/actresses, can also affect the quality.

Text Detection

Q: What is Text Detection?

Text detection is a capability of Amazon Rekognition that allows you to detect and recognize text within an image or a video, such as street names, captions, product names, overlaid graphics, video subtitles, and vehicular license plates. Text detection is specifically built to work with real-world images and videos, rather than document images. Amazon Rekognition’s DetectText API takes in an image and returns the text label and a bounding box for each detected string of characters, along with a confidence score. For example, in image sharing and social media applications, you can enable visual search based on an index of images that contain the same text labels. In security applications, you can identify vehicles based on license plate numbers from images taken by traffic cams. Similarly, for videos, using the StartTextDetection and GetTextDetection APIs, you can detect text and get confidence scores and timestamps for each detection. In media and entertainment applications, you can create text metadata to support search for relevant content, such as news, sport scores, commercials,

and captions. You can also review the detected text for policy or compliance violations e.g. an email address or phone number that has been overlaid by spammers.

Q: What type of text does Amazon Rekognition Text Detection support?

Text Detection is specifically built to work with real-world images and videos rather than document images. It supports text in most Latin scripts and numbers embedded in a large variety of layouts, fonts and styles, and overlaid on background objects at various orientation as banners and posters. Text detection recognizes up to 50 sequences of characters per the image or video frame and lists them as words and lines. Text detection supports text rotated by up to -90 to +90 degrees from the horizontal axis.

Q: Can I limit text detection to specific regions in an image or video frame?

Yes, you can use text detection filtering options to specify up to 10 Regions of Interest (ROIs) in the API request. Amazon Rekognition will only return text that falls within these regions.

Q: Can I filter text detections by word confidence or bounding box size?

Yes, in the API request you can use the text detection filtering options to specify thresholds for minimum confidence scores or minimum bounding box dimensions.

Q: How can I give feedback to Rekognition to improve its text recognition?

Please send us your requests through AWS Customer Support. Amazon Rekognition continuously expands the types of text content recognized based on customer feedback.

PPE Detection

Q: What personal protective equipment (PPE) can Amazon Rekognition detect?

Amazon Rekognition “DetectProtectiveEquipment” can detect common types of face covers, hand covers, and head covers. To learn more, please refer to feature documentation. You can also use Amazon Rekognition Custom Labels to detect PPE such as high-visibility vests, safety goggles, and other PPE unique to your business. To learn about how you can use Amazon Rekognition Custom Labels for custom PPE detection, visit this github repo.

Q: Can Amazon Rekognition detect protective equipment locations and return bounding boxes?

Yes, Amazon Rekognition “DetectProtectiveEquipment” API can detect the location of protective equipment such as face covers, hand covers, and head covers on persons in images. You get the coordinates of the bounding box rectangle for each item of protective equipment detected, as well as a confidence score. For more details on the API response, please refer to documentation.

Q: Can the service detect if the mask is worn properly?

Amazon Rekognition “DetectProtectiveEquipment” API output provides “CoversBodyPart” value (true/false) and confidence value for the Boolean value for each detected item of protective equipment. This provides information on whether the protective equipment is on the corresponding body part of the person. The prediction about the presence of the protective

equipment on the corresponding body part helps filter out cases where the PPE is in the image but not actually on the person. It does not, however, indicate or imply that the person is adequately protected by the protective equipment or that the protective equipment itself is properly worn.

Q: Can Amazon Rekognition PPE detection identify detected persons?

No, Amazon Rekognition PPE detection does not perform facial recognition or facial comparison and cannot identify the detected persons.

Q: Where can I find more information about API limits and latency?

Please refer to Amazon Rekognition PPE detection documentation to get the latest details on API limits and latency.

Q: How do I send images from my workplace cameras to Amazon Rekognition?

You have multiple options to sample images from your workplace cameras. Please refer to the Amazon Rekognition PPE detection blog to learn more.

Q: How is PPE detection priced?

Amazon Rekognition PPE detection is priced similarly to other Amazon Rekognition Image APIs on a per image basis.

Amazon Rekognition Streaming Video Events

Q: What are Amazon Rekognition Streaming Video Events?
Amazon Rekognition Streaming Video Events uses machine learning to detect objects from connected camera to provide actionable alerts in real time. Amazon Rekognition Streaming Video events work with your new and existing Kinesis Video Streams to process video streams (up to 120 seconds per motion event) and notify you as soon a desired object of interest in detected. You can use these notifications to

- Send Smart Alerts to your end users such as “a package was detected at the front door.”
- Provide home automation capabilities such as “turning on the garage light when a person is detected.”
- Integrate with smart assistants such as Echo devices to provide Alexa announcements when an object is detected.
- Provide Smart Search capabilities such as search for all video clips where a package was detected.

Q: How does Amazon Rekognition Streaming Video Events work?
You can use your new or existing Kinesis Video Streams to get started with Amazon Rekognition Streaming Video events. When configuring your stream processor settings for Amazon Rekognition you can choose the desired labels (person, pet, or package) that you want to detect, the duration for the video (up to 120 seconds per motion event) that Rekognition should process for each event, and/or the region of interest on the frame that you want to

process through Rekognition. Rekognition Streaming Video Event APIs process video only when you send a notification to Rekognition to start processing the video stream.

When motion is detected on a connected camera, you send a notification to Rekognition to start processing the video stream. Rekognition processes the corresponding Kinesis Video Stream, post motion detection, to look for the desired objects specified by you. As soon as a desired object is detected, Amazon Rekognition will send you a notification. This notification includes the object detected, bounding box, zoomed in image of the object, and the time stamp.

Q: What labels can Amazon Rekognition Streaming Video Events support?
Amazon Rekognition Streaming Video Events can support people, pets, and packages.

Q: What pets and package types can Amazon Rekognition Streaming Video APIs detect?
Amazon Rekognition Streaming Video Event APIs support dogs and cats for pet detection. The API can detect medium and large cardboard boxes with high accuracy. The API also detects smaller boxes, bubble mailer envelopes, and folders but may miss some of these objects occasionally.

Q: Will I be charged separately for each label detected? Can I choose which labels to opt into?
No, you will not be charged separately for each label. You will be charged for the duration of streaming video processed by Rekognition. You can either opt into specific labels (pet, package) or choose to opt in to all three labels (people, pet, package) while configuring your stream processing settings.

Q: Do I need to stream video continuously to Amazon Rekognition?
No, you do not have to stream video continuously to Amazon Rekognition.

Q: Should I create new Kinesis Video Streams (KVS) to use Streaming Video Events?
Amazon Rekognition Streaming Video Events works with both new and existing Kinesis Video Streams. Simply integrate the relevant KVS streams with Amazon Rekognition Streaming Video Events API to get started with video analysis on KVS streams.

Q: When will Amazon Rekognition send me a notification?
Amazon Rekognition starts processing the video stream post motion detection. You can configure the duration for processing this video stream (up to 120 seconds per event). As soon as Amazon Rekognition detects the object of interest in the video stream, Rekognition will send you a notification. This notification includes the type of object detected, the bounding box, a zoomed in image of the object detected, and a time stamp.

Q: What resolution and fps is support for label detection?
In order to keep costs and latency low Amazon Rekognition Streaming Video Events support 1080p or lower resolution video streams. Rekognition processes the video stream at 5 fps.

Q: What codecs and file format is supported for streaming video?
Amazon Rekognition Video supports H.264 files in MPEG-4 (.mp4) or MOV format.

Q: What is the maximum duration of the video processed per event?
You can process up to 120 seconds of video per event.

Q: Can I choose a particular area of the frame to be processed for my video stream?
Yes, as a part of configuring your StreamProcessor you can choose the region of interest that you want to process on your frame. Amazon Rekognition will only process that particular area of the frame.

Q: How many concurrent video streams can I process with Amazon Rekognition?
Amazon Rekognition Streaming Video Events can support 600 concurrent sessions per AWS customer. Please reach out to your account manager if you need to increase this limit.

Amazon Rekognition Stored Video Analysis

Q: What types of entities can Amazon Rekognition Video detect?
Amazon Rekognition Video can detect objects, scenes, faces, celebrities, text, and inappropriate content in videos. You can also search for faces appearing in a video using your own repository or collection of face images.

Q: What types file formats and codecs does Amazon Rekognition Video support?
Amazon Rekognition Video supports H.264 files in MPEG-4 (.mp4) or MOV format. If your video files use a different codec, you can transcode them into H.264 using AWS Elemental MediaConvert.

Q: How do Amazon Rekognition Video asynchronous APIs work?
Amazon Rekognition Video can process videos stored in an Amazon S3 bucket. You can use an asynchronous set of operations: you start video analysis by calling a Start operation such as StartLabelDetection for detecting objects and scenes. The completion status of the request is published to an Amazon Simple Notification Service (SNS) topic. To get the completion status from the Amazon SNS topic, you can use an Amazon Simple Queue Service (SQS) queue or an AWS Lambda function. Once you have the completion status, you call a Get operation such as GetLabelDetection to get the results of the request. For a list of available Amazon Rekognition Video APIs, please see this page.

Q: How can I find the timeline for each detection in a video?
Amazon Rekognition Video returns timestamps in milliseconds for each entity it detects on the video timeline, for example, using the label detection, a 'Dog' may be found at 166ms from the start of a video. Using this information, you can understand the timeline of an object or a celebrity shown in the video. To see how a sample API response with timestamps looks like, please refer to this page on detecting object and scenes in Amazon Rekognition Video.

Q: How many concurrent video analysis jobs can I run with Amazon Rekognition Video?
You can process up to 20 concurrent jobs with Amazon Rekognition Video. For more details on limits, please see our limits page.

Q: What video resolution should I use?
Amazon Rekognition Video automatically handles a wide range of video resolutions and video quality. We recommend using a 720p (1280×720 pixels) to 1080p (1920x1080 pixels), or their equivalent resolutions in other aspect ratios for optimum results. Very low resolution (such as QVGA or 240p) and very low-quality videos may adversely impact results quality.

Q: What is People Pathing?
With Rekognition Video, you can find the path of each person across the video timeline. Rekognition Video detects persons even when the camera is in motion and, for each person, returns a bounding box and the face, along with face attributes and timestamps. For retail applications, this allows to generate customer insights anonymously, such as how customers move across aisles in a shopping mall or how long they are waiting in checkout lines.

Media Analysis using Amazon Rekognition Video

Q: What types of media analysis segments can Amazon Rekognition Video detect?

Amazon Rekognition Video can detect the following types of segments or entities for media analysis:

- Black frames: Videos often contain a short duration of empty black frames with no audio that are used as cues to insert advertisements, or to demarcate the end of a program segment such as a scene or the opening credits. With Amazon Rekognition Video, you can detect such black frame sequences to automate ad insertion, package content for VOD, and demarcate various program segments or scenes. Black frames with audio (such as fade outs or voiceovers) are considered as content and not returned.
- Credits: Amazon Rekognition Video helps you automatically identify the exact frames where the opening and closing credits start and end for a movie or TV show. With this information, you can generate 'binge markers' or interactive viewer prompts such as 'Next Episode' or 'Skip Intro' in VOD applications. Amazon Rekognition Video is trained to handle a wide variety of opening and end credit styles ranging from simple rolling credits to more challenging credits alongside content, credits on scenes, or stylized credits in anime content.
- Shots: A shot is a series of interrelated consecutive pictures taken contiguously by a single camera and representing a continuous action in time and space. With Amazon Rekognition Video, you can detect the start, end, and duration of each shot, as well as a count all the shots in a piece of content. Shot metadata can be used for applications such as creating promotional videos using selected shots, generating a set of preview thumbnails that avoid transitional content between shots, and inserting ads in spots that don't disrupt viewer experience, such as the middle of a shot when someone is speaking.
- Color Bars: Amazon Rekognition Video allows you to detect sections of video that display SMPTE or EBU color bars, which are a set of colors displayed in specific patterns to ensure color is calibrated correctly on broadcast monitors, programs, and on cameras. For more information about SMPTE color bars, see SMPTE color bar. This metadata is useful to prepare content for VOD applications by removing color bar segments from the content, or to detect issues such as loss of broadcast signals in a recording, when color bars are shown continuously as a default signal instead of content.
- Slates: Slates are sections, typically at the beginning of a video, that contain text metadata about the episode, studio, video format, audio channels, and more. Amazon Rekognition can identify the start and end such slates, making it easy for operators to use the text metadata or to simply remove the slate when preparing content for final viewing.
- Studio logos: Studio logos are sequences that show the logos or emblems of the production studio involved in making the show. Amazon Rekognition can identify such sequences, making it easy for operators to review them for identifying studios.
- Content: Content refers to the portions of the TV show or movie that contain the program or related elements. Black frames, credits, color bars, slates, and studio logos are not considered to be content. Amazon Rekognition Video enables you to detect the start and end of each content segment in the video, which enables multiple uses such as finding the program run time or finding certain segments that serve specific purposes. For example, a quick recap of the previous episode at the beginning of the video is a type of content. Similarly, bonus post-credit content can appear after the credits have finished. And, some videos may have 'textless' content at the end of the video, which are a set of all program content that contains overlaid text, but with that text removed to enable internationalization in another language. Once all the content segments are detected with Amazon Rekognition Video, you can apply specific domain knowledge such as 'my videos always start with a recap' to further categorize each segment or to send them for human review.

Amazon Rekognition Video provides the start, end, duration, and timecodes for each detected entity, and provides timestamp (milliseconds), SMPTE format code, and frame number options for each.

Q: How do I get started with media analysis using Amazon Rekognition Video?

Media analysis features are available through the Amazon Rekognition Video segment detection API. This is an asynchronous API composed of two operations: *StartSegmentDetection* to start the analysis, and *GetSegmentDetection* to get the analysis results. To get started, please refer to the documentation.

If you want to visualize the results of media analysis or even try out other Amazon AI services like Amazon Transcribe with your own videos, please use the Media Insights application – a serverless framework and demo application to easily generate insights and develop applications for your video, audio, text, and image resources, using AWS Machine Learning and Media services. You can easily spin up your own demo application using the supplied AWS CloudFormation template, to try out your own videos and visualize analysis results.

Q: What is a frame accurate timecode?

Frame accurate timecodes provide the exact frame number for a relevant segment of video or entity. Media companies commonly process timecodes using the SMPTE (Society of Motion Picture and Television Engineers) format hours:minutes:seconds:frame number, for example, 00:24:53:22.

Q: Is Amazon Rekognition Video segment detection frame accurate?

Yes, the Amazon Rekognition Video segment detection API provides frame accurate SMPTE timecodes, as well as millisecond timestamps for the start and end of each detection.

Q: What types of frame rate formats can Amazon Rekognition Video segment detection handle?

Amazon Rekognition Video segment detection automatically handles integer, fractional and drop frame standards for frame rates between 15 and 60fps. For example, common frame rates such as 23.976 fps, 25fps, 29.97 fps and 30fps are supported by segment detection. Frame rate information is utilized to provide frame accurate timecodes in each case.

Q: What filtering options can I apply?

You can specify the minimum confidence for each segment type while making the API request. For example, you can filter out any segment below 70% confidence score. For black frame detection, you can also control the maximum pixel luminance that you consider to be a black pixel, for example, a value of 40 for a color range of 0 to 255. Further, you can also control what percentage of pixels in a frame need to meet this black pixel luminance criteria for the frame to be classified as a black frame, for example, 99%. These filters allow you to account for varied video quality and formats when detecting black frames. For example, videos reclaimed from tape archives might be noisy and have a different black level compared to a modern digital video. For more details, please refer to this page.

AWS Integration

Q: Does Amazon Rekognition Video work with images stored on Amazon S3?

Yes. You can start analyzing images stored in Amazon S3 by simply pointing the Amazon Rekognition API to your S3 bucket. You don't need to move your data. For more details of how to use S3 objects with Amazon Rekognition API calls, please see our Detect Labels exercise.

Q: Can I use Amazon Rekognition with images stored in an Amazon S3 bucket in another region?

No. Please ensure that the Amazon S3 bucket you want to use is in the same region as your Amazon Rekognition API endpoint.

Q: How do I process multiple image files in a batch using Amazon Rekognition?

You can process your Amazon S3 images in bulk using the steps described in our Amazon Rekognition Batch Processing example on GitHub.

Q: How can I use AWS Lambda with Amazon Rekognition?

Amazon Rekognition provides seamless access to AWS Lambda and allows you bring trigger-based image analysis to your AWS data stores such as Amazon S3 and Amazon DynamoDB. To use Amazon Rekognition with AWS Lambda, please follow the steps outlined here and select the Amazon Rekognition blueprint.

Q: Does Amazon Rekognition work with AWS CloudTrail?

Yes. Amazon Rekognition supports logging the following actions as events in CloudTrail log files: CreateCollection, DeleteCollection, CreateStreamProcessor, DeleteStreamProcessor, DescribeStreamProcessor, ListStreamProcessors, and ListCollections. For more details on the Amazon Rekognition API calls that are integrated with AWS CloudTrail, see Logging Amazon Rekonition API Calls with AWS CloudTrail.

Data Privacy

Q: Are image and video inputs processed by Amazon Rekognition stored, and how are they used by AWS?

Amazon Rekognition may store and use image and video inputs processed by the service solely to provide and maintain the service and, unless you opt out as provided below, to improve and develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies. Use of your content is important for continuous improvement of your Amazon Rekognition customer experience, including the development and training of related technologies. We do not use any personally identifiable information that may be contained in your content to target products, services or marketing to you or your end users. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information. You may opt out of having your image and video inputs used to improve or develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies by using an AWS Organizations opt-out policy. For information about how to opt out, see Managing AI services opt-out policy.

Q: Can I delete image and video inputs stored by Amazon Rekognition?

Yes. You can request deletion of image and video inputs associated with your account by contacting AWS Support. Deleting image and video inputs may degrade your Amazon Rekognition experience.

Q: Who has access to my content that is processed and stored by Amazon Rekognition?

Only authorized employees will have access to your content that is processed by Amazon Rekognition. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Do I still own my content that is processed and stored by Amazon Rekognition?

You always retain ownership of your content and we will only use your content with your consent.

Q: Is the content processed by Amazon Rekognition moved outside the AWS region where I am using Amazon Rekognition?

Any content processed by Amazon Rekognition is encrypted and stored at rest in the AWS region where you are using Amazon Rekognition. Unless you opt out as provided below, some portion of content processed by Amazon Rekognition may be stored in another AWS region solely in connection with the continuous improvement and development of your Amazon Rekognition customer experience and other Amazon machine-learning/artificial-intelligence technologies. You can request deletion of image and video inputs associated with your account by contacting AWS Support. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information. Your content will not be stored in another AWS region if you opt out of having your content used to improve and develop the quality of Amazon Rekognition and other Amazon machine-learning/artificial-intelligence technologies. For information about how to opt out, see Managing AI services opt-out policy.

Q: Can I use Amazon Rekognition in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Yes, subject to your compliance with the Amazon Rekognition Service Terms, including your obligation to provide any required notices and obtain any required verifiable parental consent under COPPA, you may use Amazon Rekognition in connection with websites, programs, or other applications that are directed or targeted, in whole or in part, to children under age 13.

Q: How do I determine whether my website, program, or application is subject to COPPA?

For information about the requirements of COPPA and guidance for determining whether your website, program, or other application is subject to COPPA, please refer directly to the resources provided and maintained by the United States Federal Trade Commission. This site

also contains information regarding how to determine whether a service is directed or targeted, in whole or in part, to children under age 13.

Q: Is Amazon Rekognition a HIPAA Eligible Service?

Amazon Rekognition is a HIPAA Eligible Service covered under the AWS Business Associate Addendum (AWS BAA). If you have an AWS BAA in place, Amazon Rekognition will use, disclose, and maintain your Protected Health Information (PHI) only as permitted by the terms of your AWS BAA.

Access Control

Q: How do I control user access for Amazon Rekognition?

Amazon Rekognition is integrated with AWS Identity and Access Management (IAM). AWS IAM policies can be used to ensure that only authorized users have access to Amazon Rekognition APIs. For more details, please see the Amazon Rekognition Authentication and Access Control page.

Report Abuse

Q: How can I report potential Amazon Rekognition abuse?

If you suspect that Amazon Rekognition is being used in manner that is abusive or illegal, or infringes on your rights or the rights of other people, please report this use and AWS will investigate the issue.

Amazon Lex

Build chatbots with conversational AI.

How it works

Amazon Lex is a fully managed artificial intelligence (AI) service with advanced natural language models to design, build, test, and deploy conversational interfaces in applications.

Use cases

Build virtual agents and voice assistants
Enable self-service capabilities with virtual contact center agents and interactive voice response (IVR). Users can change a password or schedule an appointment without speaking to a human agent.

Automate informational responses

Design conversational solutions that provide responses to frequently asked questions. Improve Connect & Lex conversation flows for tech support, HR benefits, or finance with natural language search for FAQs powered by Amazon Kendra.

Improve productivity with application bots
Automate basic user tasks in your application with powerful chatbots. Seamlessly connect with other enterprise software through AWS Lambda and maintain granular access control through IAM.

Maximize the information trapped in transcripts
Design chatbots using existing contact center transcripts. Reduce design time from weeks to hours and accelerate bot deployment.

What is Amazon Lex V2?

Amazon Lex V2 is an AWS service for building conversational interfaces for applications using voice and text. Amazon Lex V2 provides the deep functionality and flexibility of natural language understanding (NLU) and automatic speech recognition (ASR) so you can build highly engaging user experiences with lifelike, conversational interactions, and create new categories of products.

Amazon Lex V2 enables any developer to build conversational bots quickly. With Amazon Lex V2, no deep learning expertise is necessary—to create a bot, you specify the basic conversation flow in the Amazon Lex V2 console. Amazon Lex V2 manages the dialog and dynamically adjusts the responses in the conversation. Using the console, you can build, test, and publish your text or voice chatbot. You can then add the conversational interfaces to bots on mobile devices, web applications, and chat platforms (for example, Facebook Messenger).

Amazon Lex V2 provides integration with AWS Lambda, and you can integrate with many other services on the AWS platform, including Amazon Connect, Amazon Comprehend, and Amazon Kendra. Integration with Lambda provides bots access to pre-built serverless enterprise connectors to link to data in SaaS applications such as Salesforce.

For bots created after August 17, 2022, you can use conditional branching to control the conversation flow with your bot. With conditional branching you can create complex conversations without needing to write Lambda code.

Amazon Lex V2 provides the following benefits:

- **Simplicity** – Amazon Lex V2 guides you through using the console to create your own bot in minutes. You supply a few example phrases, and Amazon Lex V2 builds a complete natural language model through which the bot can

interact using voice and text to ask questions, get answers, and complete sophisticated tasks.

- **Democratized deep learning technologies** – Amazon Lex V2 provides ASR and NLU technologies to create a Speech Language Understanding (SLU) system. Through SLU, Amazon Lex V2 takes natural language speech and text input, understands the intent behind the input, and fulfils the user intent by invoking the appropriate business function.

  Speech recognition and natural language understanding are some of the most challenging problems to solve in computer science, requiring sophisticated deep learning algorithms to be trained on massive amounts of data and infrastructure. Amazon Lex V2 puts deep learning technologies within reach of all developers. Amazon Lex V2 bots convert incoming speech to text and understand the user intent to generate an intelligent response so you can focus on building your bots with added value for your customers and define entirely new categories of products made possible through conversational interfaces.

- **Seamless deployment and scaling** – With Amazon Lex V2, you can build, test, and deploy your bots directly from the Amazon Lex V2 console. Amazon Lex V2 enables you to publish your voice or text bots for use on mobile devices, web apps, and chat services (for example, Facebook Messenger). Amazon Lex V2 scales automatically. You don't need to worry about provisioning hardware and managing infrastructure to power your bot experience.

- **Built-in integration with the AWS platform** – Amazon Lex V2 operates natively with other AWS services, such as AWS Lambda and Amazon CloudWatch. You can take advantage of the power of the AWS platform for security, monitoring, user authentication, business logic, storage, and mobile app development.

- **Cost-effectiveness** – With Amazon Lex V2, there are no upfront costs or minimum fees. You are charged only for the text or speech requests that are made. The pay-as-you-go pricing and the low cost per request make the service a cost-effective way to build conversational interfaces. With the

Amazon Lex V2 free tier, you can easily try Amazon Lex V2 without any initial investment.

Amazon Lex: How It Works

It is recommended to use Amazon Lex V2 to take advantage of the latest features.

Amazon Lex enables you to build applications using a speech or text interface powered by the same technology that powers Amazon Alexa. Following are the typical steps you perform when working with Amazon Lex:

1. Create a bot and configure it with one or more intents that you want to support. Configure the bot so it understands the user's goal (intent), engages in conversation with the user to elicit information, and fulfills the user's intent.
2. Test the bot. You can use the test window client provided by the Amazon Lex console.
3. Publish a version and create an alias.
4. Deploy the bot. You can deploy the bot on platforms such as mobile applications or messaging platforms such as Facebook Messenger.

Before you get started, familiarize yourself with the following Amazon Lex core concepts and terminology:

- **Bot** – A bot performs automated tasks such as ordering a pizza, booking a hotel, ordering flowers, and so on. An Amazon Lex bot is powered by Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) capabilities. Each bot must have a unique name within your account.

  Amazon Lex bots can understand user input provided with text or speech and converse in natural language. You can create Lambda functions and add them as code hooks in your intent configuration to perform user data validation and fulfillment tasks.

- **Intent** – An intent represents an action that the user wants to perform. You create a bot to support one or more related intents. For example, you might create a bot that orders pizza and drinks. For each intent, you provide the following required information:

- **Intent name**– A descriptive name for the intent. For example, **OrderPizza**. Intent names must be unique within your account.
- **Sample utterances** – How a user might convey the intent. For example, a user might say "Can I order a pizza please" or "I want to order a pizza".
- **How to fulfill the intent** – How you want to fulfill the intent after the user provides the necessary information (for example, place order with a local pizza shop). We recommend that you create a Lambda function to fulfill the intent.

  You can optionally configure the intent so Amazon Lex simply returns the information back to the client application to do the necessary fulfillment.

In addition to custom intents such as ordering a pizza, Amazon Lex also provides built-in intents to quickly set up your bot. For more information, see Built-in Intents and Slot Types.

- **Slot** – An intent can require zero or more slots or parameters. You add slots as part of the intent configuration. At runtime, Amazon Lex prompts the user for specific slot values. The user must provide values for all *required* slots before Amazon Lex can fulfill the intent.

  For example, the OrderPizza intent requires slots such as pizza size, crust type, and number of pizzas. In the intent configuration, you add these slots. For each slot, you provide slot type and a prompt for Amazon Lex to send to the client to elicit data from the user. A user can reply with a slot value that includes additional words, such as "large pizza please" or "let's stick with small." Amazon Lex can still understand the intended slot value.

- **Slot type** – Each slot has a type. You can create your custom slot types or use built-in slot types. Each slot type must have a unique name within your account. For example, you might create and use the following slot types for the OrderPizza intent:

- Size – With enumeration values Small, Medium, and Large.
- Crust – With enumeration values Thick and Thin.

Amazon Lex also provides built-in slot types. For example, AMAZON.NUMBER is a built-in slot type that you can use for the number of pizzas ordered. For more information, see Built-in Intents and Slot Types.

For a list of AWS Regions where Amazon Lex is available, see AWS Regions and Endpoints in the *Amazon Web Services General Reference*.

Amazon Lex FAQs

General

Q: What is Amazon Lex?

Amazon Lex is a service for building conversational interfaces using voice and text. Powered by the same conversational engine as Alexa, Amazon Lex provides high quality speech recognition and language understanding capabilities, enabling addition of sophisticated, natural language 'chatbots' to new and existing applications. Amazon Lex reduces multi-platform development effort, allowing you to easily publish your speech or text chatbots to mobile devices and multiple chat services, like Facebook Messenger, Slack, Kik, or Twilio SMS. Native interoperability with AWS Lambda and Amazon CloudWatch and easy integration with many other services on the AWS platform including Amazon Cognito, and Amazon DynamoDB makes bot development effortless.

Q: How can I get started with Amazon Lex?

To start using Amazon Lex, simply sign into the AWS Management Console and navigate to "Lex" under the "Artificial Intelligence" category. You must have an Amazon Web Services account to start using Amazon Lex. If you do not already have one, you will be prompted to create one during the sign-up process. Please refer to the Amazon Lex V2 Getting Started Guide for more information.

Q: What are the most common use cases for Amazon Lex?

The most common use-cases include:

- Self-service voice assistants and chatbots – build a call center bot
- Informational bot – build an automated customer support agent or bot that answers questions
- Application/Transactional bot – build a stand-alone pizza ordering agent or a travel bot
- Enterprise Productivity bot – build custom bots to connect to enterprise data resources
- Device Control bot– use Amazon Lex to issue control commands to connected devices

Q: How does Amazon Lex work with other AWS services?

Amazon Lex leverages AWS Lambda for Intent fulfillment, Amazon Cognito for user authentication, and Amazon Polly for text to speech.

Q: Do I have to be a machine learning expert to use Amazon Lex?

No machine learning expertise is necessary to use Amazon Lex. Developers can declaratively specify the conversation flow and Amazon Lex will take care of the speech recognition and natural language understanding functionality. Developers provide some sample utterances in plain English and the different parameters (slots) that they would like to collect from their user with the corresponding prompts. The language model gets built automatically.
Q: In which AWS regions is Amazon Lex available?

For a list of the supported Amazon Lex AWS regions, please visit the AWS Region Table for all AWS global infrastructure. Also for more information, see Regions and Endpoints in the AWS General Reference.
Q. What is the maximum bandwidth supported on Amazon Lex?

Amazon Lex scales to your needs and does not impose bandwidth constraints.
Q: Is Amazon Lex a managed service?

Amazon Lex is a completely managed service so you don't have to manage scaling of resources or maintenance of code. Your interaction schema and language models are automatically backed up. We also provide comprehensive versioning capability for easy rollback. Amazon Lex architecture does not require storage or backups of end user data.
Q: When do I use Amazon Polly vs. Amazon Lex?

Amazon Polly converts text inputs to speech. Amazon Lex is a service for building conversational interfaces using voice and text.
Q: Does Amazon Lex get more intelligent over time?

Yes. Amazon Lex uses deep learning to improve over time.

Bot Structure

Q: How do I create a bot in Amazon Lex?

To create a bot, you will first define the actions performed by the bot. These actions are the intents that need to be fulfilled by the bot. For each intent, you will add sample utterances and slots. Utterances are phrases that invoke the intent. Slots are input data required to fulfill the intent. Lastly, you will provide the business logic necessary to execute the action. An Amazon Lex bot can be created both via Console and REST APIs.
Q: Can I implement business logic on the client?

Yes. Amazon Lex provides the option of returning parsed intent and slots back to the client for business logic implementation.
Q: How can I validate user input?

Amazon Lex provides deep integration with AWS Lambda and you can validate user input using the initialization and validation codeHook. This code gets executed at every turn of the conversation. The codehook can be used to set up session parameters, validate user input and customize responses.
Q: What is an Intent?

To build an Amazon Lex bot, you will need to identify a set of actions - known as 'intents' -- that you want your bot to fulfill. A bot can have multiple intents. For example, a 'BookTickets' bot can have intents to make reservations, cancel reservations and review reservations.
Q: What is an utterance?

An 'utterance' is the spoken or typed phrase to invoke an intent. For example, to invoke the intent to make reservations, you would provide a sample utterance such as, "Can I make a reservation?"

Q: What are slots?

To fulfill an intent, the Amazon Lex bot needs information from the user. This information is captured in ‘slots’. For example, you would define show name and time as slots for intent to make reservations.
Q: What are prompts?

Amazon Lex elicits the defined ‘slots’ by using the ‘prompts’ provided. For example, to elicit value for the slot ‘time’ you will define a prompt such as “What show time would you like to reserve?”. Amazon Lex is capable of eliciting multiple slot values via a multi-turn conversation.
Q: How is an action fulfilled?

Amazon Lex integrates with AWS Lambda for ‘fulfillment’ of the action or business logic. Alternately, you can configure Amazon Lex to return parsed intent and slot values to the client for action fulfillment.
Q: How do I monitor and track my bot?

You can track metrics for your bot on the ‘Monitoring’ dashboard in the Amazon Lex Console. Currently, you can track the number of missed utterances, request latency and traffic by channel for your bot. You can view list of utterances that were not recognized by your bot, aka 'missed utterances'. With these monitoring capabilities, you view how your users are interacting with the bot and make improvements over time.

Automated chatbot designer

Q: How do I use the automated chatbot designer?

The automated chatbot designer helps you create a bot design in just a few clicks. You first provide a link to the S3 location that contains your conversation transcripts via the Lex Console (or the SDK). The automated chatbot designer will then process these transcripts to surface a chatbot design that includes user intents, sample phrases associated with those intents, and a list of all the information required to fulfill them. You can then review the results provided by the automated chatbot designer and add the intents and slot types that are best suited to your bot.
Q: What transcript formats are supported by the automated chatbot designer?

The transcripts must contain conversations between a caller and an agent in standardized JSON format. You can find a sample transcript in this format in Amazon Lex documentation. Amazon Connect customers using Contact Lens can directly use conversation transcripts in their original format. Conversation transcripts from other transcription services may require a simple conversion. You can find details about the conversion process here.
Q: Which languages are supported by the automated chatbot designer?

All English locales (US, UK, AU, IN, SA) supported by Amazon Lex are supported by the automated chatbot designer. At preview, the automated chatbot designer supports US English.

Build and Test

Q: What happens when I ‘build’ a bot?

Building a bot triggers machine learning and creates the models for your bot. A new version of your intents and slot types is created. Once created a version is immutable.
Q: How can I test an Amazon Lex bot?

You can test your Amazon Lex bot via the test window on the console. Any business logic implemented in AWS Lambda can be tested via this console as well. All supported browsers allow for testing text with your Amazon Lex bot; voice can be tested from a Chrome browser.

Deployment

Q: How can I create Amazon Lex bots for mobile?

Amazon Lex provides SDKs for iOS and Android. You can develop bots for your mobile use cases with these SDKs. User authentication can be enabled via Amazon Cognito.
Q: How can I make Amazon Lex bots available on messaging services?

Amazon Lex bots can be published to messaging platforms like Facebook Messenger, Slack, Kik, and Twilio SMS. To publish the bot you can provide the tokens for authentication in the console, and we will store it securely and provide a callback URL that you can provide to the chat service.
Q: Do I have to submit my bot for certification prior to deployment?

You don't need to certify your bot with Amazon prior to deployment.
Q: Can I have an Amazon Lex bot version deployed for use by end users while I continue to develop on a different version?

Yes. You can build and deploy a version of your bot into production while you continue to develop on a different version. Every version of an Amazon Lex bot will have an ARN. Each version can be associated with a different alias. You can use these tools to set up dev, stage and prod environments.
Q: Can I choose different versions while deploying to different messaging services?

Yes. You can deploy a specific version to each messaging service. Every version of Amazon Lex will have an ARN. Each version can be associated with an alias. You can use different aliases for deployment to different messaging service. Also, you can have multiple bots deployed to the same messaging service.
Q: What is the maximum duration of speech input?

Amazon Lex bots are designed for a request and response interaction or a continuous streaming conversation. With the request and response interaction, Amazon Lex supports up to 15 seconds of speech input. In a streaming conversation, all user inputs across multiple turns are processed in one streaming API call up to 15 seconds of speech input including silence.
Q: Can I configure for speech input and text output?

Yes, you can just choose the PostContent API to provide voice input and choose text output.
Q: How many languages are supported on Amazon Lex?

Currently, Amazon Lex supports US English, Spanish, French, German, Italian, Japanese, Australian English, British English, Canadian French, Latin American Spanish, and US Spanish.
Q: What audio formats does Amazon Lex support?

Amazon Lex supports the following formats for input audio: LPCM and Opus; Supported output audio formats: MPEG, OGG, PCM.
Q: Can I use Amazon Lex in VPC?

Amazon Lex can be accessed from VPC via public endpoints for building and running a bot. Currently, Amazon Lex does not provide a VPC endpoint.
Q: Can I access Amazon Lex bots locally i.e. without an Internet connection?

No. End users will need to access the Amazon Lex runtime endpoint over the Internet.

Lex V2

Q: What are the usability improvements offered in the V2 enhanced console and APIs?

Lex V2 console and APIs use an updated information architecture (IA) to deliver simplified versioning, support for multiple languages in a bot, and streaming capabilities. Additional improvements include the saving of partially completed bot configurations, renaming of resources, simplified navigation, bulk upload of utterances, and granular debugging.

Q: How can I use the streaming capability?

You can use the streaming API to conduct a continually streaming conversation with a Lex bot. With streaming conversation, the bot continuously listens and can be designed to respond proactively to user interruptions and pauses. For example, you can configure the bot to keep a conversation going when a user needs more time to respond by sending periodic messages such as “Take your time. Let me know once you are ready.”

Q: What are the pricing details for the V2 APIs?

Amazon Lex bots are designed for a request and response interaction or a continuous streaming conversation. With the request and response interaction, each user input (voice or text) is processed as a separate API call. In a streaming conversation, all user inputs across multiple turns are processed in one streaming API call.

Q: Can I integrate bots created using V2 APIs with Amazon Connect contact flows?

Yes, Amazon Connect contact flows work with both Lex V2 and V1 APIs. You can use the Lex V2 console to create and integrate bots with Amazon Connect.

Q: Can I take advantage of V2 API features for my existing bots?

No. If you want to take advantage of V2 features, you will need to recreate your bot with V2 APIs. The Lex V1 APIs are not compatible because V2 APIs use an updated information architecture to enable simplified resource versioning and support for multiple languages within a bot. Converting to V2 APIs is easy, so get started with this step by step migration guide.

Q: Which regions and languages do the V2 APIs support?

The Amazon Lex V2 APIs and enhanced console experience is available in all existing 8 regions and languages including US English, Spanish, French, German, Italian, Japanese, Australian English, British English, Canadian French, Latin American Spanish, and US Spanish. For a list of the supported Amazon Lex AWS regions, please visit the AWS Region Table.

Q: Will the support for new features such as simplified versioning and multiple languages in a bot be available in the existing APIs?

No. These features are only available in the V2 APIs. If you want to take advantage of these features, you can migrate to V2 APIs by following this migration guide.

Q: Will I be able to access the V1 console?

Yes, you can access the V1 console in the AWS Management Console. Once in the Lex console, you can navigate between the V1 and V2 console. The bots created in the V1 console will only be visible within the V1 Console. You will not be able to access your V1 bots in the V2 console until you recreate them in the V2 console. Migrating your bots to V2 is easy, here is a step by step migration guide.

Q: How do I access the V2 console?

You can click on the link in the left navigation bar to choose V1 or V2 as your console.

Q: Can I still use Lex V1 APIs?

Yes. The existing Lex V1 APIs are still supported. You can continue to use them to build and conduct your bot conversations.

Amazon Alexa

Q: How is this different from Alexa Skills Kit?

Alexa Skills Kit (ASK) is used to build skills for use in the Alexa ecosystem and devices and lets developers take advantage of all Alexa capabilities such as the Smart Home and Flash Briefing API, streaming audio and rich GUI experiences. Amazon Lex bots support both voice and text and can be deployed across mobile and messaging platforms.
Q: Do I need a wake word to invoke an Amazon Lex intent?

Amazon Lex does not support wake word functionality. The app that integrates with Amazon Lex will be responsible for triggering the microphone, i.e. push to talk.
Q: Can an Amazon Lex bot respond using Alexa's voice?

Currently we do not support the Alexa voice for Amazon Lex responses. However, there are 7 other voices from which to choose.
Q: Can I create an Alexa Skill from an Amazon Lex bot?

Amazon Lex provides the ability for you to export your Amazon Lex bot schema into a JSON file that is compatible with Amazon Alexa. Once downloaded as JSON, you need to log in to the Alexa developer portal, navigate to the 'Interaction Model' tab, launch the Alexa Skill Builder, and paste the bot schema into the Code Editor of your Alexa Skill. More details and steps can be found in the Amazon Lex documentation.
Q: When exporting my Amazon Lex bot schema to use in an Alexa skill, are my AWS Lambda functions exported and included in the bot schema?

No. Only the bot definition will be downloaded.
Q: I have created an Alexa Skill from an Amazon Lex bot using the schema export feature. Which Alexa platforms support the Amazon Lex bot schema?

All Alexa platforms that support Alexa skills can be used: The Amazon Echo, Amazon Dot, Amazon Look, Amazon Tap, Amazon Echo Show and any third-party Alexa-enabled devices.

Data and Security

Q: Are voice and text inputs processed by Amazon Lex stored, and how are they used by AWS?

Amazon Lex may store and use voice and text inputs processed by the service solely to provide and maintain the service and to improve and develop the quality of Amazon Lex and other Amazon machine-learning/artificial-intelligence technologies. Use of your content is necessary for continuous improvement of your Amazon Lex customer experience, including the development and training of related technologies. We do not use any personally identifiable information that may be contained in your content to target products, services or marketing to you or your end users. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you.

Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information. You may opt out of having your content used to improve and develop the quality of Amazon Lex and other Amazon machine-learning/artificial-intelligence technologies by using an AWS Organizations opt-out policy. For information about how to opt out, see Managing AI services opt-out policy.
Q: Can I delete voice and text inputs stored by Amazon Lex?

Yes. You can request deletion of voice and text inputs associated with your account by contacting opting out. Deleting voice and text inputs may degrade your Amazon Lex experience. For information about how to opt out, see Managing AI services opt-out policy.

Q: Who has access to my content that is processed and stored by Amazon Lex?

Only authorized employees will have access to your content that is processed by Amazon Lex. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.
Q: Do I still own my content that is processed and stored by Amazon Lex?

You always retain ownership of your content and we will only use your content with your consent.

Q: Is the content processed by Amazon Lex moved outside the AWS region where I am using Amazon Lex?

Any content processed by Amazon Lex is encrypted and stored at rest in the AWS region where you are using Amazon Lex. Some portion of content processed by Amazon Lex may be stored in another AWS region solely in connection with the continuous improvement and development of your Amazon Lex customer experience and other Amazon machine-learning/artificial-intelligence technologies. You can request deletion of voice and text inputs associated with your account by contacting AWS Support. Your trust, privacy, and the security of your content are our highest priority and we implement appropriate and sophisticated technical and physical controls, including encryption at rest and in transit, designed to prevent unauthorized access to, or disclosure of, your content and ensure that our use complies with our commitments to you. Please see https://aws.amazon.com/compliance/data-privacy-faq/ for more information.

Q: Can I use Amazon Lex in connection with websites, programs or other applications that are directed or targeted to children under age 13 and subject to the Children's Online Privacy Protection Act (COPPA)?

Yes, subject to your compliance with the Amazon Lex Service Terms, including your obligation to provide any required notices and obtain any required verifiable parental consent under COPPA, you may use Amazon Lex in connection with websites, programs, or other applications that are directed or targeted, in whole or in part, to children under age 13. Amazon Lex does not store or retain voice or text utterance information from websites, programs, or applications that are identified by customers in accordance with the Amazon Lex Service Terms as being directed or targeted, in whole or in part, to children under age 13 and subject to COPPA.
Q: How do I determine whether my website, program, or application is subject to COPPA?

For information about the requirements of COPPA and guidance for determining whether your website, program, or other application is subject to COPPA, please refer directly to the resources provided and maintained by the United States Federal Trade Commission. This site

also contains information regarding how to determine whether a service is directed or targeted, in whole or in part, to children under age 13. whole or in part, to children under age 13.

SDK

Q: What SDKs are supported for Amazon Lex?

Amazon Lex currently supports SDKs for runtime services. IoS and Android SDKs, as well as Java, JS, Python, CLI, .Net, Ruby, PHP, Go, and CPP support both text and speech input.
Q: Can I use SDKs to build bots?

You can build bots using SDKs: Java, JavaScript, Python, CLI, .NET, Ruby on Rails, PHP, Go, and CPP.

AWS DeepRacer

AWS DeepRacer is a reinforcement learning (RL)-enabled autonomous 1/18th-scale vehicle with supporting services in the AWS Machine Learning ecosystem. It offers an interactive learning system for users of all levels to acquire and refine their skill set in machine learning in general, and reinforcement learning in particular. You can use the AWS DeepRacer console to train and evaluate deep reinforcement learning models in simulation and then deploy them to an AWS DeepRacer vehicle for autonomous driving. You can also join the AWS DeepRacer League to race in the online Virtual Circuit or the in-person events.

Developers of all skill levels can get hands on with machine learning through a cloud based 3D racing simulator, fully autonomous 1/18th scale race car driven by reinforcement learning, and global racing league.

AWS DeepRacer gives an interesting and fun way to get started with reinforcement learning (RL). RL is an advanced machine learning (ML) technique that takes a very different approach to training models than other machine learning methods. Its super power is that it learns very complex behaviors without requiring any labeled training data, and can make short term decisions while optimizing for a longer term goal.

Simulator
Build models in Amazon SageMaker and train, test, and iterate quickly and easily on the track in the AWS DeepRacer 3D racing simulator.

Car
Experience the thrill of the race in the real-world when you deploy your reinforcement learning model onto AWS DeepRacer.

League
Compete in the world's first global, autonomous racing league, to race for prizes and glory and a chance to advance to the Championship Cup.

AWS DeepRacer Evo

AWS DeepRacer Evo is the next generation in autonomous racing. It comes fully equipped with stereo cameras and LiDAR sensor to enable object avoidance and head-to-head racing, giving developers everything they need to take their racing to the next level. In object avoidance races, developers use the sensors to detect and avoid obstacles placed on the track. In head-to-head, developers race against another DeepRacer on the same track and try to avoid it while still turning in the best lap time. Forward facing left and right cameras make up the stereo cameras, which helps the car learn depth information in images. This information can then be used to sense and avoid objects being approached on the track. The LiDAR sensor is backward facing and detects objects behind and beside the car.

What Is AWS DeepRacer?

AWS DeepRacer is an integrated learning system for users of all levels to learn and explore reinforcement learning and to experiment and build autonomous driving applications. It consists of the following components:

- AWS DeepRacer Console: an AWS Machine Learning service to train and evaluate reinforcement learning models in a simulated autonomous-driving environment.
- AWS DeepRacer Vehicle: a 1/18th scale RC car capable of running inference on a trained AWS DeepRacer model for autonomous driving.
- AWS DeepRacer League: the world's first global, autonomous racing league. Race for prizes, glory, and a chance to advance to the Championship Cup.

The AWS DeepRacer Console

The AWS DeepRacer console is a graphical user interface to interact with the AWS DeepRacer service. You can use the console to train a reinforcement learning model and to evaluate the model performance in the AWS DeepRacer simulator built upon AWS RoboMaker. In the console, you can also download a trained model for deployment to your AWS DeepRacer vehicle for autonomous driving in a physical environment.

In summary, the AWS DeepRacer console supports the following features:

- Create a training job to train a reinforcement learning model with a specified reward function, optimization algorithm, environment, and hyperparameters.
- Choose a simulated track to train and evaluate a model by using SageMaker and AWS RoboMaker.
- Clone a trained model to improve training by tuning hyperparameters to optimize your model's performance.
- Download a trained model for deployment to your AWS DeepRacer vehicle so it can drive in a physical environment.
- Submit your model to a virtual race and have its performance ranked against other models in a virtual leaderboard.

The AWS DeepRacer Vehicle

The AWS DeepRacer vehicle is a Wi-Fi enabled, physical vehicle that can drive itself on a physical track by using a reinforcement learning model.

- You can manually control the vehicle, or deploy a model for the vehicle to drive autonomously.
- The autonomous mode runs inference on the vehicle's compute module. Inference uses images that are captured from the camera that is mounted on the front.

- A Wi-Fi connection allows the vehicle to download software. The connection also allows the user to access the device console to operate the vehicle by using a computer or mobile device.

The AWS DeepRacer League

The AWS DeepRacer League is an important component of AWS DeepRacer. The AWS DeepRacer League is intended to foster communal learning and collaborative exploration through sharing and competition.

With the AWS DeepRacer League, you can have your development effort compared with other AWS DeepRacer developers in a physical or virtual racing event. Not only do you get a chance to win prizes, you also have a way to measure your reinforcement learning model. You can create opportunities to share your insights with other participants, to learn from each other, and to inspire each other.

AWS DeepRacer FAQs

General

Q: What is AWS DeepRacer

AWS DeepRacer is the fastest way to get rolling with reinforcement learning (RL), literally, with a fully autonomous 1/18th scale race car driven by reinforcement learning, 3D racing simulator, and a global racing league. Developers can train, evaluate, and tune RL models in the online simulator, deploy their models onto AWS DeepRacer for a real-world autonomous experience and compete in the AWS DeepRacer League for a chance to win the AWS DeepRacer Championship Cup. AWS DeepRacer Device Terms, Warranties, and Notices »

Q: What are the new features of AWS DeepRacer?

AWS DeepRacer customers can take on their next machine learning challenge using AWSDeepRacer with the launch of multi-car racing and object avoidance capabilities in the AWSDeepRacer console. Customers can now build models for object avoidance and dual-car head-to-head races by experimenting with multiple sensor inputs and the latest reinforcement learning algorithms and neural network configurations.

Q: How is AWS DeepRacer different from other robotic scale cars in the market?

AWS DeepRacer is the first autonomous scale car specifically developed to help developers get hands-on with reinforcement learning. AWS DeepRacer gives developers a simple way to learn RL, experiment with new RL algorithms and simulation-to-real domain transfer methods, and experience RL in the real world.

Q: What is the relationship between AWS DeepRacer and AWS Robocar Rally?

After the AWS DeepRacer launch, all Robocar Rally events will be replaced with AWS DeepRacer events. Robocar Rally inspired AWS DeepRacer, but unlike Robocar Rally, AWS DeepRacer focuses on RL and not on behavioral cloning.

Q: How can I get an AWS DeepRacer?

You can access the AWS DeepRacer 3D racing simulator in the AWS DeepRacer console to train your models, evaluate them, and take part in the AWS DeepRacer League. The AWS DeepRacer car is available for purchase on on Amazon.com.

Q: How can I get an AWS DeepRacer Evo?

The AWS DeepRacer, AWS DeepRacer Evo, and AWS DeepRacer Evo sensor kit are available to order on Amazon.com for shipping within, as well as outside of, the USA (excluding the United Kingdom, Israel, Kazakhstan, Russia, and countries subject to OFAC sanctions). If you would like to be notified when it is available in your region, you can sign up to the AWS DeepRacer Evo interest list.

Q: Which geographic regions is AWS DeepRacer available in?

AWS Customers can access the AWS DeepRacer simulator from the US East (N. Virginia) Region.

Q: What is the AWS DeepRacer League?

The AWS DeepRacer League is the world's first global autonomous racing league for developers. AWS Customers can use their AWS DeepRacer RL models to compete in a global championship, racing for prizes, glory, and a chance to lift the Championship Cup.

Learn more about the league in the dedicated FAQ section below

Getting Started

Q: How do I get started with AWS DeepRacer simulator?

The AWS DeepRacer simulator provides a tutorial to get you started with reinforcement learning and training your first model. You will then also be able to evaluate and tune your models, before racing them in the AWS DeepRacer League. The AWS DeepRacer Developer Documentation provides additional details on building your first model and also how to improve your models.

Q: How can customers get started with multi-car racing and object avoidance?

Customers can login into the AWS DeepRacer console from anywhere in the world. Here they can experiment with new sensor configurations in the Garage section of the console and build reinforcement learning models for head to head and object avoidance racing.

Q: How can I start my own Community Race?

Customers can create their own virtual races the Community Races section of the AWS DeepRacer console.

Q: Do I need an AWS DeepRacer car in order to use the AWS DeepRacer simulator?

No. You can train models, evaluate them without owning an AWS DeepRacer car. Furthermore, you can race your models in both the AWS DeepRacer League Virtual and Summit Circuits without owning an AWS DeepRacer car.

Q: Does my DeepRacer need to be connected to the internet to race autonomously?

No. AWS DeepRacer uses the deployed RL model and input from the camera to run inference locally. AWS DeepRacer must be connected to the same Wi-Fi network as the device used to start and stop autonomous driving. Details on how to set up your vehicle can be found in the Developer Documentation.

AWS DeepRacer League

Q: How do I race in the DeepRacer League?

Join the AWS DeepRacer League from anywhere in the world via the AWS DeepRacer console and put your skills to the test on virtual tracks monthly. Race for prizes and glory and a chance to advance to compete in the AWS DeepRacer Championship Cup. New for 2021 Developers will start in the Open Division of the league. Each month the top 10% in the Open Division will advance to the Pro Division the folowing month. The top 16 racers at the end of the monthly Pro Division race qualify for the Pro Finale, where they will race live to determine the winners for that month. The top 10 in the Pro Finale receive AWS DeepRacer Evo devices, while the top 3 will be eligible to receive an expenses-paid trip to advance to compete in the AWS DeepRacer Championship Cup at re:Invent.

Q: How many times can I enter the League?

There is no limit on the number of races or race formats you can enter, in-person or virtually. In fact, participating in multiple racing events, increases your chances of winning one of the prizes to advance to compete for the Championship Cup.

Q: What are the prizes?

Details of the AWS DeepRacer League prizes can be found here. For full details please read the AWS DeepRacer League official rules.

Q: Do I have to build my own model to race?

No, you do not need to build your own model to compete in the AWS DeepRacer League. There are tutorials inside the AWS DeepRacer simulator on how to build your own model, as well as pre-trained sample models you can use to get started.

Q: How long does a Virtual Circuit race last?

Each virtual race will last one month (from the 1st business day of the month to the last day of the month).

Q: How many Virtual Circuit races are there?

Starting on March 1st, the Virtual Circuit will have 8 monthly races, with three racing formats in each(time trial, object avoidance, and head-to-head). Each race will be revealed in the console on the first business day of each month leading up to re:Invent 2021 (March – October).

Q: How do I score points?

In 2021, the AWS DeepRacer league will no longer track cumulative points. Winners will be awarded based on single race performance only.

Q: What is AWS Summit Online?

The AWS Summit Online is a free online event designed to bring the cloud computing community together to connect, collaborate, and learn about AWS. At a Summit Online event, you can learn how to choose the right database, modernize your data warehouse, and drive digital transformation with AI using AWS services.

AWS DeepRacer Student

Q: What is AWS DeepRacer Student?

AWS DeepRacer Student, presented by Intel, lets high school and college-enrolled students around the globe develop their ML skills in a fun, hands-on autonomous racing league exclusively for students with no AWS account required. Participants can leverage 20 hours of ML educational material, 10 hours of monthly model training compute resources, and 5 GB of model storage to train a RL model in a simulated 3D racing environment.

Q: When does the AWS DeepRacer Student League begin?

The official AWS DeepRacer Student League season, kicked off on March 1, 2022. Competitions run monthly within AWS DeepRacer Student until September 30, 2022. The top participants in each monthly race will qualify for the Global AWS DeepRacer Student League Championships in October, 2022.

Q: Who can participate in the AWS DeepRacer Student League?

Any student over the age of 16 and currently enrolled in high school or undergraduate higher education at an accredited institution globally may apply, subject to terms and conditions.

Q: Does it cost anything to compete in the AWS DeepRacer Student League?

No, AWS DeepRacer Student is free for students to learn, train, and participate in the AWS DeepRacer Student League. Students can compete monthly to win prizes such as AWS DeepRacer devices and funds applied towards postsecondary tuition.

Q: Where can I find more details regarding AWS DeepRacer Student?

Go to AWS DeepRacer Student for more details.

Q: Where can I find the official AWS DeepRacer Student League rules?
For full details, please read the AWS DeepRacer Student League Official Rules.

Product Details

Q: What are the product specifications of the AWS DeepRacer device?

Car: 1/18th scale 4WD with monster truck chassis
- CPU: Intel Atom™ Processor
- Memory: 4GB RAM
- Storage: 32GB (expandable)
- Wi-Fi: 802.11ac
- Camera: 4 MP camera with MJPEG
- Software: Ubuntu OS 16.04.3 LTS, Intel® OpenVINO™ toolkit, ROS Kinetic
- Drive battery: 7.4V/1100mAh lithium polymer
- Compute battery: 13600mAh USB-C PD
- Ports: 4x USB-A, 1x USB-C, 1x Micro-USB, 1x HDMI
- Sensors: Integrated accelerometer and gyroscope

Additional sensors on AWS DeepRacer Evo are LiDAR and 1 stereo 4 MP camera with MJPEG.

Q: If I already own an AWS DeepRacer car will it be able to master multi-car racing and object avoidance?

Developers who already own a DeepRacer car can buy an easy-to-install sensor kit from Amazon.com to give their car the same capabilities as AWS DeepRacer Evo.

Q: What does the LIDAR and stereo camera do?

LIDAR stands for light detection and ranging. It provides a continuous light beam providing the reinforcement model with inputs about whether a car is fast approaching from behind. The stereo camera adds depth perception to allow the car to detect objects in the road and be more responsive to its environment. Combining these sensory inputs with advanced algorithms, and updated reward functions, developers can build models that will not only detect obstacles (including other cars), but will also decide when to overtake to beat the other car to the finish line.

Q: Which AWS services are integrated with AWS DeepRacer?

AWS DeepRacer integrates with Amazon SageMaker for reinforcement learning model training, AWS RoboMaker to provide the racing simulator, Amazon Kinesis Video Streams for video streaming of virtual simulation footage, Amazon S3 for model storage, and Amazon CloudWatch for log capture.

Q: How do I add new tracks to the AWS DeepRacer simulator?

Currently, developers cannot add additional tracks to the AWS DeepRacer simulator. AWS DeepRacer will release a number of new racing tracks throughout 2021.

Q: Can I train my models on the AWS DeepRacer device?

No. Training an RL model requires feedback regarding the outcome of actions taken by the model. This feedback loop exists in the AWS DeepRacer simulator, but not in the real-world.

Q: Can I train my models locally on my own machine as opposed to the AWS Cloud?

Currently AWS DeepRacer does not support local training.

Q: Can I train AWS DeepRacer RL models directly on Amazon SageMaker?

Yes. You can use the AWS DeepRacer Distributed Training SageMaker Notebook to create and train RL models. You will be able to deploy these models to your AWS DeepRacer manually, but won't yet be able to import them into the AWS DeepRacer console.

AWS Identity and Access Management (IAM) and AWS IAM Identity Center (Successor to AWS Single Sign-On)

AWS Identity and Access Management (IAM)

Securely manage identities and access to AWS services and resources.

AWS IAM Identity Center (Successor to AWS Single Sign-On)

Centrally manage workforce access to multiple AWS accounts and applications.

AWS Identity and Access Management (IAM)

How it works

With AWS Identity and Access Management (IAM), you can specify who or what can access services and resources in AWS, centrally manage fine-grained permissions, and analyze access to refine permissions across AWS.

Why use IAM?

Use IAM to manage and scale workload and workforce access securely supporting your agility and innovation in AWS.

Use cases

Apply fine-grained permissions and scale with attribute-based access control
Create granular permissions based on user attributes—such as department, job role, and team name—by using attribute-based access control.

Manage per-account access or scale access across AWS accounts and applications
Manage per-account identities with IAM or use IAM Identity Center to provide multi-account access and application assignments across AWS.

Establish organization-wide and preventative guardrails on AWS
Use service control policies to establish permissions guardrails for IAM users and roles, and implement a data perimeter around your accounts in AWS Organizations.

Set, verify, and right-size permissions toward least privilege
Streamline permissions management and use cross-account findings as you set, verify, and refine policies on the journey toward least privilege.

What is IAM?

AWS Identity and Access Management (IAM) is a web service that helps you securely control access to AWS resources. You use IAM to control who is authenticated (signed in) and authorized (has permissions) to use resources.

When you create an AWS account, you begin with one sign-in identity that has complete access to all AWS services and resources in the account. This identity is called the AWS account *root user* and is accessed by signing in with the email address and password that you used to create the account. We strongly recommend that you do not use the root user for your everyday tasks. Safeguard your root user credentials and use them to perform the tasks that only the root user can perform. For the complete list of tasks that require you to sign in as the root user, see Tasks that require root user credentials in the *AWS General Reference*.

IAM features

IAM gives you the following features:

Shared access to your AWS account

You can grant other people permission to administer and use resources in your AWS account without having to share your password or access key.

Granular permissions

You can grant different permissions to different people for different resources. For example, you might allow some users complete access to Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple Storage Service (Amazon S3), Amazon DynamoDB, Amazon Redshift, and other AWS services. For other users, you can allow read-only access to just some S3 buckets, or permission to administer just some EC2 instances, or to access your billing information but nothing else.

Secure access to AWS resources for applications that run on Amazon EC2

You can use IAM features to securely provide credentials for applications that run on EC2 instances. These credentials provide permissions for your application to access other AWS resources. Examples include S3 buckets and DynamoDB tables.

Multi-factor authentication (MFA)

You can add two-factor authentication to your account and to individual users for extra security. With MFA you or your users must provide not only a password or access key to work with your account, but also a code from a specially configured device. If you already use a FIDO security key with other services, and it has an AWS supported configuration. For more information, see Supported configurations for using FIDO security keys.

Identity federation

You can allow users who already have passwords elsewhere—for example, in your corporate network or with an internet identity provider—to get temporary access to your AWS account.

Identity information for assurance

If you use AWS CloudTrail, you receive log records that include information about those who made requests for resources in your account. That information is based on IAM identities.

PCI DSS Compliance

IAM supports the processing, storage, and transmission of credit card data by a merchant or service provider, and has been validated as being compliant with Payment Card Industry (PCI) Data Security Standard (DSS). For more information about PCI DSS, including how to request a copy of the AWS PCI Compliance Package, see PCI DSS Level 1.

Integrated with many AWS services

For a list of AWS services that work with IAM, see AWS services that work with IAM.

Eventually Consistent

IAM, like many other AWS services, is eventually consistent. IAM achieves high availability by replicating data across multiple servers within Amazon's data centers around the world. If a request to change some data is successful, the change is committed and safely stored. However, the change must be replicated across IAM, which can take some time. Such changes include creating or updating users, groups, roles, or policies. We recommend that you do not include such IAM changes in the critical, high-availability code paths of your application. Instead, make IAM changes in a separate initialization or setup routine that you

run less frequently. Also, be sure to verify that the changes have been propagated before production workflows depend on them. For more information, see Changes that I make are not always immediately visible.

Free to use

AWS Identity and Access Management (IAM) and AWS Security Token Service (AWS STS) are features of your AWS account offered at no additional charge. You are charged only when you access other AWS services using your IAM users or AWS STS temporary security credentials.

Accessing IAM

You can work with AWS Identity and Access Management in any of the following ways.

AWS Management Console

The console is a browser-based interface to manage IAM and AWS resources. For more information about accessing IAM through the console, see Signing in to the AWS Management Console as an IAM user or root user. For a tutorial that guides you through using the console, see Creating your first IAM admin user and user group.

AWS Command Line Tools

You can use the AWS command line tools to issue commands at your system's command line to perform IAM and AWS tasks. Using the command line can be faster and more convenient than the console. The command line tools are also useful if you want to build scripts that perform AWS tasks.

AWS provides two sets of command line tools: the AWS Command Line Interface (AWS CLI) and the AWS Tools for Windows PowerShell. For information about installing and using the AWS CLI, see the AWS Command Line Interface User Guide. For information about installing and using the Tools for Windows PowerShell, see the AWS Tools for Windows PowerShell User Guide.

AWS SDKs

AWS provides SDKs (software development kits) that consist of libraries and sample code for various programming languages and platforms (Java, Python, Ruby, .NET, iOS, Android, etc.). The SDKs provide a convenient way to create

programmatic access to IAM and AWS. For example, the SDKs take care of tasks such as cryptographically signing requests, managing errors, and retrying requests automatically. For information about the AWS SDKs, including how to download and install them, see the Tools for Amazon Web Services page.

IAM HTTPS API

You can access IAM and AWS programmatically by using the IAM HTTPS API, which lets you issue HTTPS requests directly to the service. When you use the HTTPS API, you must include code to digitally sign requests using your credentials. For more information, see Calling the IAM API using HTTP query requests and the IAM API Reference.

AWS IAM Identity Center (Successor to AWS Single Sign-On)

How it works

AWS IAM Identity Center (successor to AWS Single Sign-On) helps you securely create or connect your workforce identities and manage their access centrally across AWS accounts and applications. IAM Identity Center is the recommended approach for workforce authentication and authorization on AWS for organizations of any size and type.

Why use IAM Identity Center?

Use IAM Identity Center to securely scale access across accounts and applications, supporting your workforce agility and workload innovation on AWS.

Use cases

Enable multi-account access to your AWS accounts
Your users can use their directory credentials for single sign-on access to multiple AWS accounts. Their personalized web user portal shows their assigned roles in AWS accounts in one place. Users can also single sign-on through the AWS Command Line Interface (CLI), AWS SDKs, or AWS Console Mobile Application using their directory credentials for a consistent authentication experience.

Enable single sign-on access to your AWS applications
IAM Identity Center is integrated with applications such as Amazon SageMaker Studio, AWS Systems Manager Change Manager, and AWS IoT SiteWise for zero-configuration authentication and authorization. These integrated applications share a consistent view of users and groups for resource sharing and collaboration all within the application.

Enable single sign-on access to Amazon EC2 Windows instances
Securely access your Amazon EC2 Windows instances with existing corporate user names, passwords, and MFA devices. You are not required to share administrator credentials, access credentials multiple times, or configure remote access client software. You can centrally grant and revoke access to your EC2 Windows instances at scale across multiple AWS accounts.

Enable single sign-on access to cloud-based applications
You can more easily configure single sign-on access to applications that support SAML 2.0 using the IAM Identity Center application configuration wizard. IAM Identity Center also provides preconfigured settings for many cloud applications, including Salesforce, Box, and Microsoft 365.

What is IAM Identity Center?

AWS IAM Identity Center (successor to AWS Single Sign-On) expands the capabilities of AWS Identity and Access Management (IAM) to provide a central place that brings together administration of users and their access to AWS accounts and cloud applications. Although the service name AWS Single Sign-On has been retired, the term *single sign-on* is still used throughout this guide to describe the authentication scheme that allows users to sign in one time to access multiple applications and websites.

With IAM Identity Center you can manage sign-in security for your workforce by creating or connecting your users and groups to AWS in one place. With *multi-account permissions* you can assign your *workforce identities* access to AWS accounts. You can use *application assignments* to assign your users access to software as a service (SaaS) applications. With a single click, IAM Identity Center enabled application admins can assign access to your workforce users, and can also use application assignments to assign your users access to software as a service (SaaS) applications.

IAM Identity Center features

IAM Identity Center includes the following core features:

Workforce identities

Human users that are members of your organization are also known as workforce identities or *workforce users*. You can create workforce users and groups in IAM Identity Center, or connect and synchronize to an existing set of users and groups in your own identity source for use across all your AWS accounts and applications. Supported identity sources include Microsoft Active Directory Domain Services, and external identity providers such as Okta Universal Directory or Microsoft Azure AD.

Application assignments for SAML applications

With application assignments, you can grant your workforce users in IAM Identity Center single sign-on access to SAML 2.0 applications, such as Salesforce and Microsoft 365. Your users can access these applications in a single place, without the need for you to set up separate federation.

Identity Center enabled applications

AWS applications and services, such as Amazon Managed Grafana, Amazon Monitron, and Amazon SageMaker Studio Notebooks, discover and connect to IAM Identity Center automatically to receive sign-in and user directory services. This provides users with a consistent single sign-on experience to these applications with no additional configuration of the applications. Because the applications share a common view of users, groups, and group membership, users also have a consistent experience when sharing application resources with others.

Multi-account permissions

With multi-account permissions you can plan for and centrally implement IAM permissions across multiple AWS accounts at one time without needing to configure each of your accounts manually. You can create fine-grained permissions based on common job functions or define custom permissions that meet your security needs. You can then assign those permissions to workforce users to control their access over specific accounts.

AWS access portal

The AWS access portal provides your workforce users with one-click access to all their assigned AWS accounts and cloud applications through a simple web portal.

IAM Identity Center rename

On July 26, 2022, AWS Single Sign-On was renamed to AWS IAM Identity Center (successor to AWS Single Sign-On). For existing customers, the following table is meant to describe some of the more common term changes that have been updated throughout this guide as a result of the rename.

Legacy term	Current term
AWS SSO user *or* SSO user	workforce user *or* user
AWS SSO user portal *or* user portal	AWS access portal
AWS SSO-integrated applications	Identity Center enabled applications
AWS SSO directory	Identity Center directory
AWS SSO store or AWS SSO identity store	identity store used by IAM Identity Center

The following table describes the applicable user, developer and API reference guide name changes that also took place as a result of this rename.

Legacy guide	Current guide
AWS Single Sign-On User Guide	IAM Identity Center User Guide
AWS Single Sign-On SCIM Implementation Developer Guide	IAM Identity Center SCIM Implementation Developer Guide
AWS Single Sign-On API Reference Guide	IAM Identity Center API Reference
AWS Single Sign-On Identity Store API Reference Guide	Identity Store API Reference
AWS Single Sign-On OIDC API Reference Guide	IAM Identity Center OIDC API Reference
AWS Single Sign-On Portal API Reference Guide	IAM Identity Center Portal API Reference

Legacy namespaces remain the same

The `sso` and `identitystore` API namespaces along with the following related namespaces **remain unchanged** for backward compatibility purposes.

- CLI commands
  - aws configure sso
  - identitystore
  - sso
  - sso-admin
  - sso-oidc
- Managed policies containing AWSSSO and AWSIdentitySync prefixes
- Service endpoints containing sso and identitystore
- AWS CloudFormation resources containing AWS::SSO prefixes
- Service-linked role containing AWSServiceRoleForSSO
- Console URLs containing sso and singlesignon
- Documentation URLs containing singlesignon

AWS Identity and Access Management (IAM) FAQs

General

Q: What is AWS Identity and Access Management (IAM)?
IAM provides fine-grained access control across all of AWS. With IAM, you can control access to services and resources under specific conditions. Use IAM policies to manage permissions for your workforce and systems to ensure least privilege. IAM is offered at no additional charge. For more information, see What is IAM?

Q: How does IAM work and what can I do with it?
IAM provides authentication and authorization for AWS services. A service evaluates if an AWS request is allowed or denied. Access is denied by default and is allowed only when a policy explicitly grants access. You can attach policies to roles and resources to control access across AWS. For more information, see Understanding how IAM works.

Q: What are least-privilege permissions?
When you set permissions with IAM policies, grant only the permissions required to perform a task. This practice is known as granting least privilege. You can apply least-privilege permissions in IAM by defining the actions that can be taken on specific resources under specific conditions. For more information, see Access management for AWS resources.

Q: How do I get started with IAM?
To get started using IAM to manage permissions for AWS services and resources, create an IAM role and grant it permissions. For workforce users, create a role that can be assumed by your identity provider. For systems, create a role that can be assumed by the service you are using, such as Amazon EC2 or AWS Lambda. After you create a role, you can attach a policy to the role to grant permissions that meet your needs. When you are just starting out, you might not know the specific permissions you need, so you can start with broader permissions. AWS managed policies provide permissions to help you get started and are available in all AWS accounts. Then, reduce permissions further by defining customer managed policies specific to your use cases. You can create and manage policies and roles in the IAM console, or via AWS APIs or the AWS CLI. For more information, see Getting started with IAM.

IAM resources

Q: What are IAM roles and how do they work?
AWS Identity and Access Management (IAM) roles provide a way to access AWS by relying on temporary security credentials. Each role has a set of permissions for making AWS service requests, and a role is not associated with a specific user or group. Instead, trusted entities such as identity providers or AWS services assume roles. For more information, see IAM roles.

Q: Why should I use IAM roles?
You should use IAM roles to grant access to your AWS accounts by relying on short-term credentials, a security best practice. Authorized identities, which can be AWS services or users from your identity provider, can assume roles to make AWS requests. To grant permissions to a role, attach an IAM policy to it. For more information, see Common scenarios for roles.

Q: What are IAM users and should I still be using them?
IAM users are identities with long-term credentials. You might be using IAM users for workforce users. In this case, AWS recommends using an identity provider and federating into AWS by assuming roles. You also can use roles to grant cross-account access to services and features such as AWS Lambda functions. In some scenarios, you might require IAM users with access keys that have long-term credentials with access to your AWS account. For these scenarios, AWS recommends using IAM access last used information to rotate credentials often and remove credentials that are not being used. For more information, see Overview of AWS identity management: Users.

Q: What are IAM policies?
IAM policies define permissions for the entities you attach them to. For example, to grant access to an IAM role, attach a policy to the role. The permissions defined in the policy determine whether requests are allowed or denied. You also can attach policies to some resources, such as Amazon S3 buckets, to grant direct, cross-account access. And you can attach policies to an AWS organization or organizational unit to restrict access across multiple accounts. AWS evaluates these policies when an IAM role makes a request. For more information, see Identity-based policies.

Granting access

Q: How do I grant access to services and resources by using IAM?
To grant access to services and resources by using AWS Identity and Access Management (IAM), attach IAM policies to roles or resources. You can start by attaching AWS managed policies, which are owned and updated by AWS and are available in all AWS accounts. If you know the specific permissions required for your use cases, you can create customer managed policies and attach them to roles. Some AWS resources provide a way to grant access by defining a policy attached to resources, such as Amazon S3 buckets. These resource-based policies allow you to grant direct, cross-account access to the resources they are attached to. For more information, see Access management for AWS resources.

Q: How do I create IAM policies?
To assign permissions to a role or resource, create a policy, which is a JavaScript Object Notation (JSON) document that defines permissions. This document includes permissions statements that grant or deny access to specific service actions, resources, and conditions. After you create a policy, you can attach it to one or more AWS roles to grant permissions to your AWS account. To grant direct, cross-account access to resources, such as Amazon S3 buckets, use resource-based policies. Create your policies in the IAM console or via AWS APIs or the AWS CLI. For more information, see Creating IAM policies.

Q: What are AWS managed policies and when should I use them?
AWS managed policies are created and administered by AWS and cover common use

cases. Getting started, you can grant broader permissions by using the AWS managed policies that are available in your AWS account and common across all AWS accounts. Then, as you refine your requirements, you can reduce permissions by defining customer managed policies specific to your use cases with the goal of achieving least-privilege permissions. For more information, see AWS managed policies.

Q: What are customer managed policies and when should I use them?
To grant only the permissions required to perform tasks, you can create customer managed policies that are specific to your use cases and resources. Use customer managed policies to continue refining permissions for your specific requirements. For more information, see Customer managed policies.

Q: What are inline policies and when should I use them?
Inline policies are embedded in and inherent to specific IAM roles. Use inline policies if you want to maintain a strict one-to-one relationship between a policy and the identity to which it is applied. For example, you can grant administrative permissions to ensure they are not attached to other roles. For more information, see Inline policies.

Q: What are resource-based policies and when should I use them?
Resource-based policies are permissions policies that are attached to resources. For example, you can attach resource-based policies to Amazon S3 buckets, Amazon SQS queues, VPC endpoints, and AWS Key Management Service encryption keys. For a list of services that support resource-based policies, see AWS services that work with IAM. Use resource-based policies to grant direct, cross-account access. With resource-based policies, you can define who has access to a resource and which actions they can perform with it. For more information, see Identity-based policies and resource-based policies.

Q: What is role-based access control (RBAC)?
RBAC provides a way for you to assign permissions based on a person's job function, known outside of AWS as a role. IAM provides RBAC by defining IAM roles with permissions that align with job functions. You then can grant individuals access to assume these roles to perform specific job functions. With RBAC, you can audit access by looking at each IAM role and its attached permissions. For more information, see Comparing ABAC to the traditional RBAC model.

Q: How do I grant access with RBAC?
As a best practice, grant access only to the specific service actions and resources required to perform each task. This is known as granting least privilege. When employees add new resources, you must update policies to allow access to those resources.

Q: What is attribute-based access control (ABAC)?
ABAC is an authorization strategy that defines permissions based on attributes. In AWS, these attributes are called tags, and you can define them on AWS resources, IAM roles, and in role sessions. With ABAC, you define a set of permissions based on the value of a tag. You can grant fine-grained permissions to specific resources by requiring the tags on the role or session to match the tags on the resource. For example, you can author a policy that grants developers access to resources tagged with the job title "developers." ABAC is helpful in environments that are growing rapidly by granting permissions to resources as they are created with specific tags. For more information, see Attribute-Based Access Control for AWS.

Q: How do I grant access by using ABAC?
To grant access by using ABAC, first define the tag keys and values you want to use for access control. Then, ensure your IAM role has the appropriate tag keys and values. If multiple identities use this role, you also can define session tag keys and values. Next, ensure that your resources have the appropriate tag keys and values. You also can require users to create resources with appropriate tags and restrict access to modify them. After your tags are in place,

define a policy that grants access to specific actions and resource types, but only if the role or session tags match the resource tags. For a detailed tutorial that demonstrates how to use ABAC in AWS, see IAM tutorial: Define permissions to access AWS resources based on tags.

Restricting access

Q: How do I restrict access by using IAM?
With AWS Identity and Access Management (IAM), all access is denied by default and requires a policy that grants access. As you manage permissions at scale, you might want to implement permissions guardrails and restrict access across your accounts. To restrict access, specify a Deny statement in any policy. If a Deny statement applies to an access request, it always prevails over an Allow statement. For example, if you allow access to all actions in AWS but deny access to IAM, any request to IAM is denied. You can include a Deny statement in any type of policy, including identity-based, resource-based, and service control policies with AWS Organizations. For more information, see Controlling access with AWS Identity and Access Management.

Q: What are AWS Organizations service control policies (SCPs) and when should I use them?
SCPs are similar to IAM policies and use almost the same syntax. However, SCPs don't grant permissions. Instead, SCPs allow or deny access to AWS services for individual AWS accounts with Organizations member accounts, or for groups of accounts within an organizational unit. The specified actions from an SCP affect all IAM users and roles, including the root user of the member account. For more information, see Policy evaluation logic.

Analyzing access

Q: How do I work toward least-privilege permissions?
When you get started granting permissions, you can start with broader permissions as you explore and experiment. As your use cases mature, AWS recommends that you refine permissions to grant only the permissions required with the goal of achieving least-privilege permissions. AWS provides tools to help you refine your permissions. You can start with AWS managed policies, which are created and administered by AWS and include permissions for common use cases. As you refine your permissions, define specific permissions in customer managed policies. To help you determine the specific permissions you require, use AWS Identity and Access Management (IAM) Access Analyzer, review AWS CloudTrail logs, and inspect last access information. You also can use the IAM policy simulator to test and troubleshoot policies.

Q: What is IAM Access Analyzer?
Achieving least privilege is a continuous cycle to grant the right fine-grained permissions as your requirements evolve. IAM Access Analyzer helps you streamline permissions management in each step of this cycle. Policy generation with IAM Access Analyzer generates a fine-grained policy based on the access activity captured in your logs. This means that after you build and run an application, you can generate policies that grant only the required permissions to operate the application. Policy validation with IAM Access Analyzer uses more than 100 policy checks to guide you to author and validate secure and functional policies. You can use these checks while creating new policies or to validate existing policies. Public and cross-account findings with IAM Access Analyzer help you verify and refine access allowed by your resource policies from outside your AWS organization or account. For more information, see Using IAM Access Analyzer.

Q: How do I remove unused permissions?
You might have IAM users, roles, and permissions that you no longer require in your AWS account. We recommend that you remove them with the goal of achieving least-privilege

access. For IAM users, you can review password and access key last used information. For roles, you can review role last used information. This information is available through the IAM console, APIs, and SDKs. Last used information helps you identify users and roles that are no longer in use and safe to remove. You also can refine permissions by reviewing service and last accessed information to identify unused permissions. For more information, see Refining permissions in AWS using last accessed information.

Q: What is the IAM policy simulator and when should I use it?
The IAM policy simulator evaluates policies you choose and determines the effective permissions for each of the actions you specify. Use the policy simulator to test and troubleshoot identity-based and resource-based policies, IAM permissions boundaries, and SCPs. For more information, see Testing IAM policies with the IAM policy simulator.

AWS IAM Identity Center (Successor to AWS Single Sign-On) FAQs

General

Q: What is AWS IAM Identity Center (successor to AWS Single Sign-On)?

IAM Identity Center is built on top of AWS Identity and Access Management (IAM) to simplify access management to multiple AWS accounts, AWS applications, and other SAML-enabled cloud applications. In IAM Identity Center, you create, or connect, your workforce users for use across AWS. You can choose to manage access just to your AWS accounts, just to your cloud applications, or to both. You can create users directly in IAM Identity Center, or you can bring them from your existing workforce directory. With IAM Identity Center, you get a unified administration experience to define, customize, and assign fine-grained access. Your workforce users get a user portal to access their assigned AWS accounts or cloud applications.

Q: What are the benefits of IAM Identity Center?

You can use IAM Identity Center to quickly and easily assign and manage your employees' access to multiple AWS accounts, SAML-enabled cloud applications (such as Salesforce, Microsoft 365, and Box), and custom-built in-house applications, all from a central place. Employees can be more productive by signing in with their existing credentials or credentials that you configure in IAM Identity Center. They can use a single personalized user portal. You'll get better visibility into cloud application use because you can monitor and audit sign-in activity centrally from AWS CloudTrail.

Q: What problems does IAM Identity Center solve?

IAM Identity Center eliminates the administrative complexity of federating and managing permissions separately for each AWS account. It allows you to set up AWS applications from a single interface, and to assign access to your cloud applications from a single place.
IAM Identity Center also helps improve access visibility by integrating with AWS CloudTrail and providing a central place for you to audit single sign-on access to AWS accounts and SAML-enabled cloud applications, such as Microsoft 365, Salesforce, and Box.

Q: Why should I use IAM Identity Center?

IAM Identity Center is our recommended front door into AWS. It should be your primary tool to manage the AWS access of your workforce users. It allows you to manage your identities in your preferred identity source, connect them once for use in AWS, allows you to define fine-grained permissions and apply them consistently across accounts. As the number of your

accounts scales, IAM Identity Center gives you the option to use it as a single place to manage user access to all your cloud applications.

Q: What can I do with IAM Identity Center?

You can use IAM Identity Center to quickly and easily assign your employees access to AWS accounts within AWS Organizations, business cloud applications (such as Salesforce, Microsoft 365, and Box), and custom applications that support Security Assertion Markup Language (SAML) 2.0. Employees can sign in with their existing corporate credentials or credentials they configure in IAM Identity Center to access their business applications from a single user portal. IAM Identity Center also allows you to audit users' access to cloud services by using AWS CloudTrail.

Q: Who should use IAM Identity Center?

IAM Identity Center is for administrators who manage multiple AWS accounts and business applications, want to centralize user access management to these cloud services, and want to provide employees a single location to access these accounts and applications without them having to remember yet another password.

Q: How do I start using IAM Identity Center?

As a new IAM Identity Center customer, you:

1. Sign in to the AWS Management Console of the management account in your AWS account and navigate to the IAM Identity Center console.
2. Select the directory you use for storing the identities of your users and groups from the IAM Identity Center console. IAM Identity Center provides you a directory by default that you can use to manage users and groups in IAM Identity Center. You can also change directory to connect to a Microsoft AD directory by clicking through a list of Managed Microsoft AD and AD Connector instances that IAM Identity Center discovers in your account automatically. If you want to connect to a Microsoft AD directory, see Getting Started with AWS Directory Service.
3. Grant users single sign-on access to AWS accounts in your organization by selecting the AWS accounts from a list populated by IAM Identity Center, and then selecting users or groups from your directory and the permissions you want to grant them.
4. Give users access to business cloud applications by:
   a. Selecting one of the applications from the list of pre-integrated applications supported in IAM Identity Center.
   b. Configuring the application by following the configuration instructions.
   c. Selecting the users or groups that should be able to access this application.
5. Give your users the IAM Identity Center sign-in web address that was generated when you configured the directory so that they can sign in to IAM Identity Center and access accounts and business applications.

Q: How much does IAM Identity Center cost?

IAM Identity Center is offered at no extra charge.

Q: In which regions is IAM Identity Center available?

See the AWS Region Table for IAM Identity Center availability by Region.

Identity sources and applications support

Q: What identity sources can I use with IAM Identity Center?

With IAM Identity Center, you can create and manage user identities in IAM Identity Center's identity store, or easily connect to your existing identity source including Microsoft Active Directory, Okta Universal Directory, Azure Active Directory (Azure AD), or another supported IdP. See the IAM Identity Center User Guide to learn more.

Q: Can I connect more than one identity source to IAM Identity Center?

No. At any given time, you can have only one directory or one SAML 2.0 identity provider connected to IAM Identity Center. But, you can change the identity source that is connected to a different one.

Q: What SAML 2.0 IdPs can I use with IAM Identity Center?

You can connect IAM Identity Center to most SAML 2.0 IdPs, such as Okta Universal Directory or Azure Active Directory. See the IAM Identity Center User Guide to learn more.

Q: How can I provision identities from my existing IdPs into IAM Identity Center?

Identities from your existing IdP must be provisioned into IAM Identity Center before you can assign permissions. You can synchronize user and group information from Okta Universal Directory, Azure AD, OneLogin, and PingFederate automatically using the System for Cross-domain Identity Management (SCIM) standard. For other IdPs, you can provision users from your IdP using the IAM Identity Center console. See the IAM Identity Center User Guide to learn more.

Q: Can I automate identity synchronization into IAM Identity Center?

Yes. If you use Okta Universal Directory, Azure AD, OneLogin, or PingFederate, you can use SCIM to synchronize user and group information from your IdP to IAM Identity Center automatically. See the IAM Identity Center User Guide to learn more.

Q: How do I connect IAM Identity Center to my Microsoft Active Directory?

You can connect IAM Identity Center to your on-premises Active Directory (AD) or to an AWS Managed Microsoft AD directory using AWS Directory Service. See the IAM Identity Center User Guide to learn more.

Q: I manage my users and groups in Active Directory on-premises. How can I leverage these users and groups in IAM Identity Center?

You have two options for connecting Active Directory–hosted on-premises to IAM Identity Center: (1) use AD Connector, or (2) use an AWS Managed Microsoft AD trust relationship. AD Connector simply connects your existing on-premises Active Directory to AWS. AD Connector is a directory gateway with which you can redirect directory requests to your on-premises Microsoft Active Directory without caching any information in the cloud. To connect an on-premises directory using AD Connector, see the AWS Directory Service Administration Guide. AWS Managed Microsoft AD makes it easy to set up and run Microsoft Active Directory in AWS. It can be used to set up a forest trust relationship between your on-premises directory and AWS Managed Microsoft AD. To set up a trust relationship, see the AWS Directory Service Administration Guide.

Q: Can I use my Amazon Cognito User Pools as the identity source in IAM Identity Center?

Amazon Cognito is a service that helps you manage identities for your customer facing applications; it is not a supported identity source in IAM Identity Center. You can create and manage your workforce identities in IAM Identity Center or in your external identity source including Microsoft Active Directory, Okta Universal Directory, Azure Active Directory (Azure AD), or another supported IdP.

Q: Does IAM Identity Center support the browser command line and mobile interfaces?

Yes, you can use IAM Identity Center to control access to the AWS Management Console and CLI v2. IAM Identity Center enables your users to access the CLI and AWS Management Console through a single sign-on experience. The AWS Mobile Console app also supports IAM Identity Center so you get a consistent sign-in experience across browser, mobile, and command line interfaces.

Q: Which cloud applications can I connect to IAM Identity Center?

You can connect the following applications to IAM Identity Center:

1. IAM Identity Center-integrated applications: IAM Identity Center-integrated applications such as SageMaker Studio and IoT SiteWise use IAM Identity Center for authentication and work with the identities you have in IAM Identity Center. There is no need for additional configuration to synchronize identities into these applications or to set up federation to separately.
2. Pre-integrated SAML applications: IAM Identity Center comes pre-integrated with commonly used business applications. For a comprehensive list, see the IAM Identity Center console.
3. Custom SAML applications: IAM Identity Center supports applications that allow identity federation using SAML 2.0. You can enable IAM Identity Center to support these applications by using the custom application wizard.

Single sign-on access to AWS accounts

Q: Which AWS accounts can I connect to IAM Identity Center?

You can add any AWS account managed using AWS Organizations to IAM Identity Center. You need to enable all features in your organizations to manage your accounts single sign-on.

Q: How do I setup single sign-on to AWS accounts in an organizational unit (OU) within my organization?

You can pick accounts within the organization or filter accounts by OU.

Q: How do I control what permissions my users get when they use IAM Identity Center to access their accounts?

When granting access to your users, you can limit the users' permissions by picking a permission set. Permission sets are a collection of permissions that you can create in IAM Identity Center, modelling them based on AWS managed policies for job functions or any AWS managed policies. AWS managed policies for job functions are designed to closely align to common job functions in the IT industry. If required, you can also fully customize the permission set to meet your security requirements. IAM Identity Center applies these permissions to the selected accounts automatically. As you change the permission sets, IAM Identity Center enables you to apply the changes to the relevant accounts easily. When your users access the accounts through the AWS access portal, these permissions restrict what they can do within

those accounts. You can also grant multiple permission sets to your users. When they access the account through the user portal, they can pick which permission set they want to assume for that session.

Q: How do I automate permissions management across multiple accounts?

IAM Identity Center provides APIs and AWS CloudFormation support to automate permissions management in multi-account environments, and retrieve the permissions programmatically for audit and governance purposes.

Q: How do I select which user attributes to use for ABAC?

To implement ABAC, you can select attributes from the IAM Identity Center's identity store for IAM Identity Center users and users synchronized from Microsoft AD or external SAML 2.0 IdPs including Okta Universal Directory, Azure AD, OneLogin, or PingFederate. When using an IdP as your identity source, you can optionally send the attributes as a part of a SAML 2.0 assertion.

Q: For which AWS accounts can I get AWS CLI credentials?

You can get AWS CLI credentials for any AWS account and user permissions that your IAM Identity Center administrator has assigned to you. These CLI credentials can be used for programmatic access to the AWS account.

Q: How long are the AWS CLI credentials from the AWS access portal valid?

AWS CLI Credentials fetched through IAM Identity Center are valid for 60 minutes. You can get a fresh set of credentials as often as needed.

Single sign-on access to business applications

Q: How do I set up IAM Identity Center to business applications, such as Salesforce?

From the IAM Identity Center console, navigate to the applications pane, choose Configure new application, and choose an application from the list of cloud applications that are preintegrated with IAM Identity Center. Follow the on-screen instructions to configure the application. Your application is now configured and you may assign access to it. Choose the groups or users that you want to provide with access to the application and Choose Assign Access to complete the process.

Q: My company uses business applications that are not in IAM Identity Center's preintegrated application list. Can I still use IAM Identity Center?

Yes. If your application supports SAML 2.0, you can configure your application as a custom SAML 2.0 application. From the IAM Identity Center console, navigate to the applications pane, choose Configure new application, and choose Custom SAML 2.0 application. Follow the instructions to configure the application. Your application is now configured and you may assign access to it. Choose the groups or users that you want to provide with access to the application, and choose Assign Access to complete the process.

Q: My application supports OpenID Connect (OIDC) only. Can I use it with IAM Identity Center?

No. IAM Identity Center supports only SAML 2.0–based applications.

Q: Does IAM Identity Center support single sign-on to native mobile and desktop applications?

No. IAM Identity Center supports single sign-on to business applications through web browsers only.

Miscellaneous

Q: What data will IAM Identity Center store on my behalf?

IAM Identity Center will store data about which AWS accounts and cloud applications are assigned to which users and groups, as well as what permissions have been granted for accessing AWS accounts. IAM Identity Center will also create and manage IAM roles in individual AWS accounts for each permission set you grant access for your users.

Q: What multi-factor authentication (MFA) capabilities can I use with IAM Identity Center?

With IAM Identity Center, you can enable standard-based strong authentication capabilities for all your users across all identity sources. If you use a supported SAML 2.0 IdP as your identity source, you can enable multi-factor authentication capabilities of your provider. When using IAM Identity Center or Active Directory as your identity source, IAM Identity Center supports the Web Authentication specification to help you secure user access to AWS accounts and business applications with FIDO-enabled security keys, such as YubiKey, and built-in biometric authenticators, such as Touch ID on Apple MacBooks and facial recognition on PCs. You can also enable one-time-passwords (TOTPs) using authenticator apps such as Google Authenticator or Twilio Authy.

You can also use your existing Remote Authentication Dial-In User Service (RADIUS) MFA configuration with IAM Identity Center and AWS Directory Services to authenticate your users as a secondary form of verification. To learn more about configuring MFA with IAM Identity Center, visit the IAM Identity Center User Guide.

Q: Does IAM Identity Center support the Web Authentication specification?

Yes. For user identities in IAM Identity Center's identity store and Active Directory, IAM Identity Center supports the Web Authentication (WebAuthn) specification to help you secure user access to AWS accounts and business applications with FIDO-enabled security keys, such as YubiKey, and built-in biometric authenticators, such as Touch ID on Apple MacBooks and facial recognition on PCs. You can also enable one-time-passwords (TOTPs) using authenticator apps such as Google Authenticator or Twilio Authy.

Q: How do my employees get started using IAM Identity Center?

Employees can get started with IAM Identity Center by visiting the access portal that is generated when you configure your identity source in IAM Identity Center. If you manage your users in IAM Identity Center, your employees can use their email address and password they configured with IAM Identity Center to sign into the user portal. If you connect IAM Identity Center to a Microsoft Active Directory or a SAML 2.0 identity provider, your employees can sign in to user portal with their existing corporate credentials and then view the accounts and applications assigned to them. To access an account or application, employees choose the associated icon from the access portal.

Q: Is there an API available for IAM Identity Center?

Yes. IAM Identity Center provides account assignment APIs to help you automate permissions

management in multi-account environments, and retrieve the permissions programmatically for audit and governance purposes.

Amazon Cognito

Implement secure, frictionless customer identity and access management that scales.

How it works

With Amazon Cognito, you can add user sign-up and sign-in features and control access to your web and mobile applications. Amazon Cognito provides an identity store that scales to millions of users, supports social and enterprise identity federation, and offers advanced security features to protect your consumers and business. Built on open identity standards, Amazon Cognito supports various compliance regulations and integrates with frontend and backend development resources.

Use cases

Engage customers with flexible authentication
Allow customers to sign in directly, or through social or enterprise identity providers, to a hosted UI with your branding.

Manage B2B identities
Use a variety of multi-tenancy options that provide different levels of policy and tenant isolation for your business.

Secure machine-to-machine authentication
Develop modern, secure, microservice-based applications, and more easily connect your application to backend resources and web services.

Get role-based access to AWS resources
Gain secure, role-based access to AWS services, such as Amazon S3, Amazon DynamoDB, and AWS Lambda.

What is Amazon Cognito?

Amazon Cognito provides authentication, authorization, and user management for your web and mobile apps. Your users can sign in directly with a user name and password, or through a third party such as Facebook, Amazon, Google or Apple.

The two main components of Amazon Cognito are user pools and identity pools. User pools are user directories that provide sign-up and sign-in options for your app users. Identity pools enable you to grant your users access to other AWS services. You can use identity pools and user pools separately or together.

An Amazon Cognito user pool and identity pool used together

See the diagram for a common Amazon Cognito scenario. Here the goal is to authenticate your user, and then grant your user access to another AWS service.

1. In the first step your app user signs in through a user pool and receives user pool tokens after a successful authentication.
2. Next, your app exchanges the user pool tokens for AWS credentials through an identity pool.
3. Finally, your app user can then use those AWS credentials to access other AWS services such as Amazon S3 or DynamoDB.

For more examples using identity pools and user pools, see Common Amazon Cognito scenarios.

Amazon Cognito is compliant with SOC 1-3, PCI DSS, ISO 27001, and is HIPAA-BAA eligible. For more information, see AWS services in scope. See also Regional data considerations.

Features of Amazon Cognito

User pools

A user pool is a user directory in Amazon Cognito. With a user pool, your users can sign in to your web or mobile app through Amazon Cognito, or federate through a third-party identity provider (IdP). Whether your users sign in directly or through a third party, all members of the user pool have a directory profile that you can access through an SDK.

User pools provide:

- Sign-up and sign-in services.
- A built-in, customizable web UI to sign in users.
- Social sign-in with Facebook, Google, Login with Amazon, and Sign in with Apple, and through SAML and OIDC identity providers from your user pool.
- User directory management and user profiles.
- Security features such as multi-factor authentication (MFA), checks for compromised credentials, account takeover protection, and phone and email verification.
- Customized workflows and user migration through AWS Lambda triggers.

For more information about user pools, see Getting started with user pools and the Amazon Cognito user pools API reference.

Identity pools

With an identity pool, your users can obtain temporary AWS credentials to access AWS services, such as Amazon S3 and DynamoDB. Identity pools support anonymous guest users, as well as the following identity providers that you can use to authenticate users for identity pools:

- Amazon Cognito user pools
- Social sign-in with Facebook, Google, Login with Amazon, and Sign in with Apple
- OpenID Connect (OIDC) providers
- SAML identity providers
- Developer authenticated identities

To save user profile information, your identity pool needs to be integrated with a user pool.

Amazon Cognito FAQs

General

Q: What is Amazon Cognito?
Amazon Cognito lets you easily add user sign-up and authentication to your mobile and web apps. Amazon Cognito also enables you to authenticate users through an external identity provider and provides temporary security credentials to access your app's backend resources in AWS or any service behind Amazon API Gateway. Amazon Cognito works with external identity providers that support SAML or OpenID Connect, social identity providers (such as Facebook, Twitter, Amazon) and you can also integrate your own identity provider.

In addition, Amazon Cognito enables you to synchronize data across a user's devices so that their app experience remains consistent when they switch between devices or upgrade to a new device. Your app can save data locally on users' devices allowing your applications to work even when the devices are offline and then automatically synchronize the data when the device is back online.

With Amazon Cognito, you can focus on creating great app experiences instead of worrying about building, securing, and scaling a solution to handle user management, authentication, and sync across platforms and devices.

Q: Who should use Amazon Cognito?
Amazon Cognito is designed for developers who want to add user management and sync functionality to their mobile and web apps. Developers can use Cognito Identity to add sign-up and sign-in to their apps and to enable their users to securely access their app's resources. Cognito also enables developers to sync data across devices, platforms, and applications.

Q: How do I start using Amazon Cognito?
You can easily get started by visiting the AWS Console. If you do not have an Amazon Web Services account, you can create an account when you sign in to the console. Once you have created a user pool for user management or an identity pool for federated identities or sync operations, you can download and integrate the AWS Mobile SDK with your app. Alternatively you can call the Cognito server-side APIs directly, instead of using the SDK. See our developer guide for more information.

Q: Does Amazon Cognito expose server-side APIs?
Yes. Cognito exposes server-side APIs. You can create your own custom interface to Cognito by calling these APIs directly. The server-side APIs are described in the Developer Guide.

Q: Which platforms does Amazon Cognito support?
Support for Cognito is included in the optional AWS Mobile SDK, which is available for iOS, Android, Unity, and Kindle Fire. Cognito is also available in the AWS SDK for JavaScript. Cognito Your User Pools is currently supported in the AWS Mobile SDKs for iOS and Android and in the JavaScript AWS SDK for Cognito. Visit our resource page to download the SDKs.

Q: Do I have to use the AWS Mobile SDK?
No. Cognito exposes its control and data APIs as web services. You can implement your own client library calling the server-side APIs directly.

Add User Sign-up & Sign-in to your mobile and web apps

Q: Can I have my own identity provider to support user sign-up and sign-in?

Yes, you can easily and securely add sign-up and sign-in functionality to your apps with Cognito Identity. Your users can sign-up and sign-in using email, phone number, or user name. You can also implement enhanced security features, such as email verification, phone number verification, and multi-factor authentication. Cognito Identity also enables you to customize workflows by, for example, adding app-specific logic to user registration for fraud detection and user validation through AWS Lambda. To learn more, visit our docs.

Q: What is a User Pool?

A User Pool is your user directory that you can configure for your web and mobile apps. A User Pool securely stores your users' profile attributes. You can create and manage a User Pool using the AWS console, AWS CLI, or AWS SDK.

Q: What user profile information is supported by Cognito Identity?

Developers can use either standard OpenID Connect-based user profile attributes (such as user name, phone number, address, time zone, etc.) or customize to add app-specific user attributes.

Q: Can I enable my application's users to sign up or sign in with an email address or phone number?

Yes, you can use the aliasing feature to enable your users to sign up or sign in with an email address and a password or a phone number and a password. To learn more, visit our docs.

Q: Can I set up password policies?

Yes, you can set up password policies, such as strength of password and character type requirements, when setting up or configuring your user pool.

Q: Can I verify the email addresses and phone numbers of my application's users?

Yes, with Cognito Identity you can require your users' email addresses and phone numbers to be verified prior to providing them access to your application. During sign-up, a verification code will be sent to the user's phone number or email address, and the user must input the verification code to complete sign-up and become confirmed.

Q: Does Cognito Identity support SMS-based multi-factor authentication (MFA)?

Yes, you can enable the end users of your application to sign in with SMS-based MFA. With SMS-based MFA enabled, your users will be prompted for their password (the first factor—what they know), and for a security code that can only be received on their mobile phone via SMS (the second factor—what they have).

Q: Is it possible to customize user sign-up and sign-in workflows?

Yes, you can customize sign-up and sign-in by adding app-specific logic to the user sign-up and sign-in flows using AWS Lambda. For example, you can create AWS Lambda functions to identify fraud or perform additional validations on user data. You are able to trigger developer-provided Lambda functions at pre-registration, at post-confirmation, at pre-authentication, during authentication to customize the challenges, and at post-authentication. You can also use Lambda functions to customize messages sent as part of email or phone number verification and multi-factor authentication.

Q: Can I remember the devices associated with my application's users in a Cognito user pool?

Yes, you can opt to remember devices used to access your application, and you associate these remembered devices with your application's users in a Cognito user pool. You can also opt to use remembered devices to supress second factor challenges for your users when you have set up multi-factor authentication.

Q: How can I migrate my existing users into an Amazon Cognito user pool?

You can use our import tool to migrate your existing users into an Amazon Cognito user pool. User attribute values are imported from a .csv file, which can be uploaded through the console, our APIs, or CLI. When imported users first sign in, they confirm their account and create a new password with a code sent to their email address or phone. There is no additional cost for using the import tool. To learn more, see the import tool documentation.

The import tool does not migrate passwords. If you want to retain your users' current passwords, you might consider an alternative approach to migrate users one at a time as they sign-in to your app during a transition period. With this approach, your app first tries to sign-in the user with your Cognito user pool. If that user doesn't exist in the user pool, your app will sign the user in with your existing identity system and temporarily retain the username and

password used to do so. After a user successfully signs in with your existing identity system, your app would then use the same username and password to create the user in your Cognito user pool. This approach requires maintaining your existing identity system during the transition period, but after the transition period ends, you can use our import tool to import the remaining users (without their passwords).

Federate identities and provide secure access to AWS resources

Q: Can I use Cognito Identity to federate identities and secure access to AWS resources?
Yes, Cognito Identity enables you to authenticate users through an external identity provider and provides temporary security credentials to access your app's backend resources in AWS or any service behind Amazon API Gateway. Amazon Cognito works with external identity providers that support SAML or OpenID Connect, social identity providers (such as Facebook, Twitter, Amazon) and you can also integrate your own identity provider.

Q: Which public identity providers can I use with Amazon Cognito Identity?
You can use Amazon, Facebook, Twitter, Google and any other OpenID Connect compatible identity provider.

Q: What is an Identity Pool?
Identity pools are the containers that Cognito Identity uses to keep your apps' federated identities organized. Identity Pool associates federated identities from social identity providers with a unique user specific identifier. Identity Pools do not store any user profiles. An identity pool can be associated with one or many apps. If you use two different identity pools for two apps then the same end user will have a different unique identifier in each Identity Pool.

Q: How does the login flow work with public identity providers?
Your mobile app authenticates with an Identity Provider (IdP) using the provider's SDK. Once the end user is authenticated with the IdP, the OAuth or OpenID Connect token or the SAML assertion returned from the IdP is passed by your app to Cognito Identity, which returns a new Cognito ID for the user and a set of temporary, limited-privilege AWS credentials.

Q: Can I register and authenticate my own users?
Cognito Identity can integrate with your existing authentication system. With a simple API call you can retrieve a Cognito ID for your end users based on your own unique identifier for your users. Once you have retrieved the Cognito ID and OpenID Token Cognito Identity provides, you can use the Cognito Identity client SDK to access AWS resources and synchronize user data. Cognito Identity is a fully managed identity provider to make it easier for you to implement user sign-up and sign-in for your mobile and web apps.

Q: How does Cognito Identity help me control permissions and access AWS services securely?
Cognito Identity assigns your users a set of temporary, limited privilege credentials to access your AWS resources so you do not have to use your AWS account credentials. The permissions for each user are controlled through AWS IAM roles that you create. You can define rules to choose the IAM role for each user, or if you are using groups in a Cognito user pool, you can assign IAM roles based on groups. Cognito Identity also allows you to define a separate IAM role with limited permissions for guest users who are not authenticated. In addition, you can use the unique identifier that Cognito generates for your users to control access to specific resources. For example you can create a policy for an S3 bucket that only allows each user access to their own folder within the bucket.

Q: When using public identity providers, does Amazon Cognito Identity store users' credentials?

No, your app communicates directly with the supported public identity provider (Amazon, Facebook, Twitter, Google, or an Open ID Connect-compliant provider) to authenticate users. Cognito Identity does not receive or store user credentials. Cognito Identity uses the token from the identity provider to obtain a unique identifier for the user and then hashes it using a one-way

hash so that the same user can be recognized again in the future without storing the actual user identifier.

Q: Does Cognito Identity receive or store confidential information about my users from the identity providers?
No. Cognito Identity does not receive any confidential information (such as email address, friends list, etc.) from the identity providers.

Q: Do I still need my own backend authentication systems with Cognito Identity?
No. Cognito Identity supports login through Amazon, Facebook, Twitter, and Google, as well as providing support for unauthenticated users. With Cognito Identity you can support federated authentication, profile data sync store and AWS access token distribution without writing any backend code.

Q: What if I don't want to force my users to log in?
Cognito Identity supports the creation and token vending process for unauthenticated users as well as authenticated users. This removes the friction of an additional login screen in your app, but still enables you to use temporary, limited privilege credentials to access AWS resources.

Q: What are unauthenticated users?
Unauthenticated users are users who do not authenticate with any identity provider, but instead access your app as a guest. You can define a separate IAM role for these users to provide limited permissions to access your backend resources.

Q: Does Cognito Identity support separate identities for different users on the same device?
Yes. Cognito Identity supports separate identities on a single device, such as a family iPad. Each identity is treated separately and you have complete control over how your app logs users in and out and how local and remote app data is stored.

Q: How do I store data associated with Cognito Identity?
You can programmatically create a data set associated with Cognito Identity and start saving data in the form of key/value pairs. The data is stored both locally on the device and in the Cognito sync store. Cognito can also sync this data across all of the end user's devices.

Q: Does the number of identities in the Cognito Identity console tell me how many users are using my app?
The number of identities in the Cognito Identity console shows you how many identities were created via the Cognito Identity APIs. For Authenticated Identities (those logging in with a login provider such as Facebook or an OpenID Connect provider), each call to Cognito Identity's GetId API will only ever create a single identity for each user. However, for Unauthenticated identities, each time the client in an app calls the GetId API will generate a new identity. Therefore, if your app calls GetId for unauthenticated identities multiple times for a single user it will appear that a single user has multiple identities. So it is important that you cache the response from GetId when using unauthenticated identities and not call it multiple times per user.

The Mobile SDK provides the logic to cache the Cognito Identity automatically so you don't have to worry about this. If you're looking for a complete analytics solution for your app, including the ability to track unique users, please look at Amazon Mobile Analytics.

Q: Why should I use attributes for access control in Amazon Cognito?
Amazon Cognito Identity Pools now enables you to use attributes from social and corporate identity providers to make access control decisions. You can create a simple permissions policy for users with the same attribute instead of creating multiple policies for each user. This will simplify your permissions management to AWS resources.

Store and sync data across devices

Q: What is the Amazon Cognito sync store?
The Amazon Cognito Sync store is a key/value pair store linked to an Amazon Cognito identity. There is no limit to the number of identities you can create in your identity pools and sync store. Each Amazon Cognito identity within the sync store has its own user information store.

Q: Is data saved directly to the Amazon Cognito sync store?
No. The optional AWS Mobile SDK saves your data to an SQLite database on the local device, this way the data is always accessible to your app. The data is pushed to the Amazon Cognito sync store by calling the synchronize() method and, if push synchronization is enabled, all other devices linked to an identity are notified of the data change in the sync store via Amazon SNS.

Q: How is data stored in the Amazon Cognito sync store?
Data associated with an Amazon Cognito identity are organized as key/value pairs. A key is a label e.g. “MusicVolume”, and a value e.g. “11”. Key/value pairs are grouped and categorized using data sets. Data sets are a logical partition of key/value pairs and the most granular entity used by Amazon Cognito to perform sync operations.

Q: What is the maximum size of a user information store within the Amazon Cognito sync store?
Each user information store can have a maximum size of 20MB. Each data set within the user information store can contain up to 1MB of data. Within a data set you can have up to 1024 keys.

Q: What kind of data can I store in a data set?
Both keys and values within a data set are alphanumeric strings. There is no limit to the length of the strings other than the total amount of values in a dataset cannot exceed 1MB. Binary data can be stored as a base64 encoded string as a value provided it does not exceed the 1MB limit.

Q: Why are data sets limited to 1MB?
Limiting the data set size to 1MB increases the chances of a synchronization task completing successfully even when bandwidth is limited without lots of retries that consume battery life and data plans.

Q: Are user identities and user information stores shared across developers?
No, a user identity and information store is tied to a specific AWS account. If there are multiple apps from different publishers on a particular device that use Amazon Cognito, each app will use the information store created by each publisher.

Q: How can I analyze and query the data stored in the Cognito Sync store?
With Cognito Streams, you can push sync store data to a Kinesis stream in your AWS account. You can then consume this stream and store the data in a way that makes it easy for you to analyze such as a Amazon Redshift database, an RDS instance you own or even an S3 file. We have published sample Kinesis consumer application to show how to store the updates data in Amazon Redshift.

Q: Why should I use Kinesis stream instead of a database export?
By streaming the data to Kinesis you can receive all of the history of changes to your datasets in real-time. This means you receive all the changes an end user makes to a dataset and gives you the flexibility to store this data in a tool of your choice.

Q: What if I already have data stored in Cognito?
When you enable the Kinesis stream feature you will be able to start a bulk publish. This process asynchronously sends all of the data currently stored in your Cognito sync store to the Kinesis stream you selected.

Q: What is the price of this feature?
Cognito pushes the data to a Kinesis stream you own. There is no difference in Cognito's per-synchronization price if this feature is enabled. You will be charged Kinesis' standard rates for your shards.

Q: Can I validate data before it is saved?
Amazon Cognito Events allows developers to run an AWS Lambda function in response to important events in Cognito. The Sync Trigger event is an event that occurs when any dataset is synchronized. Developers can write an AWS Lambda function to intercept the synchronization event. The function can evaluate the changes to the underlying Dataset and manipulate the data before it is stored in the cloud and synchronized back to the user's other devices. Alternatively, the AWS Lambda function could fail the sync operation so that the data is not synchronized to the user's other devices.

Q: How is data synchronized with Amazon Cognito?
You can programmatically trigger the sync of data sets between client devices and the Amazon Cognito sync store by using the synchronize() method in the AWS Mobile SDK. The synchronize() method reads the latest version of the data available in the Amazon Cognito sync store and compares it to the local, cached copy. After comparison, the synchronize() method writes the latest updates as necessary to the local data store and the Amazon Cognito sync store. By default Amazon Cognito maintains the last-written version of the data. You can override this behavior and resolve data conflicts programmatically. In addition, push synchronization allows you to use Amazon Cognito to send a silent push notification to all devices associated with an identity to notify them that new data is available.

Q: What is a silent push notification?
Amazon Cognito uses the Amazon Simple Notification Service (SNS) to send silent push notifications to devices. A silent push notification is a push message that is received by your application on a user's device that will not be seen by the user.

Q: How do I use push synchronization?
To enable push synchronization you need to declare a platform application using the Amazon SNS page in the AWS Management Console. Then, from the identity pool page in the Amazon Cognito page of the AWS Management Console, you can link the SNS platform application to your Cognito identity pool. Amazon Cognito automatically utilizes the SNS platform application to notify devices of changes.

Q: How are conflicts in the synchronization process handled?
By default Amazon Cognito maintains the last-written version of the data. You can override this behavior by choosing to respond to a callback from the AWS Mobile SDK which will contain both versions of the data. Your app can then decide which version of the data (the local one or the one in the Amazon Cognito sync store) to keep and save to the Amazon Cognito sync store.

Amazon Simple Notification Service (Amazon SNS) and Amazon Simple Queue Service (Amazon SQS)

Amazon Simple Notification Service (Amazon SNS)

Fully managed Pub/Sub service for A2A and A2P messaging.

Amazon Simple Queue Service (Amazon SQS)

Fully managed message queuing for microservices, distributed systems, and serverless applications.

Amazon Simple Notification Service (Amazon SNS)

Amazon Simple Notification Service (Amazon SNS) is a web service that enables applications, end-users, and devices to instantly send and receive notifications from the cloud.

How it works

Amazon Simple Notification Service (SNS) sends notifications two ways, A2A and A2P. A2A provides high-throughput, push-based, many-to-many messaging between distributed systems, microservices, and event-driven serverless applications. These applications include Amazon Simple Queue Service (SQS), Amazon Kinesis Data Firehose, AWS Lambda, and other HTTPS endpoints. A2P functionality lets you send messages to your customers with SMS texts, push notifications, and email.

Pub/Sub

SMS

Mobile push

Use cases

Integrate your applications with FIFO messaging
Deliver messages in a strictly ordered, first in, first out (FIFO) manner to maintain accuracy and consistency across independent applications.

Securely encrypt notification message delivery
Encrypt messages with AWS Key Management Service (KMS), ensure traffic privacy with AWS PrivateLink, and control access with resource policies and tags.

Capture and fan out events from over 60 AWS services
Fan out events across AWS categories, such as analytics, compute, containers, databases, IoT, machine learning (ML), security, and storage.

Send SMS texts to customers across over 240 countries
Use worldwide SMS, with redundancy across providers. Set SMS origination identity with a sender ID, long code, short code, TFN, or 10DLC.

Amazon SNS features

Topic types

Standard Topics
Standard topics can be used in many scenarios, as long as your application can process messages that arrive more than once and out of order, for example: fanning out messages to media encoding, fraud detection, tax calculation, search index, and critical alerting applications.

Maximum throughput: Standard topics support a nearly unlimited number of messages per second.

Best-effort ordering: Occasionally, messages might be delivered in an order different from which they were published.

Best-effort deduplication: A message is delivered at least once, but occasionally more than one copy of a message is delivered.

Multiple subscription types: Messages can be delivered to application-to-application (A2A) endpoints (Amazon SQS, Amazon Kinesis Data Firehose, AWS Lambda, HTTPS) as well as application-to-person (A2P) endpoints (SMS, mobile push, and email).

Message fanout: Each account can support 100,000 Standard topics and each topic supports up to 12.5M subscriptions.

FIFO Topics
FIFO topics are designed to enhance messaging between applications when the order of operations and events is critical, or where duplicates can't be tolerated, for example: fanning out messages to bank transaction logging, stock monitoring, flight tracking, inventory management, and price update applications.

High throughput: FIFO topics support up to 300 messages per second or 10 MB per second per FIFO topic (whichever comes first).

Strict ordering: The order in which messages are published and delivered is strictly preserved (i.e. first-in-first-out).

Strict deduplication: Duplicate messages aren't delivered. Deduplication happens within a 5-minute interval, from the message publish time.

SQS FIFO subscriptions: Messages can be delivered to Amazon SQS FIFO queues.

Message fanout: Each account can support 1,000 FIFO topics and each topic supports up to 100 subscriptions.

Event sources and destinations

Event-driven computing is a model in which subscriber services automatically perform work in response to events triggered by publisher services. This paradigm can be applied to automate workflows while decoupling the services that collectively and independently work to fulfill these workflows. Amazon SNS is an event-driven hub that has native integration with a wide variety of AWS event sources and event destinations.

Message publishing and batching

Message publishing enables you to send data, in the form of messages, to an Amazon SNS topic which delivers the messages asynchronously to the applications that are subscribed to the topic. You can publish from 1 to 10 messages per API request. You may choose to batch messages together to reduce your Amazon SNS costs. Each message can contain up to 256KB of data. If your use case requires larger data payloads, the Amazon SNS Extended Client Library stores the payload (up to 2GB) in an Amazon S3 bucket and publishes the reference of the stored Amazon S3 object to the Amazon SNS topic.

Message filtering

Message filtering empowers your subscriber applications to create filter policies, so that these applications can receive only the notifications that they are interested in, as opposed to receiving every message published to the topic. This enables you to simplify your architecture, offloading the message filtering logic from subscriber applications as well as the message routing logic from publisher applications.

Message fanout and delivery

When you publish a message to a topic, Amazon SNS replicates and delivers the message to applications subscribed to the topic. Amazon SNS supports application-to-application (A2A) and application-to-person (A2P) message delivery. Amazon SNS also supports cross-region

and cross-account message delivery, in addition to message delivery status logging with Amazon CloudWatch.

Message durability

Amazon SNS uses a number of mechanisms that work together to provide message durability. To start, published messages are stored across multiple, geographically-separated servers and data centers. If a subscribed endpoint isn't available, Amazon SNS executes a message delivery retry policy. To preserve any messages that aren't delivered before the delivery retry policy ends, you can use a dead-letter queue powered by Amazon SQS. Moreover, you can subscribe Amazon Kinesis Data Firehose delivery streams to Amazon SNS topics, which allows messages to be sent to durable endpoints such as Amazon S3 buckets or Amazon Redshift tables.

Message encryption

Amazon SNS provides encrypted topics to protect your messages from unauthorized and anonymous access. When you publish messages to encrypted topics, Amazon SNS immediately encrypts your messages. The encryption takes place on the server, using a 256-bit AES-GCM algorithm and a Customer Master Key (CMK) issued with AWS Key Management Service (KMS). The messages are stored in encrypted form, and decrypted as they are delivered to subscribing endpoints, such as Amazon SQS queues, Amazon Kinesis Data Firehose streams, AWS Lambda functions, HTTP/S endpoints, phone numbers, mobile apps, and email addresses.

Message privacy

Amazon SNS supports VPC Endpoints (VPCE) via AWS PrivateLink. You can use VPC Endpoints to privately publish messages to Amazon SNS topics, from an Amazon Virtual Private Cloud (VPC), without traversing the public internet. This feature brings additional security, helps promote data privacy, and aligns with assurance programs. When you use AWS PrivateLink, you don’t need to set up an Internet Gateway (IGW), Network Address Translation (NAT) device, or Virtual Private Network (VPN) connection. You don’t need to use public IP addresses, either.

Message data protection

Amazon SNS message data protection empowers topic owners to define data protection policies that can discover and protect sensitive data that is transmitted via their topics. This can help you to simplify your architecture by offloading data protection logic from your applications, while helping support your compliance objectives, for example, with regulations such as the Health Insurance Portability and Accountability Act (HIPAA), General Data Privacy Regulation (GDPR), Payment Card Industry Compliance (PCI), and Federal Risk and Authorization Management Program (FedRAMP). You can audit messages that are inbound to a topic to determine how much sensitive data they contain, prevent them from being delivered to downstream subscribers via blocking, or de-identify specific data in the payload via redaction or masking.

Message archiving and analytics

Amazon SNS provides a direct connection to Amazon Kinesis Data Firehose, allowing message storage in services such as Amazon S3, Amazon Redshift, Amazon OpenSearch Service, and MongoDB. This feature also enables message storage in analytics services, such as Datadog, New Relic, and Splunk.

SMS text messages

Amazon SNS supports the ability to send SMS text messages at scale to 200+ countries, using a highly available and durable service, with redundancy across multiple SMS providers. With Amazon SNS, you are be able to control your originating identity by using a sender ID, long

codes, or short codes. Moreover, you can use the Amazon SNS sandbox to validate your SMS workloads before moving them to production.

Mobile push notifications

Amazon SNS mobile notifications make it simple and cost effective to fan out mobile push notifications to iOS, Android, Fire, Windows, and Baidu devices. Mobile notifications can be triggered from user-driven actions or business logic. Amazon SNS delivers mobile push notifications through Amazon Device Messaging (ADM), Apple Push Notification Service (APNs), Baidu Cloud Push (Baidu), Firebase Cloud Messaging (FCM), Microsoft Push Notification Service for Windows Phone (MPNS), and Windows Push Notification Services (WNS).

Email notifications

Amazon SNS supports the delivery of notifications to email addresses subscribed to topics. This feature supports a variety of use cases. For example, you can use Amazon SNS to receive application alerts, as email notifications, to bring visibility into your DevOps workflows. Thus, you can be notified immediately when an event occurs, such as a specific change to your Amazon EC2 Auto Scaling group, or a new file uploaded to your Amazon S3 bucket, or a metric threshold breached in Amazon CloudWatch.

What is Amazon SNS?

Amazon Simple Notification Service (Amazon SNS) is a managed service that provides message delivery from publishers to subscribers (also known as *producers* and *consumers*). Publishers communicate asynchronously with subscribers by sending messages to a *topic*, which is a logical access point and communication channel. Clients can subscribe to the SNS topic and receive published messages using a supported endpoint type, such as Amazon Kinesis Data Firehose, Amazon SQS, AWS Lambda, HTTP, email, mobile push notifications, and mobile text messages (SMS).

Amazon Simple Queue Service (Amazon SQS)

How it works

Amazon Simple Queue Service (SQS) lets you send, store, and receive messages between software components at any volume, without losing messages or requiring other services to be available.

Why Amazon SQS?

First-in-first-out (FIFO) queues help make sure the messages you send to systems are published in the correct order.

Use cases

Increase application reliability and scale
Amazon SQS provides a simple and reliable way for customers to decouple and connect components (microservices) together using queues.

Decouple microservices and process event-driven applications
Separate frontend from backend systems, such as in a banking application. Customers immediately get a response, but the bill payments are processed in the background.

Ensure work is completed cost-effectively and on time
Place work in a single queue where multiple workers in an autoscale group scale up and down based on workload and latency requirements.

Maintain message ordering with deduplication
Process messages at high scale while maintaining the message order, allowing you to deduplicate messages.

Amazon SQS features

Queue types

Amazon SQS offers two queue types for different application requirements:

Standard Queues

Unlimited Throughput: Standard queues support a nearly unlimited number of transactions per second (TPS) per API action.

At-Least-Once Delivery: A message is delivered at least once, but occasionally more than one copy of a message is delivered.

Best-Effort Ordering: Occasionally, messages might be delivered in an order different from which they were sent.

You can use standard message queues in many scenarios, as long as your application can process messages that arrive more than once and out of order, for example:

- Decouple live user requests from intensive background work: Let users upload media while resizing or encoding it.
- Allocate tasks to multiple worker nodes: Process a high number of credit card validation requests.
- Batch messages for future processing: Schedule multiple entries to be added to a database.

FIFO Queues

High Throughput: By default, FIFO queues support up to 300 messages per second (300 send, receive, or delete operations per second). When you batch 10 messages per operation (maximum), FIFO queues can support up to 3,000 messages per second. If you require higher throughput, you can enable high throughput mode for FIFO on the Amazon SQS console, which will support up to 30,000 messages per second with batching, or up to 3,000 messages per second without batching.

Exactly-Once Processing: A message is delivered once and remains available until a consumer processes and deletes it. Duplicates aren't introduced into the queue.

First-In-First-Out Delivery: The order in which messages are sent and received is strictly preserved (i.e. First-In-First-Out).

FIFO queues are designed to enhance messaging between applications when the order of operations and events is critical, or where duplicates can't be tolerated, for example:

- Ensure that user-entered commands are executed in the right order.
- Display the correct product price by sending price modifications in the right order.

- Prevent a student from enrolling in a course before registering for an account.

Functionality

- Unlimited queues and messages: Create unlimited Amazon SQS queues with an unlimited number of messages in any Region
- Payload Size: Message payloads can contain up to 256KB of text in any format. Each 64KB 'chunk' of payload is billed as 1 request. For example, a single API call with a 256KB payload will be billed as four requests. To send messages larger than 256KB, you can use the Amazon SQS Extended Client Library for Java, which uses Amazon Simple Storage Service (S3) to store the message payload. A reference to the message payload is sent using SQS.
- Batches: Send, receive, or delete messages in batches of up to 10 messages or 256KB. Batches cost the same amount as single messages, meaning SQS can be even more cost effective for customers that use batching.
- Long polling: Reduce extraneous polling to minimize cost while receiving new messages as quickly as possible. When your queue is empty, long-poll requests wait up to 20 seconds for the next message to arrive. Long poll requests cost the same amount as regular requests.
- Retain messages in queues for up to 14 days.
- Send and read messages simultaneously.
- Message locking: When a message is received, it becomes "locked" while being processed. This keeps other computers from processing the message simultaneously. If the message processing fails, the lock will expire and the message will be available again.
- Queue sharing: Securely share Amazon SQS queues anonymously or with specific AWS accounts. Queue sharing can also be restricted by IP address and time-of-day.
- Server-side encryption (SSE): Protect the contents of messages in Amazon SQS queues using keys managed in the AWS Key Management Service (AWS KMS). SSE encrypts messages as soon as Amazon SQS receives them. The messages are stored in encrypted form and Amazon SQS decrypts messages only when they are sent to an authorized consumer.
- Dead Letter Queues (DLQ): Handle messages that a consumer has not successfully processed with dead- letter queues (DLQs). When a message's maximum receive count is exceeded, Amazon SQS moves the message to the DLQ associated with the original queue. DLQs must be of the same type as the source queue (standard or FIFO). You can inspect the messages in DLQs to understand why your consumer has not successfully received them. Once you have remediated the issues, you can move the messages from the DLQ to their respective source queues.

Using Amazon SQS with other AWS infrastructure web services

Amazon SQS message queuing can be used with other AWS services such as Amazon Redshift, Amazon DynamoDB, Amazon Relational Database Service (RDS), Amazon Elastic Compute Cloud (EC2), Amazon Elastic Container Service (ECS), AWS Lambda, and Amazon S3, to make distributed applications more scalable and reliable. Below are some common design patterns:

- Work Queues: Decouple components of a distributed application that may not all process the same amount of work simultaneously.

- Buffer and Batch Operations: Add scalability and reliability to your architecture, and smooth out temporary volume spikes without losing messages or increasing latency.
- Request Offloading: Move slow operations off of interactive request paths by enqueing the request.
- Fanout: Combine SQS with Simple Notification Service (SNS) to send identical copies of a message to multiple queues in parallel.
- Priority: Use separate queues to provide prioritization of work.
- Scalability: Because message queues decouple your processes, it's easy to scale up the send or receive rate of messages — simply add another process.
- Resiliency: When part of your system fails, it doesn't need to take the entire system down. Message queues decouple components of your system, so if a process that is reading messages from the queue fails, messages can still be added to the queue to be processed when the system recovers.

Amazon Simple Notification Service (Amazon SNS) FAQs

Overview

Q: What is Amazon Simple Notification Service (Amazon SNS)?

Amazon Simple Notification Service (Amazon SNS) is a web service that makes it easy to set up, operate, and send notifications from the cloud. It provides developers with a highly scalable, flexible, and cost-effective capability to publish messages from an application and immediately deliver them to subscribers or other applications. It is designed to make web-scale computing easier for developers. Amazon SNS follows the "publish-subscribe" (pub-sub) messaging paradigm, with notifications being delivered to clients using a "push" mechanism that eliminates the need to periodically check or "poll" for new information and updates. With simple APIs requiring minimal up-front development effort, no maintenance or management overhead and pay-as-you-go pricing, Amazon SNS gives developers an easy mechanism to incorporate a powerful notification system with their applications.

Q: How can I get started using Amazon SNS?

You can create an Amazon SNS topic and publish messages in a few steps by completing our 10-minute tutorial, Send Fanout Notifications.

For additional detail, see the Amazon SNS Developer Guide, and sample code in the Resource Center.

Q: What are the benefits of using Amazon SNS?

Amazon SNS offers several benefits making it a versatile option for building and integrating loosely-coupled, distributed applications:

- Instantaneous, push-based delivery (no polling)
- Simple APIs and easy integration with applications
- Flexible message delivery over multiple transport protocols
- Inexpensive, pay-as-you-go model with no up-front costs

- Web-based AWS Management Console offers the simplicity of a point-and-click interface

Q: What are some example uses for Amazon SNS notifications?

The Amazon SNS service can support a wide variety of needs including event notification, monitoring applications, workflow systems, time-sensitive information updates, mobile applications, and any other application that generates or consumes notifications. For example, Amazon SNS can be used in workflow systems to relay events among distributed computer applications, move data between data stores or update records in business systems. Event updates and notifications concerning validation, approval, inventory changes and shipment status are immediately delivered to relevant system components as well as end-users. A common pattern is to use SNS to publish messages to Amazon SQS message queues to reliably send messages to one or many system components asynchronously. Another example use for Amazon SNS is to relay time-critical events to mobile applications and devices. Since Amazon SNS is both highly reliable and scalable, it provides significant advantages to developers who build applications that rely on real-time events.

Q: How does Amazon SNS work?

It is very easy to get started with Amazon SNS. Developers must first create a “topic” which is an “access point” – identifying a specific subject or event type – for publishing messages and allowing clients to subscribe for notifications. Once a topic is created, the topic owner can set policies for it such as limiting who can publish messages or subscribe to notifications, or specifying which notification protocols will be supported (i.e. HTTP/HTTPS, email, SMS). Subscribers are clients interested in receiving notifications from topics of interest; they can subscribe to a topic or be subscribed by the topic owner. Subscribers specify the protocol and end-point (URL, email address, etc.) for notifications to be delivered. When publishers have information or updates to notify their subscribers about, they can publish a message to the topic – which immediately triggers Amazon SNS to deliver the message to all applicable subscribers.

Q: How is Amazon SNS different from Amazon SQS?

Amazon Simple Queue Service (SQS) and Amazon SNS are both messaging services within AWS, which provide different benefits for developers. Amazon SNS allows applications to send time-critical messages to multiple subscribers through a “push” mechanism, eliminating the need to periodically check or “poll” for updates. Amazon SQS is a message queue service used by distributed applications to exchange messages through a polling model, and can be used to decouple sending and receiving components. Amazon SQS provides flexibility for distributed components of applications to send and receive messages without requiring each component to be concurrently available.

A common pattern is to use SNS to publish messages to Amazon SQS queues to reliably send messages to one or many system components asynchronously.

Q: How is Amazon SNS different from Amazon MQ?

Amazon MQ, Amazon SQS, and Amazon SNS are messaging services that are suitable for anyone from startups to enterprises. If you're using messaging with existing applications, and want to move your messaging to the cloud quickly and easily, we recommend you consider Amazon MQ. It supports industry-standard APIs and protocols so you can switch from any standards-based message broker to Amazon MQ without rewriting the messaging code in your applications. If you are building brand new applications in the cloud, we recommend you consider Amazon SQS and Amazon SNS. Amazon SQS and SNS are lightweight, fully managed message queue and topic services that scale almost infinitely and provide simple, easy-to-use

APIs. You can use Amazon SQS and SNS to decouple and scale microservices, distributed systems, and serverless applications, and improve reliability.

Q: How can I get started using Amazon SNS?

To sign up for Amazon SNS, click the “Sign up for Amazon SNS” button on the Amazon SNS detail page. You must have an Amazon Web Services account to access this service; if you do not already have one, you will be prompted to create one when you begin the Amazon SNS sign-up process. After signing up, please refer to the Amazon SNS documentation and Getting Started Guide to begin using Amazon SNS. Using the AWS Management Console, you can easily create topics, add subscribers, send notifications, and edit topic policies – all from your browser.

Q: Is Amazon SNS supported in the AWS Management Console?

Amazon SNS is supported in the AWS Management Console which provides a point-and-click, web-based interface to access and manage Amazon SNS. Using the AWS Management Console, you can create topics, add subscribers, and send notifications – all from your browser. In addition, the AWS Management Console makes it easy to publish messages to your endpoint of choice (HTTP, SQS, Lambda, mobile push, email, or SMS) and edit topic policies to control publisher and subscriber access.

Q: What are the Amazon SNS service access points in each region?

Please refer to the AWS Regions and Endpoints section of the AWS documentation for the latest list of all Amazon SNS service access points.

Q: Can I get a history of SNS API calls made on my account for security analysis and operational troubleshooting purposes?

Yes. SNS supports AWS CloudTrail, a web service that records AWS API calls for your account and delivers log files to you. With CloudTrail, you can obtain a history of such information as the identity of the API caller, the time of the API call, the source IP address of the API caller, the request parameters, and the response elements returned by SNS.

SNS currently supports CloudTrail auditing for authenticated calls only. CloudTrail Audit logs for unauthenticated ConfirmSubscription and Unsubscribe calls are not available at this time. For more information, see the CloudTrail section of the SNS Developer Guide.

To receive a history of SNS API calls made on your account, simply turn on AWS CloudTrail in the AWS Management Console. To learn more about AWS CloudTrail, click here.

Features and functionality

Q: What is the format of an Amazon SNS topic?

Topic names are limited to 256 characters. Alphanumeric characters plus hyphens (-) and underscores (_) are allowed. Topic names must be unique within an AWS account. After you delete a topic, you can reuse the topic name. When a topic is created, Amazon SNS will assign a unique ARN (Amazon Resource Name) to the topic, which will include the service name (SNS), region, AWS ID of the user and the topic name. The ARN will be returned as part of the API call to create the topic. Whenever a publisher or subscriber needs to perform any action on the topic, they should reference the unique topic ARN.

The following is the ARN for a topic named “mytopic” created by a user with the AWS account ID “123456789012” and hosted in the US East region:

arn:aws:sns:us-east-1:1234567890123456:mytopic Note: Users should NOT attempt to build the topic ARN from its separate components – they should always use the name returned from the API call to create the topic.

Q: What are the available operations for Amazon SNS and who can perform these operations?

Amazon SNS provides a set of simple APIs to enable event notifications for topic owners, subscribers, and publishers.

Owner operations:

- CreateTopic – Create a new topic.
- DeleteTopic – Delete a previously created topic.
- ListTopics – List of topics owned by a particular user (AWS ID).
- ListSubscriptionsByTopic – List of subscriptions for a particular topic
- SetTopicAttributes – Set/modify topic attributes, including setting and modifying publisher/subscriber permissions, transports supported, etc.
- GetTopicAttributes – Get/view existing attributes of a topic
- AddPermission – Grant access to selected users for the specified actions
- RemovePermission – Remove permissions for selected users for the specified actions

Subscriber operations:

- Subscribe – Register a new subscription on a particular topic, which will generate a confirmation message from Amazon SNS
- ConfirmSubscription – Respond to a subscription confirmation message, confirming the subscription request to receive notifications from the subscribed topic
- UnSubscribe – Cancel a previously registered subscription
- ListSubscriptions – List subscriptions owned by a particular user (AWS ID)

Publisher operations:

- Publish: Publish a new message to the topic.

Q: Why are there two different APIs to list subscriptions?

The two APIs to list subscriptions perform different functions and return different results:

- The ListSubscriptionsByTopic API allows a topic owner to see the list of all subscribers actively registered to a topic.
- The ListSubscriptions API allows a user to get a list of all their active subscriptions (to one or more topics).

Q: What are the different delivery formats/transports for receiving notifications?

“SQS” – Users can specify an SQS standard or FIFO queue as the endpoint; Amazon SNS will enqueue a notification message to the specified queue (which subscribers can then process using SQS APIs such as ReceiveMessage, DeleteMessage, etc.).

- “HTTP”, “HTTPS” – Subscribers specify a URL as part of the subscription registration; notifications will be delivered through an HTTP POST to the specified URL.
- ”Email”, “Email-JSON” – Messages are sent to registered addresses as email. Email-JSON sends notifications as a JSON object, while Email sends text-based email.
- “SQS” – Users can specify an SQS standard queue as the endpoint; Amazon SNS will enqueue a notification message to the specified queue (which subscribers can then process using SQS APIs such as ReceiveMessage, DeleteMessage, etc.).
- “SMS” – Messages are sent to registered phone numbers as SMS text messages.

Q: Can topic owners control the transports that are allowed on topics they create/own?

Topic owners can configure specific transports on their topics by setting the appropriate permissions through access control policies.

Q: How does an owner set Access Control policies?

Please refer to the Amazon SNS Getting Started Guide for an overview of setting access control policies.

Q: Can a single topic support subscriptions over multiple protocols/transports?

Subscribers to an Amazon SNS topic can receive notifications on any transport supported by the topic. A topic can support subscriptions and notification deliveries over multiple transports.

Q: Can subscribers selectively receive only a subset of messages published to a topic?

Yes, you can use message filtering on Amazon Simple Notification Service (SNS) to build simpler and more streamlined pub/sub architectures. Message filtering enables Amazon SNS topic subscribers to selectively receive only a subset of the messages they are interested in, as opposed to receiving all messages published to a topic. To monitor the usage of SNS subscription filter policies, use Amazon CloudWatch metrics, which are automatically collected for you. You can also use the AWS::SNS::Subscription resource type in AWS CloudFormation templates to quickly deploy solutions that use SNS message filtering. For more details, try our 10-minute tutorial, Filter Messages Published to Topics, or see the Filter Messages with Amazon SNS section in our documentation.

Q: Can Amazon SNS be used with other AWS services?

Amazon SNS can be used with other AWS services such as Amazon SQS, Amazon EC2 and Amazon S3. Here is an example of how an order processing workflow system uses Amazon SNS with Amazon EC2, SQS, and SimpleDB. In this workflow system, messages are sent between application components whenever a transaction occurs or an order advances through the order processing pipeline. When a customer initially places an order, the transaction is first recorded in Amazon SimpleDB and an application running on Amazon EC2 forwards the order request to a payment processor which debits the customer’s credit card or bank account. Once approved, an order confirmation message is published to an Amazon SNS topic. In this case, the topic has various subscribers over Email/HTTP – merchant, customer and supply chain partners – and notifications sent by Amazon SNS for that topic can instantly update all of them that payment processing was successful. Notifications can also be used to orchestrate an order processing system running on EC2, where notifications sent over HTTP can trigger real-time processing in related components such as an inventory system or a shipping service. By integrating Amazon SNS with Amazon SQS, all notifications delivered are also persisted in an Amazon SQS queue where they are processed by an auditing application at a future time.

Q: Is Amazon SNS available in all regions where AWS services are available?

Please refer to the AWS Regions and Endpoints section of the AWS documentation for the most up to date information on Amazon SNS availability.

Q: How soon can customers recreate topics with previously used topic names?

Topic names should typically be available for reuse approximately 30-60 seconds after the previous topic with the same name has been deleted. The exact time will depend on the number of subscriptions which were active on the topic – topics with a few subscribers will be available instantly for reuse, topics with larger subscriber lists may take longer.

FIFO topics

Q: What are SNS FIFO topics?

Similar to standard SNS topics, SNS FIFO topics allow users to publish a message to a topic, so it can be delivered to a series of subscribing endpoints. When the delivery of those messages to subscribers must be in order (first-in-first-out), and once only, and you want SNS to take care of it, SNS FIFO topics is the way to go. Amazon SNS FIFO topics deliver ordered messages to Amazon Simple Queue Service (Amazon SQS) FIFO queues to provide consistent end-to-end message ordering for distributed applications. You can now reduce the effort required to process your high throughput, consistently ordered transactions and simplify your messaging architecture. Example use cases include bank transaction logs, stock tickers, flight trackers, price updates, news broadcasting, and inventory management.

Q: When should I use SNS FIFO topics and when should I use Kinesis Data Streams?

Both SNS FIFO topics and Kinesis Streams enable you to build applications that require strictly ordered, many-to-many messaging. SNS FIFO topics can further unlock application integration use cases that require large ordered fan-out, up to 100 subscribers. Kinesis Streams, on the other hand, supports ordered fan-out up to 5 subscribers and is often used for analytics and anomaly detection use cases.

Transports

Q: How would a user subscribe for notifications to be delivered over email?

To receive email notifications for a particular topic, a subscriber should specify “Email” or “Email-JSON” as the protocol and provide a valid email address as the end-point. This can be done using the AWS Management Console or by calling the Amazon SNS API directly. Amazon SNS will then send an email with a confirmation link to the specified email address, and require the user monitoring the email address to explicitly opt-in for receiving email notifications from that particular topic. Once the user confirms the subscription by clicking the provided link, all messages published to the topic will be delivered to that email address.

Q: Why does Amazon SNS provide two different transports to receive notifications over email?

The two email transports are provided for two distinct types of customers/end-users. “Email-JSON” sends notifications as a JSON object, and is meant for applications to programmatically process emails. The ”Email” transport is meant for end-users/consumers and notifications are regular, text-based messages which are easily readable.

Q: Can a user change the Subject and Display name for notifications sent over Email/Email-JSON?

Amazon SNS allows users to specify the Subject field for emails as a parameter passed in to the Publish API call and can be different for every message published. The Display name for topics

can be set using the SetTopicAttributes API – this name applies to all emails sent from this topic.

Q: Do subscribers need to specifically configure their email settings to receive notifications from Amazon SNS?

In most cases, users should be able to receive subscription confirmations and notifications from Amazon SNS without doing anything specific. However, there could be cases where the email provider's default settings or other user-specific configurations mistakenly redirect the emails to the junk/spam folder. To ensure that users see confirmation messages and notifications sent from Amazon SNS, users can add "no-reply@sns.amazonaws.com" to their contact lists and check their junk/spam folders for messages from Amazon SNS.

Q: In the case of passing in an SQS queue as an endpoint, will users need to create the queue prior to subscribing? What permissions will the queue require?

Using the SQS console, users should create the SQS queue prior to subscribing it to a Topic. Select this queue on the console, and from the 'Queue Actions' in the menu bar, select 'Subscribe Queue to SNS Topic' from the drop-down list. In the subscribe dialog box, select the topic from the 'Choose a Topic' drop-down list, and click the 'Subscribe' button. For complete step-by-step instructions, please refer to the Amazon SNS documentation.

Q: How would a developer setup an Amazon SQS queue to receive Amazon SNS notifications?

To have Amazon SNS deliver notifications to an SQS queue, a developer should subscribe to a topic specifying "SQS" as the transport and a valid SQS standard queue as the end-point. In order to allow the SQS queue to receive notifications from Amazon SNS, the SQS queue owner must subscribe the SQS queue to the Topic for Amazon SNS to successfully deliver messages to the queue.

If the user owns both the Amazon SNS topic being subscribed to and the SQS queue receiving the notifications, nothing further is required. Any message published to the topic will automatically be delivered to the specified SQS queue. If the user owning the SQS queue is not the owner of the topic, Amazon SNS will require an explicit confirmation to the subscription request.

Please refer to the Amazon SNS documentation for further details on subscribing an SQS queue to a topic and setting access control policies for SQS queues.

Q: How can I fanout identical messages to multiple SQS queues?

Create an SNS topic first using SNS. Then create and subscribe multiple SQS standard queues to the SNS topic. Now whenever a message is sent to the SNS topic, the message will be fanned out to the SQS queues, i.e. SNS will deliver the message to all the SQS queues that are subscribed to the topic.

Q: What is the format of structured notification messages sent by Amazon SNS?

The notification message sent by Amazon SNS for deliveries over HTTP, HTTPS, Email-JSON and SQS transport protocols will consist of a simple JSON object, which will include the following information:

- MessageId: A Universally Unique Identifier, unique for each notification published.
- Timestamp: The time (in GMT) at which the notification was published.
- TopicArn: The topic to which this message was published

- Type: The type of the delivery message, set to “Notification” for notification deliveries.
- UnsubscribeURL: A link to unsubscribe the end-point from this topic, and prevent receiving any further notifications.
- Message: The payload (body) of the message, as received from the publisher.
- Subject: The Subject field – if one was included as an optional parameter to the publish API call along with the message.
- Signature: Base64-encoded “SHA1withRSA” signature of the Message, MessageId, Subject (if present), Type, Timestamp, and Topic values.
- SignatureVersion: Version of the Amazon SNS signature used.

Notification messages sent over the “Email” transport only contain the payload (message body) as received from the publisher.

Q: How would a user subscribe for notifications to be delivered over SMS?

Please refer to the 'SMS Related Question' section below.

Security

Q: How can users secure the messages sent to my topics?

All API calls made to Amazon SNS are validated for the user’s AWS ID and the signature. In addition, we recommend that users secure their data over the wire by connecting to our secure SSL end-points.

Q: Who can create a topic?

Topics can only be created by users with valid AWS IDs who have signed up for Amazon SNS. The easiest way to create a topic is to use the AWS Management Console. It can also be created through the CreateTopic API. The AWS Management Console is available at: http://aws.amazon.com/console

Q: Can multiple users publish to a single topic?

A topic owner can set explicit permissions to allow more than one user (with a valid AWS ID) to publish to a topic. By default, only topic owners have permissions to publish to a topic.

Q: How can the owner grant/revoke publish or subscribe permissions on a topic?

The AddPermission and RemovePermission APIs provide a simple interface for developers to add and remove permissions for a topic. However, for conditional access and more advanced use cases, users should use access control policies to manage permissions. The easiest way to manage permissions is to use the AWS Management Console. The AWS Management Console is available at: http://aws.amazon.com/console

Q: How does a topic owner give access to subscribers? Do subscribers have to have valid AWS IDs?

Amazon SNS makes it easy for users with and without AWS IDs to receive notifications. The owner of the topic can grant/restrict access to subscribers by setting appropriate permissions for the topic using Access Control policies. Users can receive notifications from Amazon SNS in two ways:

- Users with AWS IDs: Subscribers with valid AWS IDs (please refer to this link for details on obtaining AWS IDs) can subscribe to any topic directly – as long as the topic owner has granted them permissions to do so. The AWS IDs will be validated as part of the subscription registration.
- Other users: Topic owners can subscribe and register end-points on behalf of users without AWS IDs.

In both cases, the owner of the subscription endpoint needs to explicitly opt-in and confirm the subscription by replying to confirmation message sent by Amazon SNS.

Q: How will Amazon SNS authenticate API calls?

All API calls made to Amazon SNS will validate authenticity by requiring that requests be signed with the secret key of the AWS ID account and verifying the signature included in the requests.

Q: How does Amazon SNS validate a subscription request to ensure that notifications will not be sent to users as spam?

As part of the subscription registration, Amazon SNS will ensure that notifications are only sent to valid, registered subscribers/end-points. To prevent spam and ensure that a subscriber end-point is really interested in receiving notifications from a particular topic, Amazon SNS requires an explicit opt-in from subscribers using a 2-part handshake:

i. When a user first calls the Subscribe API and subscribes an end-point, Amazon SNS will send a confirmation message to the specified end-point.

ii. On receiving the confirmation message at the end-point, the subscriber should confirm the subscription request by sending a valid response. Only then will Amazon SNS consider the subscription request to be valid. If there is no response to the challenge, Amazon SNS will not send any notifications to that end-point. The exact mechanism of confirming the subscription varies by the transport protocol selected:

- For HTTP/HTTPS notifications, Amazon SNS will first POST the confirmation message (containing a token) to the specified URL. The application monitoring the URL will have to call the ConfirmSubscription API with the token included token.
- For Email and Email-JSON notifications, Amazon SNS will send an email to the specified address containing an embedded link. The user will need to click on the embedded link to confirm the subscription request.
- For SQS notifications, Amazon SNS will enqueue a challenge message containing a token to the specified queue. The application monitoring the queue will have to call the ConfirmSubscription API with the token.

Note: The explicit “opt-in” steps described above are not required for the specific case where you subscribe your Amazon SQS queue to your Amazon SNS topic – and both are “owned” by the same AWS account.

Q: How long will subscription requests remain pending, while waiting to be confirmed?

Token included in the confirmation message sent to end-points on a subscription request are valid for 3 days.

Q: Who can change permissions on a topic?

Only the owner of the topic can change permissions for that topic.

Q: How can users verify that notification messages are sent from Amazon SNS?

To ensure the authenticity of the notifications, Amazon SNS will sign all notification deliveries using a cryptographically secure, asymmetric mechanism (private-public key pair based on certificates). Amazon SNS will publish its certificate to a well-known location (e.g. http://sns.us-east-1.amazonaws.com/SimpleNotificationService.pem for the US East region) and sign messages with the private key of that certificate. Developers/applications can obtain the certificate and validate the signature in the notifications with the certificate's public key, to ensure that the notification was indeed sent out by Amazon SNS. For further details on certificate locations, please refer to the Amazon SNS details page.

Q: Do publishers have to sign messages as well?

Amazon SNS requires publishers with AWS IDs to validate their messages by signing messages with their secret AWS key; the signature is then validated by Amazon SNS.

Q: Can a publisher/subscriber use SSL to secure messages?

Yes, both publishers and subscribers can use SSL to help secure the channel to send and receive messages. Publishers can connect to Amazon SNS over HTTPS and publish messages over the SSL channel. Subscribers should register an SSL-enabled end-point as part of the subscription registration, and notifications will be delivered over a SSL channel to that end-point.

Q: What permissions does a subscriber need to allow Amazon SNS to send notifications to a registered endpoint?

The owner of the end-point receiving the notifications has to grant permissions for Amazon SNS to send messages to that end-point.

Q: How can subscriptions be unsubscribed?

Subscribers can be unsubscribed either by the topic owner, the subscription owner or others – depending on the mechanism used for confirming the subscription request.

- A subscription that was confirmed with the AuthenticateOnUnsubscribe flag set to True in the call to the ConfirmSubscription API call can only be unsubscribed by a topic owner or the subscription owner.
- If the subscription was confirmed anonymously without the AuthenticateOnUnsubscribe flag set to True, then it can be anonymously unsubscribed.

In all cases except when unsubscribed by the subscription owner, a final cancellation message will be sent to the end-point, allowing the endpoint owner to easily re-subscribe to the topic (if the Unsubscribe request was unintended or in error). For further details on the ConfirmSubscription API, please refer to the Amazon SNS documentation.

Compliance

Q: Is Amazon SNS HIPAA eligible?

Yes, the AWS HIPAA compliance program includes Amazon SNS as a HIPAA eligible Service. If you have an executed Business Associate Agreement (BAA) with AWS, you can now use Amazon SNS to build HIPAA-compliant applications. If you don't have a BAA or have other questions about using AWS for your HIPAA-compliant applications, contact us for more information. Please note that Amazon SNS mobile push notification and SMS functionalities are

outside the scope of the Service's HIPAA eligibility and thus not suitable for transmitting Protected Health Information (PHI).

To learn more, see the following resources:

- AWS HIPAA Compliance page
- AWS Cloud Computing in Healthcare page

To see the current list of compliance programs that Amazon SNS is in scope for, see AWS Services in Scope by Compliance Program.

Q: What else is Amazon SNS compliant with?

Please see AWS Services in Scope by Compliance Program for the latest information about SNS and other AWS services.

Reliability

Q: How durable is my data once published to Amazon SNS?

SNS provides durable storage of all messages that it receives. Upon receiving a publish request, SNS stores multiple copies (to disk) of the message across multiple Availability Zones before acknowledging receipt of the request to the sender. Each AWS Region has multiple, isolated locations known as Availability Zones. Although rare, should a failure occur in one zone, the operation of SNS and the durability of your messages continue without disruption.

Q: Will a notification contain more than one message?

No, all notification messages will contain a single published message.

Q: How many times will a subscriber receive each message?

Although most of the time each message will be delivered to your application exactly once, the distributed nature of Amazon SNS and transient network conditions could result in occasional, duplicate messages at the subscriber end. Developers should design their applications such that processing a message more than once does not create any errors or inconsistencies.

Q: Will messages be delivered to me in the exact order they were published?

The Amazon SNS service will attempt to deliver messages from the publisher in the order they were published into the topic. However, network issues could potentially result in out-of-order messages at the subscriber end.

Q: Can a message be deleted after being published?

No, once a message has been successfully published to a topic, it cannot be recalled.

Q: Does Amazon SNS guarantee that messages are delivered to the subscribed endpoint?

Yes, as long as the subscribed endpoint is accessible. A message delivery fails when Amazon SNS can't access a subscribed endpoint, due to either a client-side or a server-side error. A client-side error happens when the subscribed endpoint has been deleted by the endpoint owner, or when its access permissions have changed in a way that prevents Amazon SNS from delivering messages to this endpoint. A server-side error happens when the service that powers the subscribed endpoint is unavailable, such as Amazon SQS or AWS Lambda. When Amazon SNS receives a client-side error, or continues to receive a server-side error for a message

beyond the number of retries specified by the corresponding retry policy, Amazon SNS discards the message — unless a dead-letter queue is attached to the subscription. For more information, see Message Delivery Retries. and Amazon SNS Dead-Letter Queues.

Q: What happens to Amazon SNS messages if the subscribing endpoint is not available?

If a message cannot be successfully delivered on the first attempt, Amazon SNS executes a 4-phase retry policy: 1) retries with no delay in between attempts, 2) retries with minimum delay between attempts, 3) retries according to a back-off model, and 4) retries with maximum delay between attempts. When the message delivery retry policy is exhausted, Amazon SNS can move the message to a dead-letter queue (DLQ). For more information, see Message Delivery Retries and Amazon SNS Dead-Letter Queues.

Worldwide SMS

Q: What features are part of the new Worldwide SMS capability?

You can use Amazon SNS to deliver SMS (text) messages to 200+ countries, and you do not require recipients to explicitly opt in as before. You must obtain prior permission from recipients to send SMS messages to their phone numbers, where required by local law and regulations. Additionally, you can now mark your SMS messages as Transactional to optimize for reliable delivery, or you can mark it as Promotional to optimize for cost savings. Furthermore, you can set account and message-level spend quotas to avoid inadvertent overruns.

Q: When should I mark an SMS message as Transactional?

SMS messages that are of high priority to your business should be marked as Transactional. This ensures that messages such as those that contain one-time passwords (OTP) or PINs get delivered over routes with the highest delivery reliability. These routes tend to be more expensive than Promotional messaging routes in countries other than the US. You should never mark marketing messages as Transactional, because this violates the local regulatory policies in certain countries, and your account may be marked for abuse and suspended.

Q: When should I mark an SMS message as Promotional?

SMS messages that carry marketing messaging should be marked Promotional. Amazon SNS ensures that such messages are sent over routes that have a reasonable delivery reliability but are substantially cheaper than the most reliable routes. This also allows Amazon SNS to handle and deliver your messages in compliance with on local laws and regulation

Q: What are account-level and message-level spend quotas and how do they work?

Spend quotas can be specified for an AWS account and for individual messages, and the quotas apply only to the cost of sending SMS messages.

The default spend quota per account (if not specified) is 1.00 USD per month. If you want to raise the quota, submit an SNS Quota Increase case. For New quota value, enter your desired monthly spend quota. In the Use Case Description field, explain that you are requesting an SMS monthly spend quota increase.

Amazon SNS sends SMS messages that you publish while the total cost incurred for your SMS traffic is below your spend quota for that calendar month. Once the spend quota is exceeded, Amazon SNS stops delivering messages until you either increase the spend quota or a new calendar month begins. Similarly, you can also specify a spend quota for an individual message, and Amazon SNS will send the message only if the cost is below the quota. Amazon SNS will

not send your SMS messages if the account-level spend quota is exceeded, regardless of whether the message-level spend quota is exceeded.

Q: Is two-way SMS supported?

Amazon SNS does not currently support two-way SMS capabilities, except for opt out where required by local regulations.

Q: Do I need to subscribe phone numbers to an SNS Topic before sending an SMS message to it?

You no longer need to subscribe a phone number to an Amazon SNS topic before you publish messages to it. Now, you can directly publish messages to a phone number using the Amazon SNS console or the Publish request in the Amazon SNS API.

Q: Does AWS offer short codes for purchase?

Yes. You can reserve a dedicated short code that is assigned to your account and available exclusively to you.

To reserve a short code, create a case using the AWS Support Center. For more information, see Reserving a Dedicated Short Code for SMS Messaging in the *Amazon SNS Developer Guide*.

Q: Does AWS offer long codes for purchase?

Yes. You may purchase long codes for use with Amazon SNS as described here.

Q: Will SMS notifications come from a specific origination number?

Amazon SNS will use numbers as configured for your account. It will prioritize using a dedicated short code, followed by one of the dedicated long codes. In case you do not have dedicated numbers, Amazon SNS will fallback to using a shared set of numbers to send SMS notifications. When using the shared set, Amazon SNS attempts to use the same number when sending messages to a specific destination phone number. This is called "Sticky Sender ID". However, depending on various factors like network conditions and throughput available, a different number may be used.

To control the originating identity used by Amazon SNS to send messages you can reserve dedicated short codes or dedicated long codes. By reserving one or more numbers, you make it easier for your audience to recognize that your organization is the source of your messages. For more information, see Reserving a Dedicated Short Code for SMS Messaging or Reserving a Dedicated Long Code for SMS Messaging in the Amazon SNS Developer Guide.

Q: Which countries does Amazon SNS support for Worldwide SMS?

Amazon SNS supports more than 200 countries, and we keep growing our reach. Please refer to the SMS Supported Country List for a comprehensive list of supported calling countries. For SMS message sending to China, please Contact Us.

Q: Which AWS regions support Worldwide SMS?

Please refer to the SNS Supported Regions and Countries page of the Amazon SNS documentation for the latest list of regions where applications using Amazon SNS to send SMS can be hosted in.

Q: Do the AWS phone numbers change?

Yes. Amazon SNS will preferentially use the configured dedicated numbers of an account in priority order of short codes before long codes. If no dedicated numbers are configured one of the numbers from a shared set will be used.

Q: Why do some devices on the same carrier receive messages from different phone numbers?

Amazon SNS will preferentially use the configured dedicated numbers of an account in priority order of short codes before long codes. If no, dedicated numbers are configured one of the numbers from a shared set will be used.

Q: What is the phone number format for sending messages to other countries?

AWS strongly encourages E.164 number formatting for all phone numbers both in the 'to' and 'from' (when applicable) fields. Please refer to the SMS Supported Country List for a comprehensive list of supported countries.

Q: Does Amazon SNS determine if a phone number is a mobile, landline, or VoIP number?

No. Currently, Amazon SNS does not detect whether a phone number is mobile, landline, or VoIP.

Q: Is time-based or scheduled delivery supported for SMS messages?

No. Amazon SNS does not currently support time-based or scheduled delivery.

Q: How do I track the delivery status of my SMS messages?

By enabling the Delivery Status feature in Amazon SNS, you can get information on the following for each message: MessageID, Time Sent, Destination Phone Number, Disposition, Disposition Reason (if applicable), Price, and Dwell Time.

Q: Do you support MMS?

No. Currently Amazon SNS does not support MMS messages.

Q: What is the cost of receiving SMS messages from Amazon SNS?

Costs for receiving SMS messages depend on the Data and Messaging of the recipient's wireless / mobile carrier plans.

Q: How do recipients opt out from receiving SMS messages from AWS?

Recipients can use their devices to opt out by replying to the message with any of the following:

- ARRET (French)
- CANCEL
- END
- OPT-OUT
- OPTOUT
- QUIT
- REMOVE
- STOP

- TD
- UNSUBSCRIBE

To opt out, the recipient must reply to the same long code or short code that Amazon SNS used to deliver the message. After opting out, the recipient will no longer receive SMS messages delivered from your AWS account unless you opt in the phone number.

Q: How do I know if a recipient device has ‘opted out’ of Global SMS?

The SNS console displays the list of opted out numbers for your account. Additionally, the Amazon SNS API provides the ListPhoneNumbersOptedOut request for listing opted out phone numbers.

Q: If a user opts out, will that number be unsubscribed automatically from the SNS Topic?

No. Opt-outs do not unsubscribe a number from an Amazon SNS topic, but rather disable the subscription. This means if you opt-in a phone number you do not need to re-subscribe the phone number to the topic.

Q: How do I confirm the end user received the SMS message?

You can use our Delivery Status feature to get information on the final disposition of your SMS message. For more information on the feature and how to use it, please refer to our documentation.

Q: Does Amazon SNS provide delivery receipts for SMS messages?

Our Delivery Status feature provides information based on delivery receipts received from the destination carrier. For more information on the Delivery Status feature and how to use it, please refer to our documentation.

Q: Does SMS support delivery to VoIP services like Google Voice or Hangouts?

Yes. Amazon SNS does support delivery to VoIP services that can receive SMS messages.

10-Digit Long Codes (10DLC)

Q. What is 10DLC?

A. 10DLC is a 10-digit long code that you can use as an origination identity when sending text messages (SMS) to consumers in the US. It supports a maximum throughput of 100 text messages per second (TPS). AWS doesn't determine the throughput allocated to you. Instead, US carriers allocate throughput to you when you register for the 10DLC. To use 10DLC numbers, carriers require that you provide information about your company and your use cases (also called 10DLC campaigns).

You must register your company and 10DLC campaigns in the Amazon Pinpoint console. Once your registration is complete, and an active 10DLC is available on your account, you can use this number as an origination identity and send SMS using Amazon SNS. Amazon supports the registration process through The Campaign Registry (TCR), a third party central entity that carriers use to validate brand information and 10DLC campaign details.

Q: How long does it take to register a 10DLC campaign?

A: In some cases, registration can occur immediately. For example, if you've previously registered with The Campaign Registry (TCR), they might already have your information.

However, it can take a week or longer to receive approval for some campaigns. After your company and 10DLC campaigns are approved by TCR, you can purchase a 10DLC number and associate it with your campaigns. After you purchase a 10DLC number, it may take up to a week for activation. For more information, see 10DLC in the Amazon SNS Developer Guide.

Q. Can I procure an unregistered P2P long code to send A2P SMS to US phone numbers?

A. No. As of February 16, 2021, you cannot purchase SMS-enabled unregistered person-to-person (P2P) long codes from AWS. Starting June 1, 2021, Amazon SNS no longer supports sending application-to-person (A2P) SMS messages over unregistered US long codes to US destinations. Instead, you can purchase and use short codes, 10DLC, and/or toll-free numbers as originating identities for US destinations. For more information, see Origination numbers in the Amazon SNS Developer Guide.

Q. Should I delete the existing unregistered US long codes in my AWS account?

A. Yes. On June 1, 2021, carriers will no longer deliver messages sent via unregistered long codes to US destinations. If you do not need them for other purposes (for example, voice telephony using other AWS products), delete them from your account. To send SMS, you can convert existing unregistered long codes to 10DLC numbers by associating them with a 10DLC campaign. For more information, see Associating a long code with a 10DLC campaign in the Amazon SNS Developer Guide. Amazon SNS uses Amazon Pinpoint for managing 10DLC campaigns.

Q. I only use Amazon SNS or Amazon Cognito. Should I still use Amazon Pinpoint to register my 10DLC campaign?

A. Yes. You must use Amazon Pinpoint to register 10DLC brands and campaigns. When you complete the registration process and your 10DLC number is activated, Amazon SNS and Amazon Cognito automatically use the 10DLC in your account as the origination ID when sending SMS.

Q. Can I continue to use my long code when it is being migrated to a 10DLC number?

A. Yes. You can continue using the long code as an origination ID, when it is converted into a 10DLC number. It is important that the 10DLC process is completed before June 1, 2021, as unregistered long codes cannot be used after that date.

Q. What is a 10DLC campaign? What information do I need to provide to create one?

A. A 10DLC campaign represents a use case for which you are sending a text message to your customers. For example, you might send a notification when a customer's bill is due. Before you send the SMS, you need to register your use cases for sending text messages, and associate a 10DLC number with a 10DLC campaign. For more information, see Registering a 10DLC campaign in the Amazon SNS Developer Guide. Amazon SNS uses Amazon Pinpoint for managing 10DLC campaigns.

Q: When sending SMS, how does Amazon SNS choose from the origination identities associated with my AWS account?

A: When you publish messages to Amazon SNS, you can choose one of the registered origination identities by setting the AWS.MM.SMS.OriginationNumber attribute. AWS recommends that you specify the origination identity when publishing messages.

When you do not specify an origination identity when publishing the message to Amazon SNS, the following applies:

- If you have only one 10DLC number configured in your account for that AWS region, then all your SMS messages will be sent to US destinations using that 10DLC number.
- If you have multiple origination identities associated with your AWS account, Amazon SNS will choose an origination identity based on the following order: short code, 10DLC, toll-free number.

For information on how to specify an origination identity, see Publishing to a mobile phone, in the Amazon SNS Developer Guide.

Q. Can I use multiple 10DLC numbers for one campaign?

A: Yes. You can associate multiple 10DLC numbers with a single campaign. However, you cannot use the same 10DLC number across multiple campaigns.

Q. I registered my 10DLC company and campaign successfully. However, the associated 10DLC number is stuck in a 'Pending' state. What do I do?

A. When the 10DLC number is in pending state, AWS is working to activate your number on the 10DLC campaign. To activate a number, a valid and active 10DLC brand and 10DLC campaign are required. Activation can take a week or longer to complete. If the 10DLC number has been in pending state for more than a week, raise support case via AWS Support console.

Q. Can I use AWS API actions to request 10DLC numbers instead of using the Amazon Pinpoint console?

A. No. Currently, you can only request 10DLC numbers via the Amazon Pinpoint console. Amazon SNS uses Amazon Pinpoint for managing 10DLC campaigns.

Q. How can I use 10DLC across different AWS regions in my AWS account?

A. 10DLC company and 10DLC campaign registration is specific to an AWS account. However, a 10DLC number is specific to an AWS region. You can have multiple 10DLC numbers in an AWS region referring to the same 10DLC campaign.

Q. Can I get an 10DLC number with a specific area code?

A. No. Currently, AWS does not support choosing 10DLC numbers.

Q. Can I use 10DLC numbers as originating identities for sending SMS outside the US?

A. No. You can only use 10DLC numbers to send SMS messages to US destinations.

Q. Can I use 10DLC numbers for sending voice messages?

A. Yes. To use 10DLC numbers to send voice messages, select voice capability when provisioning these numbers. Note that Amazon SNS does not support voice messages. However, you can use these numbers in other AWS services.

Q. Can I use variables in my 10DLC campaign sample messages?

A. Yes. To use variable content in your sample messages, you can use placeholders in the template that you provide when you register the 10DLC campaign. For example, suppose you want the message to read, "Hi John. Your OTP is 1234." In this case, you would write the template as follows: "Hi {#var1}. Your OTP is {#var2}."

Q. Is there a way to capture metrics for each 10DLC campaign?

A. Currently, there is no metric in Amazon SNS to track messages sent per 10DLC campaign or phone number. For more information, see Monitoring SMS activity, in the Amazon SNS Developer Guide.

Q. Can I migrate 10DLC registrations from one AWS account to another? How long does it take?

A. Yes. To migrate 10DLC registrations between the AWS accounts that you own, create a Service quota Increase support case in the AWS Support Center. You can expect a response within two weeks.

Q. I registered my company directly with The Campaign Registry (TCR) portal. Can I use the same registration for my AWS account?

A. No. To send SMS using Amazon SNS, you must register your brand and 10DLC campaigns with AWS, using the Amazon Pinpoint console. For more information, see Getting started with 10DLC, in the Amazon SNS Developer Guide.

Q. I send SMS messages using Amazon SNS from multiple AWS regions. How do I register a 10DLC number in the AWS region I operate in?

A. 10DLC numbers are specific to an AWS region. 10DLC company and campaigns are valid across AWS regions, in the same AWS account. You can register your brand and campaigns in one AWS region, and procure new 10DLC numbers for those 10DLC campaigns, for use in other AWS regions as needed.

Q. What happens when I send SMS messages at a higher rate than my 10DLC campaign throughput quota?

A. When you exceed your throughput quota, your AWS account experiences throttling errors. The throughput quota is broken down as follows:

- Account level delivery rate of SMS, as specified in Amazon SNS endpoints and quotas in the AWS General Reference.
- Throughput quota based on 10DLC campaign capabilities, as described in 10DLC in the Amazon SNS Developer Guide.

Q. How can I register my company in two different AWS accounts?

A: 10DLC companies and campaigns reside within a single AWS account. If you have multiple accounts, you can associate those other accounts with your main account in order to use your 10DLC numbers from any of those accounts. For more information, see 10DLC cross-account access in the Amazon SNS Developer Guide. Amazon SNS uses Amazon Pinpoint for managing 10DLC campaigns.

Q: Can I use tiny URLs for 10DLC messages?

A: No. Carriers don't allow the use of tiny URLs that services like bit.ly provide. AWS recommends that you use full URLs, matching your company's domain. Alternatively, you can use URL shortening services that provide custom and/or vanity domains and are obviously related to the brand sending the messages. Be sure to provide these URL examples in the sample messages during 10DLC campaign registration.

Q: We use Amazon SNS to send SMS, and we do not set the 'OriginationNumber' attribute. How will Amazon SNS know which 10DLC campaign to use in the event we have more than one campaign in our AWS account?

A. If you have multiple 10DLC campaigns in your AWS account, AWS recommends that you use the 'OriginationNumber' parameter while sending messages via Amazon SNS to use the correct 10DLC campaign. If you don't specify this parameter, Amazon SNS chooses the origination identity for you.

Amazon SNS chooses an origination identity type in the following order of priority: short code, 10DLC, toll-free number. If you have more than one number of a particular origination ID type, Amazon SNS selects one at random within the selected origination ID type.

Q. I use Amazon Cognito to send SMS messages for phone number verification, OTP, and MFA. How can I use a specific 10DLC campaign?

A. Amazon Cognito calls Amazon SNS on your behalf to send SMS. If you have only one 10DLC number and campaign registered in your AWS account, Amazon SNS uses that number when sending SMS on your behalf.

If you have more than one 10DLC number configured in an AWS region, you can use Amazon Cognito's Custom SMS Sender Lambda trigger to choose the origination number to use when sending SMS using Amazon SNS. For more information, see Custom SMS Sender Lambda Trigger in the Amazon Cognito Developer Guide.

Q. How can I send SMS using Amazon SNS via 10DLC numbers from AWS regions that are not supported by Amazon Pinpoint?

A. After a number is configured in an AWS region, you can continue to use Amazon SNS in that region. You can register a 10DLC in an AWS region, and create a service quota increase support case, requesting the transfer of this number to another AWS region of your choice. For more information, see Requesting 10DLC numbers, toll-free numbers, and P2P long codes for SMS messaging in the Amazon SNS Developer Guide.

SMS pricing

Q: How much do you charge for sending SMS messages?

The price you pay for sending SMS messages varies based on the recipient's country or region, and may also vary based on the recipient's mobile carrier. You can find the latest rates on the SMS Pricing page.

Q: Why does the price for sending SMS messages to the same destination country and carrier keep changing?

The costs associated with sending SMS messages to different countries and regions—and even to different carriers within those countries and regions—can change frequently and with little or no notice. Carrier policies, technological changes, and even geopolitical issues can cause the prices for sending SMS messages to change.

We strive to be transparent by publishing the current SMS sending rates on the SMS Pricing Page.

Q: Am I charged if my SMS messages aren't delivered?

You may be charged for failed deliveries if the destination carrier reports that you attempted to send a message to an invalid phone number. Phone numbers can be invalid for several reasons, such as when the phone number doesn't exist, the recipient's account doesn't have sufficient credit, or the destination number is a landline number.

Q: Does the length of a message impact the price I pay?

Yes. A single SMS message can contain a maximum of 140 bytes of information. If a message contains more than 140 bytes, Amazon SNS automatically splits it into multiple messages. When Amazon SNS splits a long message into several smaller messages, you pay for each individual message.

The maximum number of characters in a single message depends on the way the characters are encoded. A message that includes characters encoded using GSM-7 (also known as GSM 03.38) encoding can include 160 characters. A message that uses ASCII encoding can contain up to 140 characters. A message that uses UCS-2 encoding can contain up to 70 characters. When you use Amazon SNS to send an SMS message, it automatically chooses the most compact encoding that supports all of the characters in that message.

For more information about sending SMS messages, see Sending an SMS Message in the *Amazon Simple Notification Service Developer Guide*.

Q: Is there an AWS Free Tier allowance for sending SMS messages?

No.

Quotas and restrictions

Q: Are there quotas for the number of topics or number of subscribers per topic?

By default, SNS offers 10 million subscriptions per topic, and 100,000 topics per account. To request a higher quota, please contact Support.

Q: How much and what kind of data can go in a message?

With the exception of SMS messages, Amazon SNS messages can contain up to 256 KB of text data, including XML, JSON and unformatted text.

The following Unicode characters are accepted:

> #x9 | #xA | #xD | [#x20 to #xD7FF] | [#xE000 to #xFFFD] | [#x10000 to #x10FFFF]
>
> (according to http://www.w3.org/TR/REC-xml/#charsets).

Each 64KB chunk of published data is billed as 1 request. For example, a single API call with a 256KB payload will be billed as four requests.

SMS messages

Each SMS message can contain up to 140 bytes, and the character limit depends on the encoding scheme. For example, an SMS message can contain:

- 160 GSM characters
- 140 ASCII characters
- 70 UCS-2 characters

If you publish a message that exceeds the size limit, Amazon SNS sends it as multiple messages, each fitting within the size limit. Messages are not cut off in the middle of a word but on whole-word boundaries. The total size limit for a single SMS publish action is 1600 bytes.

Q: How many message filters can be applied to a topic?

By default, 200 filter policies per account per region can be applied to a topic. Please contact us if more is required.

Q: Are there TCP ports that should be used for cross-region communication between SNS and EC2?

Yes, cross-region communication between SNS and EC2 on ports other than 80/443/4080/8443 is not guaranteed to work and should be avoided.

Raw message delivery

Q: What is raw message delivery?

You can opt-in to get your messages delivered in raw form, i.e. exactly as you published them. By default, messages are delivered encoded in JSON that provides metadata about the message and topic. Raw message delivery can be enabled by setting the "RawMessageDelivery" property on the subscriptions. This property can be set by using the AWS Management Console, or by using the API SetSubscriptionAttributes.

Q: What is the default behavior if the raw message delivery property on the subscription is not set?

By default, if this property is not set, messages will be delivered in JSON format, which is the current behavior. This ensures existing applications will continue to operate as expected.

Q: Which types of endpoints support raw message delivery?

Raw message delivery support is supported with SQS and HTTP(S) endpoints. Deliveries to Lambda, email, and SMS endpoints will behave the same independent of the "RawMessageDelivery" property.

Q: How will raw messages be delivered to HTTP endpoints?

When raw-formatted messages are delivered to HTTP/s endpoints, the message body will be included in the body of the HTTP POST.

Mobile push notifications

Q: What is SNS Mobile Push?

SNS Mobile Push lets you use Simple Notification Service (SNS) to deliver push notifications to Apple, Google, Fire OS, and Windows devices, as well as Android devices in China with Baidu Cloud Push. With push notifications, an installed mobile application can notify its users immediately by popping a notification about an event, without opening the application. For example, if you install a sports app and enable push notifications, the app can send you the latest score of your favorite team even if the app isn't running. The notification appears on your device, and when you acknowledge it, the app launches to display more information. Users' experiences are similar to receiving an SMS, but with enhanced functionality and at a fraction of the cost.

Q: How do I get started sending push notifications?

Push notifications can only be sent to devices that have your app installed, and whose users have opted in to receive them. SNS Mobile Push does not require explicit opt-in for sending push notifications, but iOS, Android and Kindle Fire operating systems do require it. In order to send push notifications with SNS, you must also register your app and each installed device with SNS. For more information, see Using Amazon SNS Mobile Push Notifications.

Q: Which push notifications platforms are supported?

Currently, the following push notifications platforms are supported:

- Amazon Device Messaging (ADM)
- Apple Push Notification Service (APNS)
- Firebase Cloud Messaging (FCM)
- Windows Push Notification Service (WNS) for Windows 8+ and Windows Phone 8.1+
- Microsoft Push Notification Service (MPNS) for Windows Phone 7+
- Baidu Cloud Push for Android devices in China

Q: How many push notifications can I send with the SNS Free Tier?

The SNS free tier includes 1 million publishes, plus 1 million mobile push deliveries. So you can send 1 million free push notifications every month. Notifications to all mobile push endpoints are all counted together toward your 1 million free mobile push deliveries.

Q: Does enabling push notifications require any special confirmations with SNS Mobile Push?

No, they do not. End-users opt-in to receive push notifications when they first run an app, whether or not SNS delivers the push notifications.

Q: Do I have to modify my client app to use SNS Mobile Push?

SNS does not require you to modify your client app. Baidu Cloud Push requires Baidu-specific components to be added to your client code in order to work properly, whether or not you choose to use SNS.

Q: How do SNS topics work with Mobile Push?

SNS topics can have subscribers from any supported push notifications platform, as well as any other endpoint type such as SMS or email. When you publish a notification to a topic, SNS will send identical copies of that message to each endpoint subscribed to the topic. If you use platform-specific payloads to define the exact payload sent to each push platform, the publish will fail if it exceeds the maximum payload size imposed by the relevant push notifications platform.

Q: What payload size is supported for various target platforms?

SNS will support maximum payload size that is supported by the underlying native platform. Customers can use a JSON object to send platform specific messages. See Using SNS Mobile Push API for additional details.

Q: How do platform-specific payloads work?

When you publish to a topic and want to have customized messages sent to endpoints for the different push notification platforms then you need to select “Use different message body for different protocols” option on the Publish dialog box and then update the messages. You can use platform-specific payloads to specify the exact API string that is relayed to each push notifications service. For example, you can use platform-specific payloads to manipulate the badge count of your iOS application via APNS. For more information, see Using Amazon SNS Mobile Push Notifications.

Q: Can one token subscribe to multiple topics?

Yes. Each token can be subscribed to an unlimited number of SNS topics.

Q: What is direct addressing? How does it work?

Direct addressing allows you to deliver notifications directly to a single endpoint, rather than sending identical messages to all subscribers of a topic. This is useful if you want to deliver precisely targeted messages to each recipient. When you register device tokens with SNS, SNS creates an endpoint that corresponds to the token. You can publish to the token endpoint just as you would publish to a topic. You can direct publish either the text of your notification, or a platform-specific payload that takes advantage of platform-specific features such as updating the badge count of your app. Direct addressing is currently only available for push notifications endpoints.

Q: Does SNS support direct addressing for SMS or Email?

At this time, direct addressing is only supported for mobile push endpoints (APNS, FCM, ADM, WNS, MPNS, Baidu) and SMS. Email messaging requires the use of topics.

Q: How does SNS Mobile Push handle token feedback from notification services?

Push notification services such as APNS and FCM provide feedback on tokens which may have expired or may have been replaced by new tokens. If either APNS or FCM reports that a particular token has either expired or is invalid, SNS automatically "disables" the application endpoint associated with the token, and notifies you of this change via an event. FCM specifically, at times not only indicates that a token is invalid, but also provides the new token associated with the application endpoint in its response to SNS. When this happens, SNS automatically updates the associated endpoint with the new token value, leaving the endpoint enabled, and then notifies you of this change via an event.

Q: I use Google Cloud Messaging (GCM) for SNS mobile notifications. What happens when GCM is deprecated?

GCM device tokens are completely interchangeable with the newer Firebase Cloud Messaging (FCM) device tokens. If you have existing GCM tokens, you'll still be able to use them to send notifications. This statement is also true for GCM tokens that you generate in the future. For more information please visit [The End of Google Cloud Messaging, and What it Means for Your Apps] blog.

Q: Can I migrate existing apps to SNS Mobile Push?

Yes. You can perform a bulk upload of existing device tokens to Amazon SNS, either via the console interface or API. You would also register your app with SNS by uploading your credentials for the relevant push notifications services, and configure your proxy or app to register future new tokens with SNS.

Q: Can I monitor my push notifications through Amazon CloudWatch?

Yes. SNS publishes Cloudwatch metrics for number of messages published, number of successful notifications, number of failed notifications, number of notifications filtered out, and size of data published. Metrics are available on per application basis. You can access Cloudwatch metrics via AWS Management Console or CloudWatch APIs.

Q: What types of Windows Push Notifications does Amazon SNS support?

SNS supports all types of push notifications types offered by Microsoft WNS and MPNS, including toast, tile, badge and raw notifications. Use the TYPE message attribute to specify which notification type you wish to use. When you use default payloads to send the same

message to all mobile platforms, SNS will select toast notifications by default for Windows platforms. It is required to specify a notification type for Windows platforms when you use platform-specific payloads.

Q: Does SNS support Windows raw push notifications?

Yes. You must encode the notification payload as text to send raw notifications via SNS.

Q: What is Baidu Cloud Push?

Baidu Cloud Push is a third-party alternative push notifications relay service for Android devices. You can use Baidu Cloud Push to reach Android customers in China, no matter what Android app store those customers choose to use for downloading your app. For more information about Baidu Cloud Push, visit: https://push.baidu.com/.

Q: Can I publish Baidu notifications from all public AWS regions?

Yes, SNS supports Baidu push notifications from all public AWS regions.

Q: Can I use Baidu notifications to any Android app store?

Yes, Baidu push notifications work for apps installed via any Android app store.

Q: What are message attributes?

Message attributes allow you to provide structured metadata items (such as timestamps, geospatial data, signatures, and identifiers) about the message. Message attributes are optional and separate from, but sent along with, the message body. This information can be used by the receiver of the message to help decide how to handle the message without having to first process the message body.

You can use SNS message attributes in conjunction with SQS and mobile push endpoints. To learn more about message attributes, please see the SNS Getting Started Guide.

Q: What message attributes are supported in SNS?

SNS supports different message attributes for each endpoint type, depending on what the endpoint types each support themselves.

- For SQS endpoints, you can specify up to 10 name-type-value triples per message. Types supported include: String, Binary and Number (including integers, floating point, and doubles).
- For mobile push endpoints, you can take advantage of specific message attributes that each mobile platform supports (such as notification type).

Q: What is Time to Live (TTL)?

Some messages that you can send with SNS are relevant or valuable only for a limited period of time. Amazon SNS now allows you to set a TTL (Time to Live) value for each message. When the TTL expires for a given message that was not delivered and read by an end user, the message is deleted. TTL is specified in seconds and is relative to the time Publish call is made.

Q: How do I specify a TTL for my messages?

You can specify a TTL using the console or via API. TTL can be specified at publish time for a message, using the message attribute below. There is a different attribute for each platform. An

attribute specified for a platform is applicable only for notification deliveries to that platform.

Q: What is the default TTL?

SNS uses a default Time to Live (TTL) of 4 weeks for all mobile platforms.

Q: Do TTL message attributes override TTLs specified in a message payload?

Yes. Google FCM and Amazon ADM allow you to specify a TTL within the message payload. If you specify TTL within the message payload and also within a message attribute, SNS will follow the message attribute.

Q: What happens if I specify TTL=0?

Some platforms treat TTL = 0 as a special case and attempt to deliver the message immediately, else let it expire. If you specify TTL = 0, SNS will relay your message to the appropriate service with TTL = 0 in order to take advantage of this special case.

Q: What SNS endpoints support TTL?

You can use TTL with the following mobile push endpoints: APNS, APNS_Sandbox, FCM, ADM, Baidu, and WNS. Microsoft MPNS does not currently support TTL. TTL is also not supported for SQS, HTTP, email or SMS endpoints.

Q: What does the Delivery Status feature of Amazon SNS do?

The Delivery Status feature lets you collect information on success rates, failure rates and dwell times of your push notifications for the supported mobile notification platforms. The currently supported platforms include Apple (APNS), Google (FCM), Windows (WNS and MPNS), Amazon (ADM), and Baidu. The status information is captured in the Amazon CloudWatch log groups created by Amazon SNS on your behalf. Additionally, you can create actionable metrics in Amazon CloudWatch and trigger alarms based on the patterns you are interested in.

Q: Is the Delivery Status feature in Amazon SNS available only for mobile push notifications? Do you plan to support this feature for other endpoint types?

Currently the Delivery Status feature is available for mobile push notifications and SMS. We will evaluate extending this to other endpoint types based on feedback from customers.

Q: How do I activate the Delivery Status feature?

You can activate the Delivery Status feature from the Amazon SNS console. From your Application, choose the Delivery Status option in the Application Actions drop-down menu. For details, please read our documentation.

Q: Can I activate the Delivery Status feature from the Amazon SNS APIs?

Yes, you can activate this feature from Amazon SNS APIs by adding the relevant application-level attributes. Our documentation goes over the application-level attributes that you need to add and the specific API calls that need to be made to enable this feature.

Q: How much does the Delivery Status feature cost?

There is currently no additional Amazon SNS charge for using the Delivery Status feature.

However, depending upon your usage, you may incur charges for using CloudWatch since this feature creates Amazon CloudWatch log groups.

Q: Why can you only choose a sampling percentage for successful delivery attempts and not sample failed delivery attempts?

Based on feedback we received from customers, we found that most developers are interested in knowing all the delivery attempt failures for their applications – and prefer to only store sample successful deliveries rather than logging all of them.

Q: How can I set alarms based on failure metrics or dwell time metrics?

After activating the Delivery Status feature, you need to define a Log Metrics Filter in Amazon CloudWatch Logs for the log group that gets created by Amazon SNS on your behalf. This metrics filter can be defined to extract information that you are interested in, such as failure rate and dwell time. Once a Metric Filter is defined, you can create it and assign it to a Metric. This metric can then be used to set alarms or send notifications based on thresholds you define. For more information, take a look at our documentation or blog.

Q: Does Amazon SNS support HTTP/2 for mobile push notification to APNS endpoints?

Amazon SNS uses HTTP/2 with p12 certificates for sending push notifications via Apple Push Notification Service (APNS) to iOS and macOS endpoints.

Q: Do I have to modify my application due to the deprecation of APNS binary protocol as of November 2020?

Amazon SNS uses HTTP/2 with p12 certificates. As it does not rely on the legacy binary protocol, no change is required in your application that is sending push notifications via Amazon SNS.

SNS support for AWS Lambda

Q: What does support for AWS Lambda endpoints in Amazon SNS mean?

You can invoke your AWS Lambda functions by publishing messages to Amazon SNS topics that have AWS Lambda functions subscribed to them. Because Amazon SNS supports message fan-out, publishing a single message can invoke different AWS Lambda functions or invoke Lambda functions in addition to delivering notifications to supported Amazon SNS destinations such as mobile push, HTTP endpoints, SQS, email and SMS.

Q: What is AWS Lambda?

AWS Lambda is a compute service that runs your code in response to events and automatically manages the compute resources for you, making it easy to build applications that respond quickly to new information. More information on AWS Lambda and how to create AWS Lambda functions can be found here.

Q: What can I do with AWS Lambda functions and Amazon SNS?

By subscribing AWS Lambda functions to Amazon SNS topics, you can perform custom message handling. You can invoke an AWS Lambda function to provide custom message delivery handling by first publishing a message to an AWS Lambda function, have your Lambda function modify a message (e.g. localize language) and then filter and route those messages to other topics and endpoints. Apps and services that already send Amazon SNS notifications,

such as Amazon CloudWatch, can now immediately take advantage of AWS Lambda without having to provision or manage infrastructure for custom message handling. You can also use delivery to an AWS Lambda function as a way to publish to other AWS services such as Amazon Kinesis or Amazon S3. You can subscribe an AWS Lambda function to the Amazon SNS topic, and then have the Lambda function in turn write to another service.

Q: How do I activate AWS Lambda endpoint support in Amazon SNS?

You need to first create an AWS Lambda function via your AWS account and the AWS Lambda console, and then subscribe that AWS Lambda function to a topic using the Amazon SNS console or the Amazon SNS APIs. Once that is complete, any messages that you publish to the Amazon SNS topics which have Lambda functions subscribed to them will be delivered to the appropriate Lambda functions in addition to any other destinations subscribed to that topic.

Q: What does delivery of a message from Amazon SNS to an AWS Lambda function do?

A message delivery from Amazon SNS to an AWS Lambda function creates an instance of the AWS Lambda function and invokes it with your message as an input. For more information on message formats, please refer to the Amazon SNS documentation and the AWS Lambda documentation.

Q: How much does this feature cost?

Publishing a message with Amazon SNS costs $0.50 per million requests. Aside from charges incurred in using AWS services, there are no additional fees for delivering a message to an AWS Lambda function. Amazon SNS has a Free Tier of 1 million requests per month. AWS Lambda function costs are based on the number of requests for your functions and the time your code executes. The AWS Lambda Free-Tier includes 1M requests per month and 400,000 GB-seconds of compute time per month.

Q: Can I subscribe AWS Lambda functions created by someone else to Amazon SNS topics that I own?

We currently do not allow an AWS account owner to subscribe an AWS Lambda function that belongs to another account. You can subscribe your own AWS Lambda functions to your own Amazon SNS topics or subscribe your AWS Lambda functions to an Amazon SNS topic that was created by another account so long as the topic policy for that SNS topic allows it.

Q: Is there a quota to the number of AWS Lambda functions that I can subscribe to an Amazon SNS topic?

Amazon SNS treats AWS Lambda functions like any other destination. By default, SNS offers 10 million subscriptions per topic. To request a higher quota, please contact us.

Q: What data can I pass to my AWS Lambda function?

When an AWS Lambda function is invoked as a result of an Amazon SNS message delivery, the AWS Lambda function receives data such as the Message ID, the topic ARN, the message payload, and message attributes via an SNS Event. For more information on the event structure passed to the AWS Lambda function please read our blog.

Q: Can I track delivery status for message delivery attempts to AWS Lambda functions?

To track the success or failure status of message deliveries, you need to activate the Delivery Status feature of Amazon SNS. For more information about how to activate this feature please read our blog.

Q: What regions is AWS Lambda available in?

See AWS Regions and Endpoints for a complete list.

Q: Do my AWS Lambda functions need to be in the same region as my Amazon SNS usage?

You can subscribe your AWS Lambda functions to an Amazon SNS topic in any region.

Q: Are there any data transfer costs for invoking AWS Lambda functions?

Data transfer costs are applicable to message deliveries to AWS Lambda functions.

Q: Are there any quotas to the concurrency of AWS Lambda functions?

AWS Lambda currently supports 1000 concurrent executions per AWS account per region. If your Amazon SNS message deliveries to AWS Lambda contribute to crossing these concurrency quotas, your Amazon SNS message deliveries will be throttled. If AWS Lambda throttles an Amazon SNS message, Amazon SNS will retry the delivery attempts. For more information about AWS Lambda concurrency quotas, please refer to AWS Lambda documentation.

Q: Can Amazon SNS use the same AWS Lambda functions that I use with other services (e.g. Amazon S3)?

You can use the same AWS Lambda functions that you use with other services as long as the same function can parse the event formats from Amazon SNS in addition to the event format of the other services. For the SNS event format please read our blog.

VoIP iOS and Mac OS notifications

Q: What are VoIP Push Notifications for iOS?

In iOS 8 and later, voice-over-IP (VoIP) apps can register for VoIP remote notifications such that iOS can launch or wake the app, as appropriate, when an incoming VoIP call arrives for the user. The procedure to register for VoIP notifications is similar to registering for regular push notifications on iOS. For more information, please refer to our documentation.

Q: Can I use VoIP Push Notifications and other Push Notifications in the same iOS app?

Yes, you can have an iOS application that is registered to receive both types of push notifications. However, you will need to obtain the VoIP push notification certificate from Apple in addition to the regular push notification certificate and create a new Platform Application in Amazon SNS and choose Apple VoIP Push as the platform type. For more information, please refer to our documentation.

Q: What are Mac OS push notifications?

You can now send push notifications to Mac OS desktops that run Mac OS X Lion (10.7) or later using Amazon SNS. For more information, please refer to our documentation.

Amazon SQS FAQs

Overview

Q: What are the benefits of Amazon SQS over homegrown or packaged message queuing systems?

Amazon SQS provides several advantages over building your own software for managing message queues or using commercial or open-source message queuing systems that require significant upfront time for development and configuration.

These alternatives require ongoing hardware maintenance and system administration resources. The complexity of configuring and managing these systems is compounded by the need for redundant storage of messages that ensures messages are not lost if hardware fails.

In contrast, Amazon SQS requires no administrative overhead and little configuration. Amazon SQS works on a massive scale, processing billions of messages per day. You can scale the amount of traffic you send to Amazon SQS up or down without any configuration. Amazon SQS also provides extremely high message durability, giving you and your stakeholders added confidence.
Q: How is Amazon SQS different from Amazon Simple Notification Service (SNS)?

Amazon SNS allows applications to send time-critical messages to multiple subscribers through a “push” mechanism, eliminating the need to periodically check or “poll” for updates. Amazon SQS is a message queue service used by distributed applications to exchange messages through a polling model, and can be used to decouple sending and receiving components.
Q: How is Amazon SQS different from Amazon MQ?

If you're using messaging with existing applications, and want to move your messaging to the cloud quickly and easily, we recommend you consider Amazon MQ. It supports industry-standard APIs and protocols so you can switch from any standards-based message broker to Amazon MQ without rewriting the messaging code in your applications. If you are building brand new applications in the cloud, we recommend you consider Amazon SQS and Amazon SNS. Amazon SQS and SNS are lightweight, fully managed message queue and topic services that scale almost infinitely and provide simple, easy-to-use APIs.
Q: Does Amazon SQS provide message ordering?

Yes. FIFO (first-in-first-out) queues preserve the exact order in which messages are sent and received. If you use a FIFO queue, you don't have to place sequencing information in your messages. For more information, see FIFO Queue Logic in the *Amazon SQS Developer Guide*.

Standard queues provide a loose-FIFO capability that attempts to preserve the order of messages. However, because standard queues are designed to be massively scalable using a highly distributed architecture, receiving messages in the exact order they are sent is not guaranteed.
Q: Does Amazon SQS guarantee delivery of messages?

Standard queues provide at-least-once delivery, which means that each message is delivered at least once.

FIFO queues provide exactly-once processing, which means that each message is delivered once and remains available until a consumer processes it and deletes it. Duplicates are not introduced into the queue.
Q: How is Amazon SQS different from Amazon Kinesis Streams?

Amazon SQS offers a reliable, highly-scalable hosted queue for storing messages as they travel between applications or microservices. It moves data between distributed application components and helps you decouple these components. Amazon SQS provides common middleware constructs such as dead-letter queues and poison-pill management. It also provides a generic web services API and can be accessed by any programming language that the AWS SDK supports. Amazon SQS supports both standard and FIFO queues.

Amazon Kinesis Streams allows real-time processing of streaming big data and the ability to read and replay records to multiple Amazon Kinesis Applications. The Amazon Kinesis Client Library (KCL) delivers all records for a given partition key to the same record processor, making it easier to build multiple applications that read from the same Amazon Kinesis stream (for example, to perform counting, aggregation, and filtering).

For more information, see the Amazon Kinesis Documentation.
Q: Does Amazon use Amazon SQS for its own applications?

Yes. Developers at Amazon use Amazon SQS for a variety of applications that process large numbers of messages every day. Key business processes in both Amazon.com and AWS use Amazon SQS.

Features, functionality, and interfaces

Q: Can I use Amazon SQS with other AWS services?

Yes. You can make your applications more flexible and scalable by using Amazon SQS with compute services such as Amazon EC2, Amazon Elastic Container Service (ECS), and AWS Lambda, as well as with storage and database services such as Amazon Simple Storage Service (Amazon S3) and Amazon DynamoDB.
Q: How can I interact with Amazon SQS?

You can access Amazon SQS using the AWS Management Console, which helps you create Amazon SQS queues and send messages easily.

Amazon SQS also provides a web services API. It is also integrated with the AWS SDKs, allowing you to work in the programming language of your choice.
Q: What API actions are available for Amazon SQS?

For information on message queue operations, see the Amazon SQS API Reference.
Q: Who can perform operations on a message queue?

Only an AWS account owner (or an AWS account that the account owner has delegated rights to) can perform operations on an Amazon SQS message queue.
Q: Can I use Java Message Service (JMS) with Amazon SQS?

Yes. You can take advantage of the scale, low cost, and high availability of Amazon SQS without the worry and high overhead of running your own JMS cluster.

Amazon provides the Amazon SQS Java Messaging Library that implements the JMS 1.1 specification and uses Amazon SQS as the JMS provider. For more information, see Using JMS with Amazon SQS in the *Amazon SQS Developer Guide*.
Q: How does Amazon SQS identify messages?

All messages have a global unique ID that Amazon SQS returns when the message is delivered to the message queue. The ID isn’t required to perform any further actions on the message, but it is useful for tracking the receipt of a particular message in the message queue.

When you receive a message from the message queue, the response includes a receipt handle that you must provide when deleting the message.

For more information, see Queue and Message Identifiers in the Amazon SQS Developer Guide.
Q: How does Amazon SQS handle messages that can't be processed?

In Amazon SQS, you can use the API or the console to configure dead letter queues, which receive messages from other source queues. When configuring a dead letter queue, you are

required to set appropriate permissions for the dead letter queue redrive using RedriveAllowPolicy.

RedriveAllowPolicy includes the parameters for the dead-letter queue redrive permission. It defines which source queues can specify dead-letter queues as a JSON object.

Once you make a dead letter queue, it receives messages after a maximum number of processing attempts cannot be completed. You can use dead letter queues to isolate messages that can't be processed for later analysis.

For more information, see Using Amazon SQS Dead Letter Queues in the Amazon SQS Developer Guide.
Q: What is a visibility timeout?

The visibility timeout is a period of time during which Amazon SQS prevents other consuming components from receiving and processing a message. For more information, see Visibility Timeout in the *Amazon SQS Developer Guide*.
Q: Does Amazon SQS support message metadata?

Yes. An Amazon SQS message can contain up to 10 metadata attributes. You can use message attributes to separate the body of a message from the metadata that describes it. This helps process and store information with greater speed and efficiency because your applications don't have to inspect an entire message before understanding how to process it.

Amazon SQS message attributes take the form of name-type-value triples. The supported types include string, binary, and number (including integer, floating-point, and double). For more information, see Using Amazon SQS Message Attributes in the *Amazon SQS Developer Guide*.
Q: How can I determine the time-in-queue value?

To determine the time-in-queue value, you can request the SentTimestamp attribute when receiving a message. Subtracting that value from the current time results in the time-in-queue value.
Q: What is the typical latency for Amazon SQS?

Typical latencies for SendMessage, ReceiveMessage, and DeleteMessage API requests are in the tens or low hundreds of milliseconds.
Q: For anonymous access, what is the value of the SenderId attribute for a message?

When the AWS account ID is not available (for example, when an anonymous user sends a message), Amazon SQS provides the IP address.
Q: What is Amazon SQS long polling?

Amazon SQS long polling is a way to retrieve messages from your Amazon SQS queues. While the regular short polling returns immediately, even if the message queue being polled is empty, long polling doesn’t return a response until a message arrives in the message queue, or the long poll times out.

Long polling makes it inexpensive to retrieve messages from your Amazon SQS queue as soon as the messages are available. Using long polling might reduce the cost of using SQS, because you can reduce the number of empty receives. For more information, see Amazon SQS Long Polling in the *Amazon SQS Developer Guide*.
Q: Is there an additional charge for using Amazon SQS long polling?

No. Long-polling ReceiveMessage calls are billed exactly the same as short-polling ReceiveMessage calls.

Q: When should I use Amazon SQS long polling, and when should I use Amazon SQS short polling?

In almost all cases, Amazon SQS long polling is preferable to short polling. Long-polling requests let your queue consumers receive messages as soon as they arrive in your queue while reducing the number of empty ReceiveMessageResponse instances returned.

Amazon SQS long polling results in higher performance at reduced cost in the majority of use cases. However, if your application expects an immediate response from a ReceiveMessage call, you might not be able to take advantage of long polling without some modifications to your application.

For example, if your application uses a single thread to poll multiple queues, switching from short polling to long polling will probably not work, because the single thread will wait for the long-poll timeout on any empty queues, delaying the processing of any queues that might contain messages.

In such an application, it is a good practice to use a single thread to process only one queue, allowing the application to take advantage of the benefits that Amazon SQS long polling provides.
Q: What value should I use for my long-poll timeout?

In general, you should use a maximum of 20 seconds for a long-poll timeout. Because higher long-poll timeout values reduce the number of empty ReceiveMessageResponse instances returned, try to set your long-poll timeout as high as possible.

If the 20-second maximum doesn't work for your application (see the example in the previous question), set a shorter long-poll timeout, as low as 1 second.

All AWS SDKs work with 20-second long polls by default. If you don't use an AWS SDK to access Amazon SQS, or if you configured your AWS SDK to specifically have a shorter timeout, you might need to modify your Amazon SQS client to allow longer requests or to use a shorter long-poll timeout.
Q: What is the AmazonSQSBufferedAsyncClient for Java?

The AmazonSQSBufferedAsyncClient for Java provides an implementation of the AmazonSQSAsyncClient interface and adds several important features:

- Automatic batching of multiple SendMessage, DeleteMessage, or ChangeMessageVisibility requests without any required changes to the application
- Prefetching of messages into a local buffer that allows your application to immediately process messages from Amazon SQS without waiting for the messages to be retrieved

Working together, automatic batching and prefetching increase the throughput and reduce the latency of your application while reducing your costs by making fewer Amazon SQS requests. For more information, see Client-Side Buffering and Request Batching in the *Amazon SQS Developer Guide*.
Q: Where can I download the AmazonSQSBufferedAsyncClient for Java?

You can download the AmazonSQSBufferedAsyncClient as part of the AWS SDK for Java.
Q: Do I have to rewrite my application to use the AmazonSQSBufferedAsyncClient for Java?

No. The AmazonSQSBufferedAsyncClient for Java is implemented as a drop-in replacement for the existing AmazonSQSAsyncClient.

If you update your application to use the latest AWS SDK and change your client to use the AmazonSQSBufferedAsyncClient for Java instead of the AmazonSQSAsyncClient, your application will receive the added benefits of automatic batching and prefetching.
Q: How can I subscribe Amazon SQS message queues to receive notifications from Amazon SNS topics?

1. In the Amazon SQS console, select an Amazon SQS standard queue.
2. Under Queue Actions, select Subscribe Queue to SNS Topic from the drop-down list.
3. In the dialog box, select the topic from the Choose a Topic drop-down list, and click Subscribe.

For more information, see Subscribing a Queue to an Amazon SNS Topic in the *Amazon SQS Developer Guide*.
Q: Can I delete all messages in a message queue without deleting the message queue itself?

Yes. You can delete all messages in an Amazon SQS message queue using the PurgeQueue action.

When you purge a message queue, all the messages previously sent to the message queue are deleted. Because your message queue and its attributes remain, there is no need to reconfigure the message queue; you can continue using it.

To delete only specific messages, use the DeleteMessage or DeleteMessageBatch actions.

For more information, see this Tutorial: Purging Messages from an Amazon SQS Queue.

FIFO queues

Q: What regions are FIFO queues available in?

FIFO queues are available in all AWS regions where Amazon SQS is available. See here for details on Amazon SQS region availability.
Q: How many copies of a message will I receive?

FIFO queues are designed to never introduce duplicate messages. However, your message producer might introduce duplicates in certain scenarios: for example, if the producer sends a message, does not receive a response, and then resends the same message. Amazon SQS APIs provide deduplication functionality that prevents your message producer from sending duplicates. Any duplicates introduced by the message producer are removed within a 5-minute deduplication interval.

For standard queues, you might occasionally receive a duplicate copy of a message (at-least-once delivery). If you use a standard queue, you must design your applications to be idempotent (that is, they must not be affected adversely when processing the same message more than once).

For more information, see Exactly-Once Processing in the Amazon SQS Developer Guide.
Q: Are the Amazon SQS queues I used previously changing to FIFO queues?

No. Amazon SQS *standard* queues (the new name for existing queues) remain unchanged, and you can still create standard queues. These queues continue to provide the highest scalability and throughput; however, you will not get ordering guarantees and duplicates might occur.

Standard queues are appropriate for many scenarios, such as work distribution with multiple idempotent consumers.

Q: Can I convert my existing standard queue to a FIFO queue?

No. You must choose the queue type when you create it. However, it is possible to move to a FIFO queue. For more information, see Moving From a Standard Queue to a FIFO Queue in the *Amazon SQS Developer Guide*.

Q: Are Amazon SQS FIFO queues backwards-compatible?

To take advantage of FIFO queue functionality, you must use the latest AWS SDK.

FIFO queues use the same API actions as standard queues, and the mechanics for receiving and deleting messages and changing the visibility timeout are the same. However, when sending messages, you must specify a message group ID. For more information, see FIFO Queue Logic in the *Amazon SQS Developer Guide*.

Q: With which AWS or external services are Amazon SQS FIFO queues compatible?

Some AWS or external services that send notifications to Amazon SQS might not be compatible with FIFO queues, despite allowing you to set a FIFO queue as a target.

The following features of AWS services aren't currently compatible with FIFO queues:

- Auto Scaling Lifecycle Hooks
- AWS IoT Rule Actions
- AWS Lambda Dead Letter Queues

For information about compatibility of other services with FIFO queues, see your service documentation.

Q: Are Amazon SQS FIFO queues compatible with the Amazon SQS Buffered Asynchronous Client, the Amazon SQS Extended Client Library for Java, or the Amazon SQS Java Message Service (JMS) Client?

FIFO queues aren't currently compatible with the Amazon SQS Buffered Asynchronous Client.

FIFO queues are compatible with the Amazon SQS Extended Client Library for Java and the Amazon SQS Java Message Service (JMS) client.

Q: Which AWS CloudWatch metrics do Amazon SQS FIFO queues support?

FIFO queues support all metrics that standard queues support. For FIFO queues, all approximate metrics return accurate counts. For example, the following AWS CloudWatch metrics are supported:

- ApproximateNumberOfMessagesDelayed - The number of messages in the queue that are delayed and not available for reading immediately.
- ApproximateNumberOfMessagesVisible - The number of messages available for retrieval from the queue.
- ApproximateNumberOfMessagesNotVisible - The number of messages that are in flight (sent to a client but have not yet been deleted or have not yet reached the end of their visibility window).

Q: What are message groups?

Messages are grouped into distinct, ordered "bundles" within a FIFO queue. For each message group ID, all messages are sent and received in strict order. However, messages with different

message group ID values might be sent and received out of order. You must associate a message group ID with a message. If you don't provide a message group ID, the action fails.

If multiple hosts (or different threads on the same host) send messages with the same message group ID are sent to a FIFO queue, Amazon SQS delivers the messages in the order in which they arrive for processing. To ensure that Amazon SQS preserves the order in which messages are sent and received, ensure that multiple senders send each message with a unique message group ID.

For more information, see FIFO Queue Logic in the *Amazon SQS Developer Guide*.
Q: Do Amazon SQS FIFO queues support multiple producers?

Yes. One or more producers can send messages to a FIFO queue. Messages are stored in the order that they were successfully received by Amazon SQS.

If multiple producers send messages in parallel, without waiting for the success response from SendMessage or SendMessageBatch actions, the order between producers might not be preserved. The response of SendMessage or SendMessageBatch actions contains the final ordering sequence that FIFO queues use to place messages in the queue, so your multiple-parallel-producer code can determine the final order of messages in the queue.
Q: Do Amazon SQS FIFO queues support multiple consumers?

By design, Amazon SQS FIFO queues don't serve messages from the same message group to more than one consumer at a time. However, if your FIFO queue has multiple message groups, you can take advantage of parallel consumers, allowing Amazon SQS to serve messages from different message groups to different consumers.
Q: What is the throughput quota for an Amazon SQS FIFO queue?

By default, FIFO queues support up to 3,000 messages per second with batching or up to 300 messages per second (300 send, receive, or delete operations per second) without batching. If you require higher throughput, you can enable high throughput mode for FIFO on the Amazon SQS console, which will support up to 30,000 messages per second with batching, or up to 3,000 messages per second without batching.
Q: Are there any limits specific to FIFO queue attributes?

The name of a FIFO queue must end with the .fifo suffix. The suffix counts towards the 80-character queue name limits. To determine whether a queue is FIFO, you can check whether the queue name ends with the suffix.

Security and reliability

Q: How reliable is the storage of my data in Amazon SQS?

Amazon SQS stores all message queues and messages within a single, highly-available AWS region with multiple redundant Availability Zones (AZs), so that no single computer, network, or AZ failure can make messages inaccessible. For more information, see Regions and Availability Zones in the *Amazon Relational Database Service User Guide*.
Q: How can I secure the messages in my message queues?

Authentication mechanisms ensure that messages stored in Amazon SQS message queues are secured against unauthorized access. You can control who can send messages to a message queue and who can receive messages from a message queue. For additional security, you can build your application to encrypt messages before they are placed in a message queue.

Amazon SQS has its own resource-based permissions system that uses policies written in the same language as AWS Identity and Access Management (IAM) policies: for example, you can

use variables, just like in IAM policies. For more information, see Amazon SQS Policy Examples in the *Amazon SQS Developer Guide*.

Amazon SQS supports the HTTP over SSL (HTTPS) and Transport Layer Security (TLS) protocols. Most clients can automatically negotiate to use newer versions of TLS without any code or configuration change. Amazon SQS supports versions 1.0, 1.1, and 1.2 of the Transport Layer Security (TLS) protocol in all regions.
Q: Why are there separate ReceiveMessage and DeleteMessage operations?

When Amazon SQS returns a message to you, the message stays in the message queue whether or not you actually receive the message. You're responsible for deleting the message and the deletion request acknowledges that you're done processing the message.

If you don't delete the message, Amazon SQS will deliver it again when it receives another receive request. For more information, see Visibility Timeout in the *Amazon SQS Developer Guide*.
Q: Can a deleted message be received again?

No. FIFO queues never introduce duplicate messages.

For standard queues, under rare circumstances, you might receive a previously-deleted message a second time.
Q: What happens if I issue a DeleteMessage request on a previously-deleted message?

When you issue a DeleteMessage request on a previously-deleted message, Amazon SQS returns a *success* response.

Server-Side encryption (SSE)

Q: What are the benefits of SSE for Amazon SQS?

SSE lets you transmit sensitive data in encrypted queues. SSE protects the contents of messages in Amazon SQS queues using keys managed in the AWS Key Management Service (AWS KMS). SSE encrypts messages as soon as Amazon SQS receives them. The messages are stored in encrypted form and Amazon SQS decrypts messages only when they are sent to an authorized consumer.

For more information, see Protecting Data Using Server-Side Encryption (SSE) and AWS KMS in the Amazon SQS Developer Guide.
Q: Can I use SNS, Cloud Watch Events and S3 Events with encrypted queues?

Yes. To do this you need to enable compatibility between AWS services (eg. Amazon CloudWatch Events, Amazon S3, and Amazon SNS), and Queues with SSE. For detailed instructions see the Compatibility section of the SQS Developer Guide.
Q: What regions are queues with SSE available in?

Server-side encryption (SSE) for Amazon SQS is available in all AWS regions where Amazon SQS is available. See here for details on Amazon SQS region availability.
Q: How do I enable SSE for a new or existing Amazon SQS queue?

To enable SSE for a new or existing queue using the Amazon SQS API, specify the customer master key (CMK) ID: the alias, alias ARN, key ID, or key ARN of an AWS-managed CMK or a custom CMK by setting the KmsMasterKeyId attribute of the CreateQueue or SetQueueAttributes action.

For detailed instructions, see Creating an Amazon SQS Queue with Server-Side Encryption and Configuring Server-Side Encryption (SSE) for an Existing Amazon SQS Queue in the *Amazon SQS Developer Guide*.
Q: What Amazon SQS queue types can use SSE?

Both standard and FIFO queues support SSE.
Q: What permissions do I need to use SSE with Amazon SQS?

Before you can use SSE, you must configure AWS KMS key policies to allow encryption of queues and encryption and decryption of messages.

To enable SSE for a queue, you can use the AWS-managed customer master key (CMK) for Amazon SQS or a custom CMK. For more information, see Customer Master Keys in the *AWS KMS Developer Guide*.

To send messages to an encrypted queue, the producer must have the kms:GenerateDataKey and kms:Decrypt permissions for the CMK.

To receive messages from an encrypted queue, the consumer must have the kms:Decrypt permission for any CMK that is used to encrypt the messages in the specified queue. If the queue acts as a dead letter queue, the consumer must also have the kms:Decrypt permission for any CMK that is used to encrypt the messages in the source queue.

For more information, see What Permissions Do I Need to Use SSE? in the *Amazon SQS Developer Guide*.
Q: Are there any charges for using SSE with Amazon SQS?

There are no additional Amazon SQS charges. However, there are charges for calls from Amazon SQS to AWS KMS. For more information, see AWS Key Management Service Pricing.

The charges for using AWS KMS depend on the data key reuse period configured for your queues. For more information, see How Do I Estimate My AWS KMS Usage Costs? in the *Amazon SQS Developer Guide*.
Q: What does SSE for Amazon SQS encrypt and how is it encrypted?

SSE encrypts the body of a message in an Amazon SQS queue.

SSE doesn't encrypt the following components:

- Queue metadata (queue name and attributes)
- Message metadata (message ID, timestamp, and attributes)
- Per-queue metrics

Amazon SQS generates data keys based on the AWS-managed customer master key (CMK) for Amazon SQS or a custom CMK to provide envelope encryption and decryption of messages for a configurable time period (from 1 minute to 24 hours).

For more information, see What Does SSE for Amazon SQS Encrypt? in the *Amazon SQS Developer Guide.*
Q: What algorithm does SSE for Amazon SQS use to encrypt messages?

SSE uses the AES-GCM 256 algorithm.
Q: Does SSE limit the transactions per second (TPS) or number of queues that can be created with Amazon SQS?

SSE doesn't limit the throughput (TPS) of Amazon SQS. The number of SSE queues that you can create is limited by the following:

- The data key reuse period (1 minute to 24 hours).
- The AWS KMS per-account quota (100 TPS by default).
- The number of IAM users or accounts that access queues.
- The existence of a large backlog (a larger backlog requires more AWS KMS calls).

For example, let's assume the following:

- You set your data key reuse period to 5 minutes (300 seconds).
- Your KMS account has a default AWS KMS TPS quota of 100 TPS.
- You use an Amazon SQS queue without a backlog and with 1 IAM user for SendMessage or ReceiveMessage actions to all queues.

In this case, you can calculate the theoretical maximum of Amazon SQS queues with SSE as follows:

300 seconds × 100 TPS / 1 IAM user = 30,000 queues
Q: How can I estimate my AWS KMS usage costs?

To predict costs and better understand your AWS bill, you might want to know how often Amazon SQS uses your CMK.

> Note: Although the following formula can give you a very good idea of expected costs, actual costs might be higher because of the distributed nature of Amazon SQS.

To calculate the number of API requests per queue (R), use the following formula:

$$R = B / D * (2 * P + C)$$

B is the billing period (in seconds)

D is the data key reuse period (in seconds)

P is the number of producing principals that send to the Amazon SQS queue.

C is the number of consuming principals that receive from the Amazon SQS queue.

> Important: In general, producing principals incur double the cost of consuming principals. For more information, see How Does the Data Key Reuse Period Work? in the *Amazon SQS Developer Guide*.
>
> If the producer and consumer have different IAM users, the cost increases.

For more information, see How Do I Estimate My AWS KMS Usage Costs? in the *Amazon SQS Developer Guide*

Compliance

Q: Is Amazon SQS PCI DSS certified?

Yes. Amazon SQS is PCI DSS Level 1 certified. For more information, see PCI Compliance.
Q: Is Amazon SQS HIPAA-eligible?

Yes, AWS has expanded its HIPAA compliance program to include Amazon SQS as a HIPAA Eligible Service. If you have an executed Business Associate Agreement (BAA) with AWS, you can use Amazon SQS to build your HIPAA-compliant applications, store messages in transit, and transmit messages—including messages containing protected health information (PHI).

If you already have an executed BAA with AWS, you can start using Amazon SQS right away. If you don't have a BAA or have other questions about using AWS for your HIPAA-compliant applications, contact us for more information.

Note: If you prefer not to transfer PHI through Amazon SQS (or if you have messages larger than 256 KB), you can alternatively send Amazon SQS message payloads through Amazon S3 using the Amazon SQS Extended Client Library for Java (Amazon S3 is a HIPAA Eligible Service, excluding the use of Amazon S3 Transfer Acceleration). For more information, see Using the Amazon SQS Extended Client Library for Java in the *Amazon SQS Developer Guide*.

Limits and restrictions

Q: How long can I keep my messages in Amazon SQS message queues?

Longer message retention provides greater flexibility to allow for longer intervals between message production and consumption.

You can configure the Amazon SQS message retention period to a value from 1 minute to 14 days. The default is 4 days. Once the message retention quota is reached, your messages are automatically deleted.
Q: How do I configure Amazon SQS to support longer message retention?

To configure the message retention period, set the MessageRetentionPeriod attribute using the console or using the Distributiveness method. Use this attribute to specify the number of seconds a message will be retained in Amazon SQS.

You can use the MessageRetentionPeriod attribute to set the message retention period from 60 seconds (1 minute) to 1,209,600 seconds (14 days). For more information on working with this message attribute, see the *Amazon SQS API Reference*.
Q: How do I configure the maximum message size for Amazon SQS?

To configure the maximum message size, use the console or the SetQueueAttributes method to set the MaximumMessageSize attribute. This attribute specifies the number of bytes that an Amazon SQS message can contain. Set this attribute to a value between 1,024 bytes (1 KB), and 262,144 bytes (256 KB). For more information, see Using Amazon SQS Message Attributes in the *Amazon SQS Developer Guide*.

To send messages larger than 256 KB, use the Amazon SQS Extended Client Library for Java. This library lets you send an Amazon SQS message that contains a reference to a message payload in Amazon S3 that can be as large as 2 GB.
Q: What kind of data can I include in a message?

Amazon SQS messages can contain up to 256 KB of text data, including XML, JSON and unformatted text. The following Unicode characters are accepted:

#x9 | #xA | #xD | [#x20 to #xD7FF] | [#xE000 to #xFFFD] | [#x10000 to #x10FFFF]

For more information, see the XML 1.0 Specification.
Q: How large can Amazon SQS message queues be?

A single Amazon SQS message queue can contain an unlimited number of messages. However, there is a quota of 120,000 for the number of inflight messages for a standard queue and 20,000 for a FIFO queue. Messages are inflight after they have been received from the queue by a consuming component, but have not yet been deleted from the queue.
Q: How many message queues can I create?

You can create any number of message queues.
Q: Is there a size limit on the name of Amazon SQS message queues?

Queue names are limited to 80 characters.
Q: Are there restrictions on the names of Amazon SQS message queues?

You can use alphanumeric characters, hyphens (-), and underscores (_).
Q: Can I reuse a message queue name?

A message queue's name must be unique within an AWS account and region. You can reuse a message queue's name after you delete the message queue.

Queue sharing

Q: How do I share a message queue?

You can associate an access policy statement (and specify the permissions granted) with the message queue to be shared. Amazon SQS provides APIs for creating and managing access policy statements:

- AddPermission
- RemovePermission
- SetQueueAttributes
- GetQueueAttributes

For more information, see the *Amazon SQS API Reference*.
Q: Who pays for shared queue access?

The message queue owner pays for shared message queue access.
Q: How do I identify another AWS user I want to share a message queue with?

The Amazon SQS API uses the AWS account number to identify AWS users.
Q: What do I need to provide to an AWS user I want to share a message queue with?

To share a message queue with an AWS user, provide the full URL from the message queue you want to share. The CreateQueue and ListQueues operations return this URL in their responses.
Q: Does Amazon SQS support anonymous access?

Yes. You can configure an access policy that allows anonymous users to access a message queue.
Q: When should I use the permissions API?

The permissions API provides an interface for sharing access to a message queue to developers. However, this API cannot allow conditional access or more advanced use cases.
Q: When should I use the SetQueueAttributes operation with JSON objects?

The SetQueueAttributes operation supports the full access policy language. For example, you can use the policy language to restrict access to a message queue by IP address and time of

day. For more information, see Amazon SQS Policy Examples in the *Amazon SQS Developer Guide*.

Service access and regions

Q: What regions is Amazon SQS available in?

For service region availability, see the AWS Global Infrastructure Region Table.
Q: Can I share messages between queues in different regions?

No. Each Amazon SQS message queue is independent within each region.
Q: Is there a pricing difference between regions?

Amazon SQS pricing is the same for all regions, except China (Beijing). For more information, see Amazon SQS Pricing.
Q: What is the pricing structure between various regions?

You can transfer data between Amazon SQS and Amazon EC2 or AWS Lambda free of charge within a single region.

When you transfer data between Amazon SQS and Amazon EC2 or AWS Lambda in different regions, you will be charged the normal data transfer rate. For more information, see Amazon SQS Pricing.

Dead-letter queues

Q: What are dead-letter queues?

A dead-letter queue is an Amazon SQS queue to which a source queue can send messages if the source queue's consumer application is unable to consume the messages successfully. Dead-letter queues make it easier for you to handle message consumption failures and manage the life cycle of unconsumed messages. You can configure an alarm for any messages delivered to a dead-letter queue, examine logs for exceptions that might have caused them to be delivered to the queue, and analyze message contents to diagnose consumer application issues. Once you recover your consumer application, you can redrive the messages from your dead-letter queue to the source queue.
Q: How do dead-letter queues work?

When you create your source queue, Amazon SQS allows you to specify a dead-letter queue (DLQ) and the condition under which SQS should move messages to the DLQ. The condition is the number of times a consumer can receive a message from the queue, defined as maxReceiveCount. This configuration of a dead-letter queue with a source queue and the maxReceiveCount is known as the redrive policy. When the ReceiveCount for a message exceeds the maxReceiveCount for a queue, Amazon SQS is designed to move the message to a dead-letter queue (with its original message ID). For example, if the source queue has a redrive policy with maxReceiveCount set to five, and the consumer of the source queue receives a message six times without successfully consuming it, SQS moves the message to the dead-letter queue.

The redrive policy manages the first half of the life cycle of unconsumed messages by moving them from a source queue to a dead-letter queue. Now the dead-letter queue redrive to source queue efficiently completes the cycle by moving those messages back to their source queue, as shown below.

Q: How does the dead-letter queue redrive to source queue work?

First, it allows you to investigate a sample of messages available in the dead-letter queue by showing message attributes and related metadata. Then, once you have investigated the messages, you can move them back to their source queue(s). You can also select the redrive velocity to configure the rate at which Amazon SQS will move the messages from the dead-letter queue to the source queue.

Q: Can I use a dead letter queue with FIFO queues?

Yes. However, you must use a FIFO dead letter queue with a FIFO queue. (Similarly, you can use only a standard dead letter queue with a standard queue.)

Amazon Simple Email Service (SES)

Get reliable, scalable email to communicate with customers at the lowest industry prices.

How it works

Amazon Simple Email Service (SES) lets you reach customers confidently without an on-premises Simple Mail Transfer Protocol (SMTP) system.

Why Amazon SES?

Amazon SES is a cloud email service provider that can integrate into any application for bulk email sending. Whether you send transactional or marketing emails, you pay only for what you use. Amazon SES also supports a variety of deployments including dedicated, shared, or owned IP addresses. Reports on sender statistics and a deliverability dashboard help businesses make every email count.

Use cases

Automate transactional messages
Keep your customers up to date by sending automated emails, such as purchase or shipping notifications, order status updates, and policy change notices.

Deliver marketing emails globally
Tell customers around the world about products and services through newsletters, special offers, and engaging content.

Send timely notifications to customers
Send customers timely notifications about their interaction with your products and services, including daily reminders, weekly usage reports, and newsletters.

Send bulk email communications
Deliver messages—including notifications and announcements—to large groups, and track results using configuration sets.

Amazon SES Features

Building a large-scale email solution can be a complex and costly challenge for a business: you have to build your infrastructure, configure your network, warm up your IP addresses and protect your sender reputation. Many third-party email solutions require contract negotiations and significant up-front costs.

Amazon SES eliminates these challenges, allowing you to start sending email in minutes. You benefit from the years of experience and the sophisticated email infrastructure Amazon.com built to serve its own large-scale customer base.

Sender Configuration Options
Amazon SES offers several methods of sending email, including the Amazon SES console, the Simple Mail Transfer Protocol (SMTP) interface, and the Amazon SES API. You can access the API using the AWS Command Line Interface (AWS CLI), or by using an AWS Software Development Kit (SDK).

Inbox Deliverability
Deliverability, or ensuring emails reach inboxes instead of spam or junk, is a core element of a successful email strategy. Amazon Simple Email Service (SES) offers deliverability features through Virtual Deliverability Manager (VDM) that make it easier and more automated to improve inbox placement.

SES provides insights into the performance of senders' emails, makes recommendations on how to improve deliverability, and can even automatically implement email deliverability improvements for senders. SES features three core deliverability components: insights, recommended changes, and automatic implementation

Insights dashboard
SES's deliverability insights help email senders understand their deliverability performance. View at-a-glance reports on sending and delivery data in a single interface in the SES console, such as bounce rate, opens, clicks and deliverability at ISP and configuration levels.

Configuration recommendations

SES can notify senders of deliverability issues and provide actionable recommendations to help improve email deliverability, like reviewing email authentication configuration such as SPF, DKIM, and DMARC. Senders have ongoing visibility if a downgrade in email reputation or improvement opportunity is identified.

Automatic implementation

Senders have the option to allow SES to automatically implement deliverability updates, like optimizing email delivery patterns. Once SES detects an opportunity for improvement, the system will implement the change without the need for monitoring or manual adjustment.

Flexible Deployment Options

Shared IP Addresses

By default, Amazon SES **sends** email from IP addresses that are shared with other Amazon SES customers. Shared addresses are a great option for many customers who want to start sending immediately with established IPs. They are included in the base Amazon SES pricing, and their reputations are carefully monitored to ensure high deliverability.

Dedicated IP Addresses

Customers can manage their own IP reputations by leasing dedicated IP addresses for use with their Amazon SES account. Customers can use the dedicated IP pools feature in SES console to create pools of those IP addresses, and either send all traffic from these dedicated IPs or use configuration sets to align specific use cases to specific IPs. Tailored for different customer use cases and sending patterns, SES offers two different ways to implement and manage dedicated IP addresses:

- Standard - dedicated IP addresses that customers manually setup and manage. SES provides known IP addresses for leases that never change. This standard option is good for customers with continuous, predictable sending patterns.

- Managed - dedicated IP address that are automatically set up and managed by SES on your behalf to help optimize customers' cost and usage. The managed dedicated IPs reduces the burden of manual monitoring or scaling of dedicated IP pools. Dedicated IPs (managed) are suitable for customers with a less predictable or non-continuous sending pattern.

Owned IP Addresses

Amazon SES also supports Bring Your Own IP (BYOIP). This feature lets you use a range of IP addresses that you already own to send email with Amazon SES. This makes leveraging current investments and migrating from other email service providers easy.

Sender Identity Management and Security

Amazon SES supports all industry-standard authentication mechanisms, including Domain Keys Identified Mail (DKIM), Sender Policy Framework (SPF), and Domain-based Message Authentication, Reporting and Conformance (DMARC). When an internet service provider (ISP) receives an email, they check to see if it is authenticated before attempting to deliver it to the recipient. Authentication demonstrates to the ISP that you own the email address you are sending from.

Amazon SES also enables customers to connect an Amazon SES SMTP endpoint to a virtual private cloud (VPC) through a VPC endpoint powered by AWS PrivateLink. With this feature,

customers can access the Amazon SES SMTP endpoint securely without requiring an Internet Gateway in a VPC.

Sending Statistics

Amazon SES provides a few methods for monitoring your email sending activity, helping you fine-tune your email sending strategy. Amazon SES can capture information about the entire email response funnel, including the numbers of sends, deliveries, opens, clicks, bounces, complaints, and rejections. This data is shared by default in the Sending Statistics report in the Amazon SES console. Use the Global suppression list to remove bounced emails from your sending list, or configure your own account-level suppression list. Sending data can be stored in an Amazon S3 bucket or an Amazon Redshift database, sent to Amazon SNS for real-time notifications, or analyzed using Amazon Kinesis Analytics.

Reputation Dashboard

The Amazon SES console includes a reputation dashboard that you can use to track issues that could impact the delivery of your emails. This dashboard tracks the overall bounce and feedback loops for your account, and can inform you when other deliverability-impacting events occur, such as spamtrap hits, references to blocked domains in your emails, and reports from reputable anti-spam organizations.

Amazon SES automatically publishes the bounce and complaint metrics from this dashboard to Amazon CloudWatch. You can use CloudWatch to create alarms that notify you when your bounce or complaint rates reach certain thresholds. With this information, you can take immediate action on issues that could impact your sender reputation.

Email Receiving

When you use Amazon SES to receive incoming emails, you have complete control over which emails you accept, and what to do with them after you receive them. You can accept or reject mail based on the email address, IP address, or domain of the sender. Once Amazon SES has accepted the email, you can store it in an Amazon S3 bucket, execute custom code using an AWS Lambda function, or publish notifications to Amazon SNS.

Mailbox Simulator

The Amazon SES mailbox simulator makes it easy to test how your application handles certain scenarios, such as bounces or complaints, without impacting your sender reputation. Using the mailbox simulator is as easy as sending a test email to a specific address. You can use the mailbox simulator to simulate successful deliveries, hard bounces, out-of-office responses or feedback.

Amazon SES API v2

This is information about the Amazon SES API v2, including supported operations, data types, parameters, and schemas.

Amazon SES is an AWS service that you can use to send email messages to your customers.

If you're new to Amazon SES API v2, you might find it helpful to review the Amazon Simple Email Service Developer Guide. The *Amazon SES Developer Guide* provides information and code samples that demonstrate how to use Amazon SES API v2 features programmatically.

The Amazon SES API v2 is available in several AWS Regions and it provides an endpoint for each of these Regions. For more information, see Amazon SES endpoints and quotas in the *Amazon Web Services General Reference*. To learn more about AWS Regions, see Managing AWS Regions in the *Amazon Web Services General Reference*.

What is Amazon SES?

Amazon SES is an email platform that provides an easy, cost-effective way for you to send and receive email using your own email addresses and domains.

For example, you can send marketing emails such as special offers, transactional emails such as order confirmations, and other types of correspondence such as newsletters. When you use Amazon SES to receive mail, you can develop software solutions such as email autoresponders, email unsubscribe systems, and applications that generate customer support tickets from incoming emails.

Benefits

Building a large-scale email solution is often a complex and costly challenge for a business. You must deal with infrastructure challenges such as email server management, network configuration, and IP address reputation. Additionally, many third-party email solutions require contract and price negotiations, as well as significant up-front costs. Amazon SES eliminates these challenges and enables you to benefit from the years of experience and sophisticated email infrastructure Amazon.com has built to serve its own large-scale customer base.

Related services

Amazon SES integrates seamlessly with other AWS products. For example, you can:

- Add email-sending capabilities to any application. If your application runs in Amazon Elastic Compute Cloud (Amazon EC2), you can use Amazon SES to send 62,000 emails every month at no additional charge. You can send

email from Amazon EC2 by using an AWS SDK, by using the Amazon SES SMTP interface, or by making calls directly to the Amazon SES API.

- Use AWS Elastic Beanstalk to create an email-enabled application such as a program that uses Amazon SES to send a newsletter to customers.
- Set up Amazon Simple Notification Service (Amazon SNS) to notify you of your emails that bounced, produced a complaint, or were successfully delivered to the recipient's mail server. When you use Amazon SES to receive emails, your email content can be published to Amazon SNS topics.
- Use the AWS Management Console to set up Easy DKIM, which is a way to authenticate your emails. Although you can use Easy DKIM with any DNS provider, it is especially easy to set up when you manage your domain with Route 53.
- Control user access to your email sending by using AWS Identity and Access Management (IAM).
- Store emails you receive in Amazon Simple Storage Service (Amazon S3).
- Take action on your received emails by triggering AWS Lambda functions.
- Use AWS Key Management Service (AWS KMS) to optionally encrypt the mail you receive in your Amazon S3 bucket.
- Use AWS CloudTrail to log Amazon SES API calls that you make using the console or the Amazon SES API.
- Publish your email sending events to Amazon CloudWatch or Amazon Kinesis Data Firehose. If you publish your email sending events to Kinesis Data Firehose, you can access them in Amazon Redshift, Amazon OpenSearch Service, or Amazon S3.

Amazon Simple Email Service (SES) FAQs

Getting Started

Q: What is an easy way to test Amazon SES?

The Amazon SES sandbox is an area where new users can test the capabilities of Amazon SES. When your account is in the sandbox, you can only send email to verified identities. A verified identity is an email addresses or domain that you've proven that you own.

Additionally, when your account is in the sandbox, there are limits to the volume of email you can send each day, and to the number of messages you can send each second.

Q: Can I start sending large email volumes right away?

When you're ready to start sending email to non-verified recipients, submit an Amazon SES Sending Limit Increase request through the AWS Support Center. For more information, see Moving Out of the Amazon SES Sandbox in the *Amazon SES Developer Guide*.

Limits and Restrictions

Q: Can I send emails from any email address?

No. You can only use Amazon SES to send email from addresses or domains that you own.

To prove that you own an email address or domain, you have to verify it. In each AWS Region, you can verify up to 10,000 email addresses and domains, in any combination. For more information about verifying email addresses and domains, see Verifying Identities in Amazon SES in the Amazon SES Developer Guide.

Q: Is there a limit on the size of emails Amazon SES can deliver?

Amazon SES v2 API and SMTP accepts email messages up to 40MB in size including any images and attachments that are part of the message. Messages larger than 10MB are subject to bandwidth throttling, and depending on your sending rate, you may be throttled to as low as 40MB/s. For example, you could send a 40MB message at the rate of 1 message per second, or two 20MB messages per second.

Amazon SES API v1 accepts messages up to 10MB in size including any images and attachments that are part of the message.

Q: Are there any limits on how many emails I can send?

Every Amazon SES account has its own set of sending limits. These limits are:

- Sending quota—the maximum number of recipients that you can send email to in a 24-hour period.
- Maximum send rate—the maximum number of recipients that you can send email per second.

Sending limits are based on recipients rather than on messages. You can check your sending limits at any time by using the Amazon SES console.

Note: If we determine that the email you send is of poor or questionable quality (for example, if it has high bounce or complaint rates, or if it contains unsolicited or malicious content), we reserve the right to pause your ability to send email.

Authentication, Validation, and Configuration

Q: Does Amazon SES support Sender Policy Framework (SPF)?

Yes, Amazon SES supports SPF. You may need to publish an SPF record, depending on how you use Amazon SES to send email. If you don't need to comply with Domain-based Message Authentication, Reporting and Conformance (DMARC) using SPF, you don't need to publish an SPF record, because by default, Amazon SES sends your emails from a MAIL FROM domain that's owned by Amazon Web Services. If you want to comply with DMARC using SPF, you have to set up Amazon SES to use your own MAIL FROM domain and publish an SPF record.

Q: Does Amazon SES support Domain Keys Identified Mail (DKIM)?

Yes, Amazon SES supports DKIM. If you have enabled and configured Easy DKIM, Amazon SES signs outgoing messages using DKIM on your behalf. If you prefer, you can also sign your email manually. To ensure maximum deliverability, there are a few DKIM headers that you should not sign. For more information, see Manual DKIM Signing in Amazon SES in the Amazon SES Developer Guide.

Q: Can emails from Amazon SES comply with DMARC?

With Amazon SES, your emails can comply with DMARC through SPF, DKIM, or both.

For more information, see Amazon SES and Security Protocols in the Amazon SES Developer Guide.

Q: Can I specify a dedicated IP address when I send certain types of email?

If you lease several dedicated IP addresses to use with your Amazon SES account, you can use the dedicated IP pools feature to create groups (pools) of those IP addresses. You can then associate each pool with a configuration set; when you send emails using that configuration set, those emails are only sent from the IP addresses in the associated pool. For more information, see Creating Dedicated IP Pools in the *Amazon SES Developer Guide*.

Q: What is the difference between dedicated IP address (standard) and (managed)?

Both Dedicated IP options help you manage your sending reputation through reserved IP addresses. Dedicated IP Addresses (standard) requires you to manually setup and manage your IP addresses. Dedicated IP Addresses (managed) reduces the need for manual monitoring or scaling of dedicated IP pools. It also helps you to model warmup status more accurately and prevent over-sending which can impact deliverability. For more detailed benefit comparison, see Dedicated IP addresses in the *Amazon SES Developer Guide.*

Sending Capabilities

Q: Does Amazon SES provide an SMTP endpoint?

Amazon SES provides an SMTP interface for seamless integration with applications that can send email via SMTP. You can connect directly to this SMTP interface from your applications, or configure your existing email server to use this interface as an SMTP relay.

In order to connect to the Amazon SES SMTP interface, you have to create SMTP credentials. For more information about creating SMTP credentials, see Obtaining Your Amazon SES SMTP Credentials in the *Amazon SES Developer Guide*.

Q: Can I use Amazon SES to send email from my existing applications?

Amazon SES allows you to create a private SMTP relay for use with any existing SMTP client software, including software that you develop yourself, or any third-party software that can send email using the SMTP protocol.

For more information, see Using the Amazon SES SMTP Interface to Send Email in the *Amazon SES Developers Guide*.

Q: Can Amazon SES send emails with attachments?

Amazon SES supports many popular content formats, including documents, images, audio, and video.

Note: For your own safety and that of your customers, Amazon SES scans every attachment that you send for viruses and malware.

You can use an email client that supports SMTP to send email with attachments. When you configure a client to send outgoing email through Amazon SES, the client constructs the appropriate MIME parts and email headers before sending the message.

You can also send email with attachments programmatically. To include an attachment in your email, construct a new multipart email message. In the message, include a MIME part that contains an appropriate Content-Type header, along with the MIME-encoded content. Next, use the Content-Disposition header to specify whether the content is to be displayed inline or treated as an attachment.

Once you've composed your message, you can use the SendRawEmail API operation to send it.

Q: Can I test Amazon SES responses without sending email to real recipients?

You can use the Amazon SES mailbox simulator to test your sending rate and to test your ability to handle events such as bounces and complaints, without sending email to actual recipients. Messages that you send to the mailbox simulator don't count against your bounce and complaint metrics or your daily sending quota. However, we do charge you for each message you send to the mailbox simulator, just as if they were messages you sent to actual customers.

For more information about the Amazon SES mailbox simulator, see Testing Amazon SES Email Sending in the Amazon SES Developer Guide.

Deliverability

Q: How does Amazon SES help ensure reliable email delivery?

Amazon SES uses content filtering technologies to scan outgoing email messages. These content filters help ensure that the content being sent through Amazon SES meets the standards of ISPs. In order to help you further improve the deliverability of your emails, Amazon SES provides a feedback loop that includes bounce, complaint, and delivery notifications.

Q: Does Amazon SES guarantee receipt of my emails?

Amazon SES closely monitors ISP guidelines to help ensure that legitimate, high-quality email is delivered reliably to recipient inboxes. However, neither Amazon SES nor any other email-sending service can guarantee delivery of every single email. ISPs can drop or lose email messages, recipients can accidentally provide the wrong email address, and if recipients do not wish to receive your email messages, ISPs may choose to reject or silently drop them.

Q: How long does it take for emails sent using Amazon SES to arrive in recipients' inboxes?

Amazon SES attempts to deliver emails to the Internet within a few seconds of each request. However, due to a number of factors and the inherent uncertainties of the Internet, we can't predict with certainty when your email will arrive, nor can we predict the exact route the message will take to get to its destination.

For example, an ISP might be unable to deliver the email to the recipient because of a temporary condition such as "mailbox full." In these cases, Amazon attempts to redeliver the message. If the error is permanent, such as "mailbox does not exist," Amazon SES doesn't try to deliver the message again, and you receive a hard bounce notification. You can set up delivery notifications to alert you when Amazon SES successfully delivers one of your emails to a recipient's mail server.

Q: Can my email deliverability affected by bounces or complaints that are caused by other Amazon SES users?

Typically, when other Amazon SES users send messages that result in bounces or complaints, your ability to send email remains unchanged.

An exception to this rule occurs when a recipient's email address generates a hard bounce. When a recipient's email address generates a hard bounce, Amazon SES adds that address to a global suppression list. If you try to send an email to an address that is on the global suppression list, the call to Amazon SES succeeds, but Amazon SES treats the email as a hard bounce instead of attempting to send it.

Emails that you send to addresses on the global suppression list count toward your sending quota and your bounce rate. An email address can remain on the suppression list for up to 14 days.

For more information about the global suppression list, see Amazon SES and Deliverability in the *Amazon SES*

Security and Encryption

Q: Can Amazon access the emails that I send and receive?

We use in-house anti-spam technologies to filter messages that contain poor-quality content. Additionally, we scan all messages that contain attachments to check for viruses and other malicious content.

Q: Can I encrypt email messages that I receive?

Amazon SES integrates with AWS Key Management Service (KMS), which provides the ability to encrypt the mail that it writes to your Amazon S3 bucket. Amazon SES uses client-side encryption to encrypt your mail before it sends the email to Amazon S3. This means that it is necessary for you to decrypt the content on your side after you retrieve the mail from Amazon S3. The AWS Java SDK and AWS Ruby SDK provide a client that is able to handle the decryption for you.

Q: Does Amazon SES send email over an encrypted connection using Transport Layer Security (TLS)?

Amazon SES supports TLS 1.2, TLS 1.1 and TLS 1.0 for TLS connections.

By default, Amazon SES uses *opportunistic TLS*. This means that Amazon SES always attempts to make a secure connection to the receiving mail server. If Amazon SES can't establish a secure connection, it sends the message unencrypted.

You can change this behavior so that Amazon SES only sends the message to the receiving email server if it can establish a secure connection

Q: How does Amazon SES ensure that incoming mail is free of spam and viruses?

Amazon SES uses a number of spam and virus protection measures. It uses block lists to prevent mail from known spammers from entering the system in the first place. It also performs virus scans on every incoming email that contains an attachment. Amazon SES makes its spam detection verdicts available to you, enabling you to decide if you trust each message. In addition to the spam and virus verdicts, Amazon SES provides the DKIM and SPF check results.

Q: What prevents Amazon SES users from sending spam?

Amazon SES uses in-house content filtering technologies to scan email content for spam and malware.

If we determine that an account is sending spam or malicious content, we will pause that account's ability to send additional email.

Amazon SES and Other AWS Services

Q: How is Amazon SES different from Amazon SNS?

Amazon SES is for applications that need to send communications via email. Amazon SES supports custom email header fields, and many MIME types.

By contrast, Amazon Simple Notification Service (Amazon SNS) is for messaging-oriented applications, with multiple subscribers requesting and receiving "push" notifications of time-critical messages via a choice of transport protocols, including HTTP, Amazon SQS, and email. The body of an Amazon SNS notification is limited to 8192 characters of UTF-8 strings, and isn't intended to support multimedia content.

Q: Do I need to sign up for Amazon EC2 or any other AWS services to use Amazon SES?

Amazon SES users do not need to sign up for any other AWS services. Any application with Internet access can use Amazon SES to deliver email, whether that application runs in your own data center, within Amazon EC2, or as a client software solution.

AWS CloudFormation

Speed up cloud provisioning with infrastructure as code.

How it works

AWS CloudFormation lets you model, provision, and manage AWS and third-party resources by treating infrastructure as code.

Use cases

Manage infrastructure with DevOps

Automate, test, and deploy infrastructure templates with continuous integration and delivery (CI/CD) automations.

Scale production stacks

Run anything from a single Amazon Elastic Compute Cloud (EC2) instance to a complex multi-region application.

Share best practices

Define an Amazon Virtual Private Cloud (VPC) subnet or provisioning services like AWS OpsWorks or Amazon Elastic Container Service (ECS) with ease.

AWS CloudFormation Features

AWS CloudFormation gives you an easy way to model a collection of related AWS and third-party resources, provision them quickly and consistently, and manage them throughout their lifecycles, by treating infrastructure as code. A CloudFormation template describes your desired resources and their dependencies so you can launch and configure them together as a stack. You can use a template to create, update, and delete an entire stack as a single unit, as often as you need to, instead of managing resources individually. You can manage and provision stacks across multiple AWS accounts and AWS Regions.

Extensibility

Using the AWS CloudFormation Registry, you can model and provision third-party resources and modules published by AWS Partner Network (APN) Partners and the developer community. Examples of third-party resources are monitoring, team productivity, incident management, and version control tools, along with resources from APN Partners such as MongoDB, Datadog, Atlassian Opsgenie, JFrog, Trend Micro, Splunk, Aqua Security, FireEye, Sysdig, Snyk, Check Point, Spot by NetApp, Gremlin, Stackery, and Iridium. You can also browse, discover, and choose from a collection of pre-built modules by JFrog and Stackery, along with those maintained by AWS Quick Starts.

You can build your own resource providers using the AWS CloudFormation CLI, an open-source tool that streamlines the development process, including local testing and code generation capabilities.

Cross account & cross-region management

CloudFormation StackSets let you provision a common set of AWS resources across multiple accounts and regions, with a single CloudFormation template. StackSets takes care of automatically and safely provisioning, updating, or deleting stacks, no matter where they are.

Authoring with JSON/YAML

CloudFormation allows you to model your entire cloud environment in text files. You can use open-source declarative languages, such as JSON or YAML, to describe what AWS resources you want to create and configure. If you prefer to design visually, you can use AWS CloudFormation Designer to help you get started with AWS CloudFormation templates.

Authoring with familiar programming languages

With the AWS Cloud Development Kit (AWS CDK), you can define your cloud environment using TypeScript, Python, Java, and .NET. AWS CDK is an open-source software development framework that helps you model cloud application resources using familiar programming languages, and then provision your infrastructure using CloudFormation directly from your IDE. CDK provides high-level components that preconfigure cloud resources with proven defaults, so you can build cloud applications without needing to be an expert. Learn more about AWS CDK.

Build serverless applications with SAM

Build serverless applications faster with the AWS Serverless Application Model (SAM), an open-source framework that provides shorthand syntax to express functions, APIs, databases, and event source mappings. With just a few lines per resource, you can define the application you want and model it using YAML. During deployment, SAM transforms and expands the SAM syntax into CloudFormation syntax.

Safety controls

CloudFormation automates provisioning and updating your infrastructure in a safe and controlled manner. There are no manual steps or controls that can lead to errors. You can use Rollback Triggers to specify the CloudWatch alarms that CloudFormation should monitor during the stack creation and update process. If any of the alarms are triggered, CloudFormation rolls back the entire stack operation to a previously deployed state.

Using ChangeSets, you can preview the proposed changes that CloudFormation intends to make to your infrastructure and application resources prior to execution, so that your deployments go exactly as planned. CloudFormation determines the right operations to perform, provisions resources in the most efficient way possible, and rolls back automatically if errors are encountered. This returns the state of your infrastructure and application resources to the last known good state. Using Drift Detection, you can keep track of changes to resources outside CloudFormation, making sure you always have the most up-to-date picture of your infrastructure.

Preview changes to your environment

AWS CloudFormation Change Sets allow you to preview how proposed changes to a stack might affect your running resources, for example to check whether your changes will delete or replace any critical resources. CloudFormation makes the changes to your stack only after you decide to execute the Change Set.

Dependency management

AWS CloudFormation automatically manages dependencies between your resources during stack management actions. You don’t need to worry about specifying the order in which resources are created, updated, or deleted; CloudFormation determines the correct sequence of actions to take for each resource when performing stack operations.

What is AWS CloudFormation?

AWS CloudFormation is a service that helps you model and set up your AWS resources so that you can spend less time managing those resources and more time focusing on your applications that run in AWS. You create a template that describes all the AWS resources that you want (like Amazon EC2 instances or Amazon RDS DB instances), and CloudFormation takes care of provisioning and configuring those resources for you. You don't need to individually create and configure AWS resources and figure out what's dependent on what; CloudFormation handles that. The following scenarios demonstrate how CloudFormation can help.

Simplify infrastructure management

For a scalable web application that also includes a backend database, you might use an Auto Scaling group, an Elastic Load Balancing load balancer, and an Amazon Relational Database Service database instance. You might use each individual service to provision these resources and after you create the resources, you would have to configure them to work together. All these tasks can add complexity and time before you even get your application up and running.

Instead, you can create a CloudFormation template or modify an existing one.
A *template* describes all your resources and their properties. When you use that template to create a CloudFormation stack, CloudFormation provisions the Auto Scaling group, load balancer, and database for you. After the stack has been successfully created, your AWS resources are up and running. You can delete the stack just as easily, which deletes all the resources in the stack. By using CloudFormation, you easily manage a collection of resources as a single unit.

Quickly replicate your infrastructure

If your application requires additional availability, you might replicate it in multiple regions so that if one region becomes unavailable, your users can still use your application in other regions. The challenge in replicating your application is that it also requires you to replicate your resources. Not only do you need to record all the resources that your application requires, but you must also provision and configure those resources in each region.

Reuse your CloudFormation template to create your resources in a consistent and repeatable manner. To reuse your template, describe your resources once and then provision the same resources over and over in multiple regions.

Easily control and track changes to your infrastructure

In some cases, you might have underlying resources that you want to upgrade incrementally. For example, you might change to a higher performing instance type in your Auto Scaling launch configuration so that you can reduce the maximum number of instances in your Auto Scaling group. If problems occur after you complete the update, you might need to roll back your infrastructure to the original settings. To do this manually, you not only have to remember which resources were changed, you also have to know what the original settings were.

When you provision your infrastructure with CloudFormation, the CloudFormation template describes exactly what resources are provisioned and their settings. Because these templates are text files, you simply track differences in your templates to track changes to your infrastructure, similar to the way developers control revisions to source code. For example, you

can use a version control system with your templates so that you know exactly what changes were made, who made them, and when. If at any point you need to reverse changes to your infrastructure, you can use a previous version of your template.

AWS CloudFormation FAQs

General

Q: What is AWS CloudFormation?

AWS CloudFormation is a service that gives developers and businesses an easy way to create a collection of related AWS and third-party resources, and provision and manage them in an orderly and predictable fashion.

Q: What can developers do with AWS CloudFormation?

Developers can deploy and update compute, database, and many other resources in a simple, declarative style that abstracts away the complexity of specific resource APIs. AWS CloudFormation is designed to allow resource lifecycles to be managed repeatably, predictable, and safely, while allowing for automatic rollbacks, automated state management, and management of resources across accounts and regions. Recent enhancements and options allow for multiple ways to create resources, including using AWS CDK for coding in higher-level languages, importing existing resources, detecting configuration drift, and a new Registry that makes it easier to create custom types that inherit many core CloudFormation benefits.

Q: How is CloudFormation different from AWS Elastic Beanstalk?

These services are designed to complement each other. AWS Elastic Beanstalk provides an environment where you can easily deploy and run applications in the cloud. It is integrated with developer tools and provides a one-stop experience for managing application lifecycle. If your application workloads can be managed as Elastic Beanstalk workloads, you can enjoy a more turn-key experience in creating and updating applications. Behind the scenes, Elastic Beanstalk uses CloudFormation to create and maintain resources. If your application requirements dictate more custom control, the additional functionality of CloudFormation gives you more options to control your workloads.

AWS CloudFormation is a convenient provisioning mechanism for a broad range of AWS and third-party resources. It supports the infrastructure needs of many different types of applications such as existing enterprise applications, legacy applications, applications built using a variety of AWS resources, and container-based solutions (including those built using AWS Elastic Beanstalk).

AWS CloudFormation supports Elastic Beanstalk application environments as one of the AWS resource types. This allows you, for example, to create and manage an AWS Elastic Beanstalk–hosted application along with an RDS database to store the application data. Any other supported AWS resource can be added to the group as well.

Q: What new concepts does AWS CloudFormation introduce?

CloudFormation introduces four concepts: A template is a JSON or YAML declarative code file that describes the intended state of all the resources you need to deploy your application. A stack implements and manages the group of resources outlined in your template, and allows the state and dependencies of those resources to be managed together. A change set is a

preview of changes that will be executed by stack operations to create, update, or remove resources. A stack set is a group of stacks you manage together that can replicate a group.

Q: What resources does AWS CloudFormation support?

To see a complete list of supported AWS resources and their features, visit the Supported AWS Services page in the Release History of the documentation.

The AWS CloudFormation Registry and AWS CloudFormation custom resources enable management of additional AWS and third party resources.

Q: Can I manage individual AWS resources that are part of an AWS CloudFormation stack?

Yes, you can. CloudFormation does not get in the way; you retain full control of all elements of your infrastructure, and can continue using all your existing AWS and third-party tools to manage your AWS resources. However, because CloudFormation can allow for additional rules, best practices, and compliance controls, we recommend that you allow CloudFormation to manage the changes to your resources. This predictable, controlled approach helps in managing hundreds or thousands of resources across your application portfolio.

Q: What are the elements of an AWS CloudFormation template?

CloudFormation templates are JSON or YAML-formatted text files comprised of five types of elements:

1. An optional list of template parameters (input values supplied at stack creation time)
2. An optional list of output values (e.g., the complete URL to a web application)
3. An optional list of data tables used to look up static configuration values (e.g., AMI names)
4. The list of AWS resources and their configuration values
5. A template file format version number

Template parameters are used to customize aspects of your template at run time, when the stack is built. For example, the Amazon RDS database size, Amazon EC2 instance types, database and web server port numbers can be passed to AWS CloudFormation when a stack is created. Each parameter can have a default value and description, and may be marked as “NoEcho” to hide the actual value you enter on the screen and in the AWS CloudFormation event logs. When you create an AWS CloudFormation stack, the AWS Management Console will automatically synthesize and present a pop-up dialog form for you to edit parameter values.

Output values are a convenient way to present a stack’s key resources (such as the address of an Elastic Load Balancing load balancer or Amazon RDS database) to the user via the AWS Management Console, or via the command line tools. You can use simple functions to concatenate string literals and the value of attributes associated with the actual AWS resources. A template can also leverage Registry resource types, your own custom private types, your own macros, and retrieving configuration parameters from AWS Secrets Manager and AWS System Manager Parameter Store.

Q: How does AWS CloudFormation choose actual resource names?

You can assign logical names to AWS resources in a template. When a stack is created, AWS CloudFormation binds the logical name to the name of the corresponding actual AWS resource. Actual resource names are a combination of the stack and logical resource name. This allows multiple stacks to be created from a template without fear of name collisions between AWS resources.

Q: Why can’t I name all my resources?

Although AWS CloudFormation allows you to name some resources (such as Amazon S3 buckets), CloudFormation doesn't allow this for all resources. Naming resources restricts the reusability of templates and results in naming conflicts when an update causes a resource to be replaced. To minimize these issues, CloudFormation supports resource naming on a case by case basis.

Q: Can I install software at stack creation time using AWS CloudFormation?

Yes. AWS CloudFormation provides a set of application bootstrapping scripts that enable you to install packages, files, and services on your EC2 instances simply by describing them in your CloudFormation template. For more details and a how-to, see Bootstrapping Applications via AWS CloudFormation.

CloudFormation can also be integrated with Systems Manager to drive and maintain software installations with Systems Manager Automation Documents.

Q: Can I use AWS CloudFormation with Chef?

Yes. AWS CloudFormation can be used to bootstrap both the Chef Server and Chef Client software on your EC2 instances. For more details and a how-to, see Integrating AWS CloudFormation with Chef.

Q: Can I use AWS CloudFormation with Puppet?

Yes. AWS CloudFormation can be used to bootstrap both the Puppet Master and Puppet Client software on your EC2 instances. For more details and a how-to, see Integrating AWS CloudFormation with Puppet.

Q: Can I use AWS CloudFormation with Terraform?

Yes. CloudFormation can bootstrap your Terraform engine on your EC2 instances, and you can use Terraform resource providers to create resources in stacks, leveraging stack state management, dependencies, stabilization and rollback.

Q: Does AWS CloudFormation support Amazon EC2 tagging?

Yes. Amazon EC2 resources that support the tagging feature can also be tagged in an AWS template. The tag values can refer to template parameters, other resource names, resource attribute values (e.g. addresses), or values computed by simple functions (e.g., a concatenated a list of strings). CloudFormation automatically tags Amazon EBS volumes and Amazon EC2 instances with the name of the CloudFormation stack they are part of.

Q: Do I have access to the Amazon EC2 instance, or Auto Scaling Launch Configuration user-data fields?

Yes. You can use simple functions to concatenate string literals and attribute values of the AWS resources and pass them to user-data fields in your template. Please refer to our sample templates to learn more about these easy to use functions.

Q: What happens when one of the resources in a stack cannot be created successfully?

By default, the "automatic rollback on error" feature is enabled. This will direct CloudFormation to only create or update all resources in your stack if all individual operations succeed. If they do not, CloudFormation reverts the stack to the last known stable configuration. This is useful when, for example, you accidentally exceed your default limit of Elastic IP addresses, or you don't have access to an EC2 AMI that you're trying to run. This feature enables you to rely on

the fact that stacks are created either fully or not at all, which simplifies system administration and layered solutions built on top of CloudFormation.

Q: Can stack creation wait for my application to start up?

Yes. One of the options CloudFormation provides is a *WaitCondition* resource that acts as a barrier, blocking the creation of other resources until a completion signal is received from an external source such as your application or management system. Other options include creating custom logic with AWS Lambda functions.

Q: Can I save my data when a stack is deleted?

Yes. CloudFormation allows you to define deletion policies for resources in the template. You can specify that snapshots be created for Amazon EBS volumes or Amazon RDS database instances before they are deleted. You can also specify that a resource should be preserved and not deleted when the stack is deleted. This is useful for preserving Amazon S3 buckets when the stack is deleted.

Q: Can I update my stack after it has been created?

Yes. You can use CloudFormation to modify and update the resources in your existing stacks in a controlled and predictable way. By using templates to manage your stack changes, you have the ability to apply version control to your AWS infrastructure just as you do with the software running on it.

Q: Can I create stacks in a Virtual Private Cloud (VPC)?

Yes. CloudFormation supports creating VPCs, subnets, gateways, route tables and network ACLs as well as creating resources such as elastic IPs, Amazon EC2 Instances, EC2 security groups, auto scaling groups, elastic load balancers, Amazon RDS database instances and Amazon RDS security groups in a VPC.

Q: How can I participate in the CloudFormation community?

Please join the AWS CloudFormation GitHub community.

Q: Can I manage resources created outside of CloudFormation?

Yes! With Resource Import, you can bring an existing resource into AWS CloudFormation management using resource import.

Getting Started

Q: How do I sign up for AWS CloudFormation?

To sign up for CloudFormation, click Create Free Account on the CloudFormation product page. After signing up, please refer to the CloudFormation documentation, which includes our Getting Started Guide.

Q: Why am I asked to verify my phone number when signing up for AWS CloudFormation?

CloudFormation registration requires you to have a valid phone number and email address on file with AWS in case we ever need to contact you. Verifying your phone number takes only a few minutes and involves receiving an automated phone call during the registration process and entering a PIN number using the phone key pad.

Q: How do I get started after I have signed up?

The best way to get started with CloudFormation is to work through the Getting Started Guide, which is included in our technical documentation. Within a few minutes, you will be able to deploy and use one of our sample templates that illustrate how to create the infrastructure needed to run applications such as such as WordPress. There are various other sources of CloudFormation training, from thirdsparty curriculum providers to tutorials and articles on the web. For more information, check out the CloudFormation Resources.

Q: Are there sample templates that I can use to check out AWS CloudFormation?

Yes, CloudFormation includes sample templates that you can use to test drive the offering and explore its functionality. Our sample templates illustrate how to interconnect and use multiple AWS resources in concert, following best practices for multiple Availability Zone redundancy, scale out, and alarming. To get started, all you need to do is go to the AWS Management Console, click Create Stack, and follow the steps to select and launch one of our samples. Once created, select your stack in the console and review the Template and Parameter tabs to look at the details of the template file used to create the respective stack. Sample templates are also available on GitHub.

AWS CloudFormation Registry

Q: What is the AWS CloudFormation Registry?

The AWS CloudFormation Registry is a managed service that lets you register, use, and discover AWS and third-party resource types. Third-party resource types must be registered before they can be used to provision resources with AWS CloudFormation templates. Please see Using the AWS CloudFormation registry in our in the documentation for details.

Q: What are resource types in AWS CloudFormation?

A resource provider is a set of resource types with specifications and handlers that control the lifecycle of underlying resources via create, read, update, delete, and list operations. You can use resource providers to model and provision resources using CloudFormation. For example, AWS::EC2::Instance is a resource type from the Amazon EC2 provider. You can use this type to model and provision an Amazon EC2 instance using CloudFormation. Using the CloudFormation Registry, you can build and use resource providers to model and provision third-party resources such as SaaS monitoring, team productivity, or source code management resources.

Q: What is the difference between AWS and third party resource providers?

The difference between AWS and third party resource providers is their origin. AWS resource providers are built and maintained by Amazon and AWS to manage AWS resources and services. For example, three AWS resource providers help you manage Amazon DynamoDB, AWS Lambda, and Amazon EC2 resources. These providers contain resource types such as AWS::DynamoDB::Table, AWS::Lambda::Function, and AWS::EC2::Instance. For a complete reference, go to our documentation.

Third party resource providers are built by another company, organization, or the developer community. They can help you manage both AWS and non-AWS resources such as AWS application resources and non-AWS SaaS software services such as monitoring, team productivity, incident management, or version control management tools.

Q: What is a resource schema?

A resource schema defines a resource type in a structured and consistent format. This schema is also used to validate the definition of a resource type. The schema includes all the supported

parameters and attributes for a given resource type, as well as the required permissions to create the resource with the least privileges possible.

Q: How do I develop resource types?

Use the AWS CloudFormation CLI to build resource providers. You start by defining a simple declarative schema for your resources, which includes permissions required and relationships to other resources. You then use the CloudFormation CLI to generate the scaffolding for resource lifecycle handlers (Create, Read, Update, Delete, and List), along with test stubs for unit and integration testing.

Q: How do I register a resource provider?

You can either use the open source AWS CloudFormation CLI or directly call the RegisterType and related Registry APIs available via the AWS SDKs and AWS CLI. For more details, please see Using the AWS CloudFormation registry in our in the documentation. AWS resource providers are available out of the box and do not require any additional registration steps before use.

AWS CloudFormation Public Registry

Q: How does CloudFormation Public Registry relate to the CloudFormation Registry?

The CloudFormation Registry launched in November 2019 consisted of a private listing, allowing customers to extend CloudFormation for their own private use. The Public Registry extends the CloudFormation Registry and adds a public, searchable, central location for sharing, finding, consuming, and managing Resource Types and Modules <>, making it that much easier to configure and manage infrastructure and applications in a consistent manner for both AWS and third-party products.

Q: Is there a cost for using third-party Resource Types available on the CloudFormation Public Registry?

Yes.

Q: Does AWS verify publishers of third-party extensions on the CloudFormation Public Registry?

Yes. In the CloudFormation Public Registry, you have access to curated content from verified publishers. First, we verify each publisher's identity using either AWS Marketplace or third-parties such as GitHub and Bitbucket.

Q: What is the AWS CloudFormation Public Registry

CloudFormation Public Registry is a new searchable and managed catalog of extensions that contains resource types (provisioning logic) and modules published by AWS Partner Network (APN) Partners and the developer community. With CloudFormation Public Registry, anyone can now publish resource types and Modules on the Registry. Customers can easily discover and use these published resource types and modules, eliminating the need to build and maintain themselves.

Q: What is the difference between a resource and a module?

A Resource Type is a code package containing provisioning logic, which allows you to manage the lifecycle of a resource like an Amazon EC2 Instance or an Amazon DynamoDB Table from creation to deletion, abstracting away complex API interactions. Resource Types contain a

schema, which defines the shape and properties of a resource, and the necessary logic to provision, update, delete, and describe a resource. An example third-party Resource Type in the CloudFormation Public Registry is a Datadog monitor, MongoDB Atlas Project, or an Atlassian Opsgenie User among others.

Modules are building blocks that can be reused across multiple CloudFormation templates and is used just like a native CloudFormation resource. These building blocks can be for a single resource, like best practices for defining an Amazon Elastic Compute Cloud (Amazon EC2) instance or they can be for multiple resources, to define common patterns of application architecture.

Q: How do I develop and add my own resource or module to the AWS CloudFormation Registry?

You can refer to this link to develop and add your own resource or module to the AWS CloudFormation Registry. You can choose to publish it privately or to the Public Registry.

Limits and Restrictions

Q: Are there limits to the number of templates or stacks?

For more information on the maximum number of AWS CloudFormation stacks that you can create, see Stacks in AWS CloudFormation quotas. Complete our request for a higher limit here, and we will respond to your request within two business days.

Q: Are there limits to the size of description fields?

For more information, see Template Description in AWS CloudFormation quotas and Parameters, Resources and Outputs in the AWS documentation.

Q: Are there limits to the number of parameters or outputs in a template?

For more information on the number of parameters and outputs you can specify in a template, see Parameters and Outputs sections in AWS CloudFormation quotas.

Q: Are there limits to the number of resources that can be created in a stack?

For more information on the number of resources you can declare in a template, see Resources in AWS CloudFormation quotas. Creating smaller templates and stacks and modularizing your application across multiple stacks is a best practice to minimize blast radius for your resource changes, and to troubleshoot issues with multiple resource dependencies faster, since smaller groups of resources will have less complex dependencies than larger groups.

Regions and Endpoints

Q: What are the AWS CloudFormation service access points in each region?

Endpoints for each region are available in AWS CloudFormation endpoints in the technical documentation.

Q: What are the AWS regions where AWS CloudFormation is currently available?

Please refer to Regional Products and Services for details of CloudFormation availability by region.

AWS Amplify

Build full-stack web and mobile apps in hours. Easy to start, easy to scale.

How it works

AWS Amplify is a complete solution that lets frontend web and mobile developers easily build, ship, and host full-stack applications on AWS, with the flexibility to leverage the breadth of AWS services as use cases evolve. No cloud expertise needed.

Create an app backend

Create a cross-platform backend for your iOS, Android, Flutter, web, or React Native app with real-time and offline functionality in just a few clicks.

Build a frontend UI

Visually build a pixel-perfect UI and connect your frontend UI to a cloud backend in clicks. AWS Amplify saves you from writing thousands of lines of code.

Host a web app

Easily deploy your web app or website to the fast, secure, and reliable AWS content delivery network (CDN), with hundreds of points of presence globally.

AWS Amplify features

Build

Use the Amplify Studio or Amplify CLI to configure an app backend, and use the Amplify libraries and UI components to connect your app to your backend.

Authentication

Create seamless on-boarding flows with a fully-managed user directory and pre-built sign-up, sign-in, forgot password, and multi-factor auth workflows. Amplify also supports login with a social provider such as Facebook, Google Sign-In, or Login With Amazon and provides fine grained access control to mobile and web applications. Powered by Amazon Cognito.

DataStore

Use a multi-platform (iOS/Android/React Native/Web) on-device persistent storage engine that automatically synchronizes data between mobile/desktop/web apps and the cloud, powered by GraphQL. DataStore provides a programming model for leveraging shared and distributed data without writing additional code for offline and online scenarios, which makes working with distributed, cross-user data just as simple as working with local-only data. Powered by AWS AppSync.

Analytics

Understand the behavior of your web, iOS or Android users. Use auto tracking to track user sessions and web page metrics or create custom user attributes and in-app metrics. Get access to real time data stream and analyze the data for customer insights and build data driven marketing strategies to drive customer adoption, engagement, and retention. Powered by Amazon Pinpoint and Amazon Kinesis.

API

Make secure HTTP requests to GraphQL and REST endpoints to access, manipulate, and combine data from one or more data sources such Amazon DynamoDB, Amazon Aurora Serverless, and your custom data sources with AWS Lambda. Amplify enables you to easily build scalable applications that require real-time updates, local data access for offline scenarios, and data synchronization with customizable conflict resolution when devices are back online. Powered by AWS AppSync and Amazon API Gateway.

Functions

Add a Lambda function to your project which you can use alongside a REST API or as a datasource in your GraphQL API using the @function directive in the Amplify CLI. You can update the Lambda execution role policies for your function to access other resources generated and maintained by the CLI, using the CLI. Amplify CLI enables you to create, test and deploy Lambda functions across various runtimes and once a runtime is selected, you can select a function template for the runtime to help bootstrap your Lambda function.

Geo

Add location-aware features like maps and location search to your JavaScript-based web app in minutes. Amplify Geo includes pre-integrated map UI components (based on the popular MapLibre open-source library) and it updates the Amplify Command Line Interface (CLI) tool with support for provisioning all required cloud location services. You can customize embedded maps to match your app's theme, or choose from many community-developed MapLibre plugins for more flexibility and advanced visualization options. Powered by Amazon Location Service.

Interactions

Build interactive and engaging conversational bots with the same deep learning technologies that power Amazon Alexa with just a single line of code. Create great user experiences through chat bots when it comes to tasks such as automated customer chat support, product information/recommendations or streamlining common work activities etc. Powered by Amazon Lex.

Predictions

Enhance your app by adding AI/ML capabilities. You can easily achieve use cases like text translation, speech generation from text, entities recognition in image, interpretation of text, and transcribing text. Amplify enables simplified orchestration of advanced use cases like uploading images for automatic training and using GraphQL directives for chaining multiple AI/ML actions. Powered by Amazon Machine Learning services, such as Amazon Sagemaker.

PubSub

Pass messages between your app instances and your app's backend creating real-time interactive experiences. Amplify provides connectivity with cloud-based message-oriented middleware. Powered by AWS IoT services and Generic MQTT Over WebSocket Providers.

Push notifications

Improve customer engagement by using marketing and analytics capabilities. Leverage customer insights to segment and target your customers more effectively. You can tailor your content and communicate through multiple channels including email, texts as well as push notifications. Powered by Amazon Pinpoint.

Storage

Store and manage user generated content such as photos, videos securely on device or in the cloud. The AWS Amplify Storage module provides a simple mechanism for managing user content for your app in public, protected or private storage buckets. Leverage cloud scale storage so that you can easily take your application from prototype to production. Powered by Amazon S3.

Ship

Use the Amplify console or CLI to host static web applications.

Amplify Hosting

AWS Amplify offers a fully managed service for deploying and hosting fullstack web applications, with built-in CI/CD workflows that accelerate your application release cycle. A fullstack serverless app consists of a backend built with cloud resources such as GraphQL or REST APIs, file and data storage, and a frontend built with single page application frameworks such as React, Angular, Vue, or Gatsby. Simply connect your application's code repository in the Amplify console, and changes to your frontend and backend are deployed in a single workflow on every code commit.

Scale and manage

Use Amplify Studio to manage app users and content.

User management

Amplify Studio offers a way to manage authenticated users. Create/edit users and groups, update user properties, auto-verify signups, and more, without going through verification flows.

Content management

Developers can give content editors and testers access to update the app data via Amplify Studio. Admins can save content as markdown to render rich text.

Extensibility

Override generated resources

Modify fine-grained backend resource settings and override via CDK. Amplify does the heavy lifting fo your. For example, use Amplify to add new Cognito resources with default settings to your backend. Override just the settings you want with amplify override auth.

Custom AWS resources

The Amplify CLI provides escape hatches to add custom AWS resources with CDK or CloudFormation. Running the “amplify add custom” command in your Amplify project provides CDK or CloudFormation placeholders along with mechanisms to reference other Amplify-generated resources.

Access AWS resources

Amplify is built on top of Infrastructure-as-Code that deploys resources within your account. Add business logic to your backend using Amplify's Function and Container support. Grant functions access to an SNS topic to send an SMS or allow your container to access an existing database.

Import AWS resources

Integrate your existing resources, such as your Amazon Cognito user pool and federated identities (identity pool), or Storage resources like DynamoDB + S3, into an Amplify project via Amplify Studio. This will enable your API (GraphQL), Storage (S3), and other resources to leverage your existing authentication mechanism.

Command hooks

Use Command Hooks to run custom scripts before, during, and after Amplify CLI commands (“amplify push”, “amplify api gql-compile”, and more). Customers can trigger validation checks, run credential scans, or clean-up build artifacts during deployment. This allows you to extend Amplify’s best-practice defaults to meet your organization’s security guidelines and operational requirements.

Export Infrastructure-as-Code

You can use Amplify with your existing DevOps guidelines and tools to enforce deployment policies or to integrate into your in-house deployment systems. Amplify’s export feature lets you export your Amplify project to your preferred tooling, using CDK. “amplify export” exports the Amplify CLI build artifacts (including CloudFormation templates, API resolver code, and client-side code generation).

Tools

Amplify Libraries

AWS Amplify offers use case-centric open source libraries to build cloud powered mobile and web apps. Amplify libraries are powered by AWS services and can be used with new backends created with the Amplify CLI and Amplify Studio, or your existing AWS backend.

Amplify UI components

Amplify UI Components is an open-source UI toolkit that encapsulates cloud-connected workflows inside of cross-framework UI components. AWS Amplify provides drop-in UI components for authentication, storage and interactions, with a style guide for your apps that automatically integrate with your configured cloud services.

Amplify Studio

Amplify Studio offers an easy way to develop app backends and manage app content. Amplify Studio provides a visual interface for modeling data, adding authentication, authorization and managing users and groups. As you create backend resources, Amplify Studio generates automation templates that enable seamless integration with the Amplify CLI, giving you the ability to extend your app backend with additional capabilities and create multiple environments for testing and team collaboration. With Amplify Studio, you can provide access to team members without an AWS account so developers and non-developers can access the data they need to develop and manage apps more efficiently.

Amplify CLI toolchain

The Amplify Command Line Interface (CLI) is a toolchain to configure and maintain your app backend from your local desktop. Configure cloud functionality using the CLI's interactive workflow and intuitive use cases such as auth, storage, API. Test features locally and deploy multiple environments. All configured resources are available to customers as infrastructure-as-code templates enabling better team collaboration and easy integration with Amplify's CI/CD workflow.

Amplify Hosting

Host your front-end web app, create/delete backend environments, setup CI/CD on the front end and backend, and access Amplify Studio to configure your backend and manage app content and users.

AWS Amplify FAQs

General

Q: What is AWS Amplify?

AWS Amplify consists of a set of tools (open source framework, visual development environment, console) and services (web app and static website hosting) to accelerate the development of mobile and web applications on AWS.

Amplify's open source framework includes an opinionated set of libraries, UI components, and a command line interface (CLI) to build an app backend and integrate it with your iOS, Android, Web, and React Native apps. The framework leverages a core set of AWS Cloud Services to offer capabilities including offline data, authentication, analytics, push notifications, and bots at high scale.

Amplify Studio further simplifies the configuration of backends and frontend UIs with a visual point-and-click experience that works seamlessly with the Amplify CLI. Amplify Studio also includes functionality for managing app content and users.

AWS Amplify also offers a fully managed web app and static website hosting service to host your front-end web app, create/delete backend environments, setup CI/CD on the front end and

backend.

Finally, as part of the broader set of front-end web and mobile development tools and services, you can use AWS Device Farm for testing apps on real iOS devices, Android devices, and web browsers.

Q: What does it cost to use AWS Amplify?

When you use Amplify's open source framework (libraries, UI components, CLI) or Amplify Studio, you pay only for the underlying AWS services you use. There are no additional charges for using these tools.

Q: How does hosting with AWS Amplify relate to Amplify's open source framework?

AWS Amplify consists of tools (open source framework and visual development environment) and a fully managed web hosting service. The tools in the framework (libraries, UI components, CLI), Amplify Studio, the console, and the static web hosting service can be used together or individually.

For example, you can go to AWS Amplify from the AWS console to deploy and host Single Page App (SPA) frontends and static websites, whether or not they use Amplify libraries.

If you are using the Amplify CLI to configure backend resources for your app, AWS Amplify's static web hosting service offers additional functionality. On each check-in, AWS Amplify provisions or updates these backend resources prior to deploying your front end. There is support for a variety of configurations, such as isolated backend deployments per branch or shared backend deployments across branches when you use AWS Amplify's web hosting service.

Q: Where can I find the latest news on AWS Amplify?

Visit our blog and What's New page.

Tools & Features

Q: What can I do with the Amplify libraries, CLI and Amplify Studio?

With the Amplify libraries, you can quickly add features such as offline data, multifactor authentication, analytics, and others to your application with a few lines of code. You can configure the underlying cloud services like AWS AppSync, Amazon Cognito, Amazon Pinpoint, AWS Lambda, Amazon S3, or Amazon Lex directly from the Amplify CLI or Amplify Studio with intuitive guided workflows, minimizing the time required to set-up and manage your backend services.

Q: What languages and platforms do Amplify libraries support?

Amplify libraries support iOS, Android, Web, Flutter, and React Native apps. For Web apps, there is deep integration with React, Ionic, Angular, and Vue.js.

Q: Can I use the Amplify libraries even if I do not use the CLI?

Yes. The libraries can be used to access backend resources that were created without the Amplify CLI.

Q: How do Amplify features work with AWS cloud services?

Amplify features are organized based on the use cases you need to integrate with your app, such as offline data, multi factor authentication, analytics, and others. When you configure these features using the Amplify CLI or the Amplify Studio, the necessary AWS cloud services are provisioned for you. The configuration is persisted in CloudFormation templates that can be checked into source control and shared with other developers. When you add these features to your app via the Amplify libraries, the library makes the necessary calls to AWS services. For example, 'amplify add analytics' will configure Amazon Pinpoint. Then, when you use the Analytics APIs from the Amplify library in your app, the necessary calls will be made to Pinpoint.

Q: How is AWS Amplify related to the AWS Mobile SDKs for iOS and Android?

Amplify iOS and Amplify Android are the recommended ways to build iOS and Android apps that leverage AWS services, whether or not you have configured them using the Amplify CLI. Get started here. If your app is already built using the previous AWS Mobile SDKs for iOS and Android, documentation is available here.

Q. What is Amplify Studio?

Amplify Studio is a visual interface for configuring and maintaining app backends and creating frontend UIs outside the AWS console. Once you've launched your app, Amplify Studio also enables developers and non-developers to manage app content and users.

Q. Why is Amplify Studio outside the AWS console?

Amplify Studio is accessible outside the AWS console to provide front-end developers new to AWS the opportunity to engage with AWS tools quickly and more efficiently. Amplify Studio provides a simplified view of the features needed to build a cloud-connected web or mobile app, both the backend and frontend UI. Amplify Studio also provides easy access for non-developers (QA testers, PMs) to manage the app content and users without requiring developers to figure out the right IAM roles and policies.

Q. What is the Amplify console and how is it different from the Amplify Studio?

The Amplify console is the control center for your app inside the AWS management console. The AWS Amplify console shows you all the front-end environments and backend environments for your apps, whereas Amplify Studio has a unique instance tied to each individual backend environment.

The Amplify console is where you can access AWS Amplify's fully managed web hosting service to set up web hosting, full-stack CI/CD, add a custom domain, clone/delete multiple backend environments, and navigate to underlying AWS service consoles. On the other hand, Amplify Studio is used for configuring and maintaining the app backend - adding features such as auth, data, functions. After launching your app, the Amplify Studio also gives non-developers (QA, PMs) a way to manage app content and users.

Q: What is AWS Amplify's web hosting service?

In addition to AWS Amplify's development tools and features, AWS Amplify offers a fully managed hosting service for web apps and static websites that can be accessed directly from the AWS console. AWS Amplify's static web hosting service provides a complete workflow for building, deploying, and hosting single page web apps or static sites with serverless backends. Continuous deployment allows developers to deploy updates to their web app on every code commit to their Git repository. When the build succeeds, the app is deployed and hosted on an amplifyapp.com subdomain. Developers can connect their custom domain to start receiving production traffic.

Q: What type of web apps can I build and deploy?

In addition to AWS Amplify's development tools and features, AWS Amplify offers a fully managed static web hosting service for web apps and static websites that can be accessed directly from the AWS console. AWS Amplify's static web hosting service provides a complete workflow for building, deploying, and hosting single page web apps or static sites with serverless backends. Continuous deployment allows developers to deploy updates to their web app on every code commit to their Git repository. When the build succeeds, the app is deployed and hosted on an amplifyapp.com subdomain. Developers can connect their custom domain to start receiving production traffic.

Q: How do I get started with AWS Amplify web hosting?

To get started, go to AWS Amplify in the AWS console and connect your source repository. AWS Amplify automatically determines the front-end framework used, and then builds and deploys the app to a globally available content delivery network (CDN). Amplify detects backend functionality added using the Amplify CLI or Amplify Studio, and can deploy the necessary AWS resources in the same deployment as the front end. AWS Amplify will build and deploy your web app quickly, and host your web app on a globally available content delivery network (CDN) with a friendly URL (example: https://master.appname.amplifyapp.com). To get started, go to AWS Amplify on the AWS console.

Q: What is an AWS Amplify 'app'?

An AWS Amplify 'app' is your project container. Each app project contains a list of branches you have connected from your source repository. You can connect additional feature branches, a custom domain, or access your build logs from your app project.

Q: What is continuous deployment?

Continuous deployment is a DevOps strategy for software releases where every code commit to a repository is automatically released to production or staging environment. This practice reduces time to market by ensuring that your hosted web app is always a reflection of the latest code in your repository.

Q: What Git source code providers does AWS Amplify static web hosting support?

You can connect private and public repositories from GitHub, BitBucket, GitLab, and AWS CodeCommit.

Q: Does AWS Amplify web hosting store my Git access tokens?

AWS Amplify never stores access tokens from repositories. Once you authorize AWS Amplify, we fetch an access token from your source provider. We simply pass the token to our console, and from then on, all communication with the GitHub API happens straight from the browser. After configuring continuous deployment, the token is permanently discarded.

Q: Does AWS Amplify web hosting support private Git servers?

We currently do not support private Git servers.

Q: What are environment variables? How do I use them?

Environment variables are configurations required by apps at runtime. These configurations could include database connection details, third-party API keys, different customization parameters and secrets. The best way to expose these configurations is to do so with environment variables. You can add environment variables when creating an app or by going to

the app settings. All environment variables are encrypted to prevent rogue access. Add all your app environment variables in the key and value textboxes. By default, AWS Amplify applies the environment variables across all branches, so you don't have to re-enter variables when you connect a new branch. Once you enter all the variables hit Save.

Q: What happens when a build is run?

AWS Amplify will create a temporary compute container (4 vCPU, 7GB RAM), download the source code, run the commands configured in the project, deploy the generated artifact to a web hosting environment, and then destroy the compute container. During the build, AWS Amplify will stream the build output to the service console.

Q: How can I leverage AWS Amplify web hosting to work with multiple environments?

AWS Amplify leverages Git's branching model to create new environments every time a developer pushes code to a new branch. In typical development teams, developers deploy their 'master' branch to production, keep the 'dev' branch as staging, and create feature branches when working on new functionality. AWS Amplify Console can create frontend and backend environments linked to each connected branch. This allows developers to work in sandbox environments, and use 'Git' as a mechanism to merge code and resolve conflicts. Changes are automatically pushed to production once they are merged into the master (or production) branch.

Q: What are atomic deployments?

Every deployment is atomic, which means the site is ready to view after the deployment is complete. Atomic deployments eliminate maintenance windows by ensuring the web app is only updated once the entire deploy has finished. The new version of the web app is then made available instantly to end-users, without the developer having to invalidate CDN caches.

Q: How is hosting a modern web app different from a traditional web app?

Hosting a modern web app does not require web servers and can use content delivery networks to store static content (HTML, CSS and JavaScript files). AWS Amplify leverages the Amazon CloudFront Global Edge Network to distribute your web app globally.

Q: How do I connect my custom domain?

Connecting your custom domain is easy – if your domain is registered on Route53, simply pick it from a dropdown and AWS Amplify will automatically configure the DNS records to point the apex and 'www' subdomain to your website. Additionally, we automatically create subdomains for all branches that are connected. For example, connecting a 'dev' branch creates a deployment at https://dev.appname.amplifyapp.com. As part of the custom domain setup, we generate a free HTTPS certificate to ensure traffic to your site is secure.

Q: What domain registrars does AWS Amplify web hosting support?

Domains purchased through all domain registrars can be connected to an app by defining a custom domain. For developers using Amazon Route53 as their registrar, AWS Amplify automatically updates the DNS records to point to their deployed app. For 3rd party registrars, AWS Amplify provides instructions on how to update their DNS records.

Q: Is all web traffic served over HTTPS?

AWS Amplify web hosting generates a free HTTPS on all sites and will enable it automatically on all Route53-managed domains. The SSL certificate is generated by Amazon Certificate Manager and has wildcard domain support. ACM handles the complexity of creating and managing

public SSL/TLS certificates for your AWS based websites and applications. With the wildcard option, the main domain and all subdomains can be covered by a single certificate.

Q: Can I password protect my web deployments?

All web deployments can be password protected with basic access authentication. When working on new features, developers can share updates with internal stakeholders by setting up a username and password for a branch deployment.

Q: What are redirects and rewrites? How do I use them?

A redirect is a client-side request to have the web browser go to another URL. This means that the URL that you see in the browser will update to the new URL. A rewrite is a server-side rewrite of the URL. This will not change what you see in the browser because the changes are hidden from the user. Reverse proxies are cross-origin rewrites. From the AWS Amplify console settings, developers can specify redirects, HTTP response code, custom 404s, and proxies to external services.

Q: How will I be charged for my use of AWS Amplify Hosting?

AWS Amplify web hosting is priced for two features – build & deploy, and web hosting. For the build & deploy feature the price per build minute is $0.01. For the hosting feature the price per GB served is $0.15 and price per GB stored is $0.023.With the AWS Free Usage Tier, you can get started for free. Upon sign up, new AWS customers receive 1,000 build minutes per month for the build and deploy feature, and 15 GB served per month and 5 GB data storage per month for the hosting feature.

Q: Does your prices include taxes?

Except as otherwise noted, our prices are exclusive of applicable taxes and duties, including VAT and applicable sales tax. For customers with a Japanese billing address, use of AWS services is subject to Japanese Consumption Tax. Learn more.

Q: Are prices different per region?

Prices are the same across all regions.

AWS Budgets

Improve planning and cost control with flexible budgeting and forecasting.

How it works

With AWS Budgets, set custom budgets to track your costs and usage, and respond quickly to alerts received from email or SNS notifications if you exceed your threshold.

Use cases

Monitor costs and usage

Set your preferred budget period to daily, monthly, quarterly, or annually, and create specific budget limits.

Create scheduled reports

Stay informed on how actual or forecasted costs and usage progress toward your budget threshold.

Respond to thresholds

Set up custom actions to run automatically or through an approval process when a budget target is exceeded.

Cloud Financial Management with AWS

The cloud allows you to trade fixed expenses (such as data centers and physical servers) for variable expenses, and only pay for IT as you consume it. And, because of the economies of scale, the variable expenses are much lower than what you would pay to do it yourself. Whether you were born in the cloud, or you are just starting your migration journey to the cloud, AWS has a set of solutions to help you manage and optimize your spend.

During this unprecedented time, many businesses and organizations are facing disruption to their operations, budgets, and revenue. AWS has a set of solutions to help you with cost management and optimization. This includes services, tools, and resources to organize and track cost and usage data, enhance control through consolidated billing and access permission, enable better planning through budgeting and forecasts, and further lower cost with resources and pricing optimizations.

Cloud Financial Management Solutions

Organize and Report Cost and Usage based on User-Defined Methods

You need complete, near real-time visibility of your cost and usage information to make informed decisions. AWS equips you with tools to organize your resources based on your needs, visualize and analyze cost and usage data in a single pane of glass, and accurately chargeback to appropriate entities (e.g. department, project, product). Rather than centrally policing the cost, you can provide real-time cost data that makes sense to your engineering, application, and business teams. The detailed, allocable cost data allows teams to have the visibility and details to be accountable of their own spend.

Manage Billing and Control Costs

Business and organization leaders need a simple and easy way to access AWS billing information, including a spend summary, a breakdown of all service costs incurred by accounts across the organization, along with discounts and credits. Customer can choose to consolidate your bills and take advantage of higher volume discounts based on aggregated usage across your bills. Leaders also need to set appropriate guardrails in place so you can maintain control over cost, governance, and security. AWS helps organizations balance freedom and control by enabling the governance of granular user permission.

Improved Planning with Flexible Forecasting and Budgeting

Businesses and organizations need to plan and set expectations around cloud costs for your projects, applications, and more. The emergence of the cloud allowed teams to acquire and deprecate resources on an ongoing basis, without relying on teams to approve, procure and install infrastructure. However, this flexibility requires organizations to adapt to the new, dynamic forecasting and budgeting process. AWS provides forecasts based on your cost and usage history and allows you to set budget threshold and alerts, so you can stay informed whenever cost and usage is forecasted to, or exceeds the threshold limit. You can also set reservation utilization and/or coverage targets for your Reserved Instances and Savings Plans and monitor how they are progressing towards your target.

Optimize Costs with Resource and Pricing Recommendations

With AWS, customers can take control of your cost and continuously optimize your spend. There are a variety of AWS pricing models and resources you can choose from to meet requirements for both performance and cost efficiency, and adjust as needed. When evaluating AWS services for your architectural and business needs, you will have the flexibility to choose from a variety of elements, such as operating systems, instance types, availability zones, and purchase options. AWS offers resources optimization recommendations to simplify the evaluation process so you can efficiently select the cost-optimized resources. We also provide recommendations around pricing models (up to 72% with Reserved Instances and Savings Plans and up to 90% with Spot Instances) based on your utilization patterns, so you can further drive down your cost without compromising workload performance.

AWS Cloud Financial Management Services

Whether you want to organize and track your cost and usage, enhance control through consolidated billing and access permission, enable better planning through budgeting and forecasting, or further lower cost with resources and pricing optimizations, you can leverage our services, tools, and resources to help reduce your AWS bill.

Use Cases	Capabilities	AWS Resources
Organize	Construct your cost allocation strategy that aligns with your business logic	AWS Billing Conductor | AWS Cost Allocation Tags | AWS Cost Categories
Report	Raise awareness and accountability of your cloud spend with the detailed, allocable cost data	AWS Cost Explorer | AWS Cost and Usage Report | AWS Application Cost Profiler
Access	Track billing information across the organization in a consolidated view	AWS Consolidated Billing | AWS Purchase Order Management | AWS Credits
Control	Establish effective governance mechanisms with the right guardrails in place	AWS Cost Anomaly Detection | AWS Identity and Access Management | AWS Organizations | AWS Control Tower | AWS Service Catalog
Forecast	Estimate your resource utilization and spend with forecast dashboards that you create	AWS Cost Explorer (Self-Service) | AWS Budgets (Event-Driven)
Budget	Keep your spend in check with custom	AWS Budgets | AWS Budget Actions | AWS Service Catalog

	budget threshold and auto alert notification	
Purchase	Leverage free trials and programmatic discounts based on your workload pattern and needs	AWS Free Tier | AWS Reserved Instances | AWS Savings Plans | AWS Spot Instances | Amazon DynamoDB On-demand
Elasticity	Scale and schedule your services based on your expected utilization pattern and needs	AWS Instance Scheduler | Amazon Redshift pause and resume | EC2 Auto Scaling | AWS Trusted Advisor
Rightsize	Align your service allocation size to your actual workload demand	AWS Cost Explorer Right Sizing Recommendations | AWS Compute Optimizer | Amazon Redshift resize | Amazon S3 Intelligent Tiering
Inspect	Stay up-to-date with your resource deployment and cost optimization opportunities	AWS Cost Explorer

General AWS Budgets FAQs

Q: What is AWS Budgets and how does it work?

Using AWS Budgets, you can set a budget that alerts you when you exceed (or are forecasted to exceed) your budgeted cost or usage amount. You can also set alerts based on your RI or Savings Plans Utilization and Coverage using AWS Budgets.

Q: What kinds of dimensions can be used to create a budget?

AWS Budgets gives you access to a number of filtering dimensions (i.e., AWS Service, Availability Zone, and Member Account), and allows you to create budgets that are tracked on a monthly, quarterly, or yearly cadence.

Q: How many budgets can I create?

You can create up to 20,000 budgets. If you would like to increase your limit, please reach out to Customer Support.

Q: How many alerts and subscribers can I add for each budget?

For each budget, you are allowed to create up to five alerts. Each alert can be sent to 10 email subscribers and/or be published to an SNS topic.