Contents

Introduction 1

What is Cloud Computing? 1

Six Advantages of Cloud Computing 2

Types of Cloud Computing 3

- Cloud Computing Models 3
- Cloud Computing Deployment Models 4

Global Infrastructure 5

Security and Compliance 6

- Security 6
- Compliance 7

Amazon Web Services Cloud Platform 8

- AWS Management Console 8
- AWS Command Line Interface 8
- Software Development Kits 8
- Analytics 8
- Application Integration 14
- AR and VR 16
- AWS Cost Management 16
- Blockchain 18
- Business Applications 18
- Compute 20
- Customer Engagement 26
- Database 26
- Desktop and App Streaming 31
- Developer Tools 32
- Game Tech 35

Internet of Things (IoT) 35
Machine Learning 42
Management and Governance 52
Media Services 60
Migration and Transfer 62
Mobile 67
Networking and Content Delivery 69
Robotics 75
Satellite 76
Security, Identity, and Compliance 77
Storage 84
Next Steps 87
Conclusion 88
Contributors 88
Further Reading 88
Document Revisions 89

Introduction

In 2006, Amazon Web Services (AWS) began offering IT infrastructure services to businesses as web services—now commonly known as cloud computing. One of the key benefits of cloud computing is the opportunity to replace upfront capital infrastructure expenses with low variable costs that scale with your business. With the cloud, businesses no longer need to plan for and procure servers and other IT infrastructure weeks or months in advance. Instead, they can instantly spin up hundreds or thousands of servers in minutes and deliver results faster.

Today, AWS provides a highly reliable, scalable, low-cost infrastructure platform in the cloud that powers hundreds of thousands of businesses in 190 countries around the world.

What is Cloud Computing?

Cloud computing is the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the Internet with pay-as-you-go pricing. Whether you are running applications that share photos to millions of mobile users or you're supporting the critical operations of your business, a cloud services platform provides rapid access to flexible and low-cost IT resources. With cloud computing, you don't need to make large upfront investments in hardware and spend a lot of time on the heavy lifting of managing that hardware. Instead, you can provision exactly the right type and size of computing resources you need to power your newest bright idea or operate your IT department. You can access as many resources as you need, almost instantly, and only pay for what you use.

Cloud computing provides a simple way to access servers, storage, databases and a broad set of application services over the Internet. A cloud services platform, such as Amazon Web Services, owns and maintains the network-connected hardware required for these application services, while you provision and use what you need via a web application.

Six Advantages of Cloud Computing

- **Trade capital expense for variable expense** – Instead of having to invest heavily in data centers and servers before you know how you're going to use them, you can pay only when you consume computing resources, and pay only for how much you consume.
- **Benefit from massive economies of scale** – By using cloud computing, you can achieve a lower variable cost than you can get on your own. Because usage from hundreds of thousands of customers is aggregated in the cloud, providers such as AWS can achieve higher economies of scale, which translates into lower pay as-you-go prices.
- **Stop guessing capacity** – Eliminate guessing on your infrastructure capacity needs. When you make a capacity decision prior to deploying an application, you often end up either sitting on expensive idle resources or dealing with limited capacity. With cloud computing, these problems go away. You can access as much or as little capacity as you need, and scale up and down as required with only a few minutes' notice.
- **Increase speed and agility** – In a cloud computing environment, new IT resources are only a click away, which means that you reduce the time to make those resources available to your developers from weeks to just minutes. This results in a dramatic increase in agility for the organization, since the cost and time it takes to experiment and develop is significantly lower.
- **Stop spending money running and maintaining data centers** – Focus on projects that differentiate your business, not the infrastructure. Cloud computing lets you focus on your own customers, rather than on the heavy lifting of racking, stacking, and powering servers.
- **Go global in minutes** – Easily deploy your application in multiple regions around the world with just a few clicks. This means you can provide lower latency and a better experience for your customers at minimal cost.

Types of Cloud Computing

Cloud computing provides developers and IT departments with the ability to focus on what matters most and avoid undifferentiated work such as procurement, maintenance, and capacity planning. As cloud computing has grown in popularity, several different models and deployment strategies have emerged to help meet specific needs of different users. Each type of cloud service and deployment method provides you with different levels of control, flexibility, and management. Understanding the differences between Infrastructure as a Service, Platform as a Service, and Software as a Service, as well as what deployment strategies you can use, can help you decide what set of services is right for your needs.

Cloud Computing Models

Infrastructure as a Service (IaaS)

Infrastructure as a Service (IaaS) contains the basic building blocks for cloud IT and typically provides access to networking features, computers (virtual or on dedicated hardware), and data storage space. IaaS provides you with the highest level of flexibility and management control over your IT resources and is most similar to existing IT resources that many IT departments and developers are familiar with today.

Platform as a Service (PaaS)

Platform as a Service (PaaS) removes the need for your organization to manage the underlying infrastructure (usually hardware and operating systems) and allows you to focus on the deployment and management of your applications. This helps you be more efficient as you don’t need to worry about resource procurement, capacity planning, software maintenance, patching, or any of the other undifferentiated heavy lifting involved in running your application.

Software as a Service (SaaS)

Software as a Service (SaaS) provides you with a completed product that is run and managed by the service provider. In most cases, people referring to Software as a Service are referring to end-user applications. With a SaaS offering you do not have to think about how the service is maintained or how the

underlying infrastructure is managed; you only need to think about how you will use that particular piece of software. A common example of a SaaS application is web-based email which you can use to send and receive email without having to manage feature additions to the email product or maintain the servers and operating systems that the email program is running on.

Cloud Computing Deployment Models

Cloud

A cloud-based application is fully deployed in the cloud and all parts of the application run in the cloud. Applications in the cloud have either been created in the cloud or have been migrated from an existing infrastructure to take advantage of the benefits of cloud computing. Cloud-based applications can be built on low-level infrastructure pieces or can use higher level services that provide abstraction from the management, architecting, and scaling requirements of core infrastructure.

Hybrid

A hybrid deployment is a way to connect infrastructure and applications between cloud-based resources and existing resources that are not located in the cloud. The most common method of hybrid deployment is between the cloud and existing on-premises infrastructure to extend, and grow, an organization's infrastructure into the cloud while connecting cloud resources to the internal system. For more information on how AWS can help you with your hybrid deployment, please visit our hybrid page.

On-premises

The deployment of resources on-premises, using virtualization and resource management tools, is sometimes called the “private cloud.” On-premises deployment doesn’t provide many of the benefits of cloud computing but is sometimes sought for its ability to provide dedicated resources. In most cases this deployment model is the same as legacy IT infrastructure while using application management and virtualization technologies to try and increase resource utilization.

Global Infrastructure

AWS serves over a million active customers in more than 190 countries. We are steadily expanding global infrastructure to help our customers achieve lower latency and higher throughput, and to ensure that their data resides only in the AWS Region they specify. As our customers grow their businesses, AWS will continue to provide infrastructure that meets their global requirements.

The AWS Cloud infrastructure is built around AWS Regions and Availability Zones. An AWS Region is a physical location in the world where we have multiple Availability Zones. Availability Zones consist of one or more discrete data centers, each with redundant power, networking, and connectivity, housed in separate facilities. These Availability Zones offer you the ability to operate production applications and databases that are more highly available, fault tolerant, and scalable than would be possible from a single data center. The AWS Cloud operates in over 60 Availability Zones within over 20 geographic Regions around the world, with announced plans for more Availability Zones and Regions. For more information on the AWS Cloud Availability Zones and AWS Regions, see AWS Global Infrastructure.

Each Amazon Region is designed to be completely isolated from the other Amazon Regions. This achieves the greatest possible fault tolerance and stability. Each Availability Zone is isolated, but the Availability Zones in a Region are connected through low-latency links. AWS provides you with the flexibility to place instances and store data within multiple geographic regions as well as across multiple Availability Zones within each AWS Region. Each Availability Zone is designed as an independent failure zone. This means that Availability Zones are physically separated within a typical metropolitan region and are located in lower risk flood plains (specific flood zone categorization varies by AWS Region). In addition to discrete uninterruptable power supply (UPS) and onsite backup generation facilities, data centers located in different Availability Zones are designed to be supplied by independent substations to reduce the risk of an event on the power grid impacting more than one Availability Zone. Availability Zones are all redundantly connected to multiple tier-1 transit providers.

Security and Compliance

Security

Cloud security at AWS is the highest priority. As an AWS customer, you will benefit from a data center and network architecture built to meet the requirements of the most security-sensitive organizations. Security in the cloud is much like security in your on-premises data centers—only without the costs of maintaining facilities and hardware. In the cloud, you don't have to manage physical servers or storage devices. Instead, you use software-based security tools to monitor and protect the flow of information into and of out of your cloud resources.

An advantage of the AWS Cloud is that it allows you to scale and innovate, while maintaining a secure environment and paying only for the services you use. This means that you can have the security you need at a lower cost than in an on-premises environment.

As an AWS customer you inherit all the best practices of AWS policies, architecture, and operational processes built to satisfy the requirements of our most security-sensitive customers. Get the flexibility and agility you need in security controls.

The AWS Cloud enables a shared responsibility model. While AWS manages security of the cloud, you are responsible for security in the cloud. This means that you retain control of the security you choose to implement to protect your own content, platform, applications, systems, and networks no differently than you would in an on-site data center.

AWS provides you with guidance and expertise through online resources, personnel, and partners. AWS provides you with advisories for current issues, plus you have the opportunity to work with AWS when you encounter security issues.

You get access to hundreds of tools and features to help you to meet your security objectives. AWS provides security-specific tools and features across network security, configuration management, access control, and data encryption.

Finally, AWS environments are continuously audited, with certifications from accreditation bodies across geographies and verticals. In the AWS

environment, you can take advantage of automated tools for asset inventory and privileged access reporting.

Benefits of AWS Security

- **Keep Your Data Safe:** The AWS infrastructure puts strong safeguards in place to help protect your privacy. All data is stored in highly secure AWS data centers.
- **Meet Compliance Requirements:** AWS manages dozens of compliance programs in its infrastructure. This means that segments of your compliance have already been completed.
- **Save Money:** Cut costs by using AWS data centers. Maintain the highest standard of security without having to manage your own facility
- **Scale Quickly:** Security scales with your AWS Cloud usage. No matter the size of your business, the AWS infrastructure is designed to keep your data safe.

Compliance

AWS Cloud Compliance enables you to understand the robust controls in place at AWS to maintain security and data protection in the cloud. As systems are built on top of AWS Cloud infrastructure, compliance responsibilities will be shared. By tying together governance-focused, audit-friendly service features with applicable compliance or audit standards, AWS Compliance enablers build on traditional programs. This helps customers to establish and operate in an AWS security control environment.

The IT infrastructure that AWS provides to its customers is designed and managed in alignment with best security practices and a variety of IT security standards. The following is a partial list of assurance programs with which AWS complies:

- SOC 1/ISAE 3402, SOC 2, SOC 3
- FISMA, DIACAP, and FedRAMP
- PCI DSS Level 1
- ISO 9001, ISO 27001, ISO 27017, ISO 27018

AWS provides customers a wide range of information on its IT control environment in whitepapers, reports, certifications, accreditations, and other third-party attestations. More information is available in the Risk and Compliance whitepaper and the AWS Security Center.

Amazon Web Services Cloud Platform

AWS consists of many cloud services that you can use in combinations tailored to your business or organizational needs. This section introduces the major AWS services by category. To access the services, you can use the AWS Management Console, the Command Line Interface, or Software Development Kits (SDKs).

AWS Management Console

Access and manage Amazon Web Services through the AWS Management Console, a simple and intuitive user interface. You can also use the AWS Console Mobile Application to quickly view resources on the go.

AWS Command Line Interface

The AWS Command Line Interface (CLI) is a unified tool to manage your AWS services. With just one tool to download and configure, you can control multiple AWS services from the command line and automate them through scripts.

Software Development Kits

Our Software Development Kits (SDKs) simplify using AWS services in your applications with an Application Program Interface (API) tailored to your programming language or platform.

Analytics

Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

Athena is easy to use. Simply point to your data in Amazon S3, define the schema, and start querying using standard SQL. Most results are delivered within seconds. With Athena, there's no need for complex extract, transform, and load (ETL) jobs to prepare your data for analysis. This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets.

Athena is out-of-the-box integrated with AWS Glue Data Catalog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions, and maintain schema versioning. You can also use Glue's fully-managed ETL capabilities to transform data or convert it into columnar formats to optimize cost and improve performance.

Amazon EMR

Amazon EMR provides a managed Hadoop framework that makes it easy, fast, and cost-effective to process vast amounts of data across dynamically scalable Amazon EC2 instances. You can also run other popular distributed frameworks such as Apache Spark, HBase, Presto, and Flink in Amazon EMR, and interact with data in other AWS data stores such as Amazon S3 and Amazon DynamoDB. EMR Notebooks, based on the popular Jupyter Notebook, provide a development and collaboration environment for ad hoc querying and exploratory analysis.

Amazon EMR securely and reliably handles a broad set of big data use cases, including log analysis, web indexing, data transformations (ETL), machine learning, financial analysis, scientific simulation, and bioinformatics.

Amazon CloudSearch

Amazon CloudSearch is a managed service in the AWS Cloud that makes it simple and cost-effective to set up, manage, and scale a search solution for your website or application. Amazon CloudSearch supports 34 languages and popular search features such as highlighting, autocomplete, and geospatial search.

Amazon Elasticsearch Service

Amazon Elasticsearch Service makes it easy to deploy, secure, operate, and scale Elasticsearch to search, analyze, and visualize data in real-time. With Amazon Elasticsearch Service, you get easy-to-use APIs and real-time

analytics capabilities to power use-cases such as log analytics, full-text search, application monitoring, and clickstream analytics, with enterprise-grade availability, scalability, and security. The service offers integrations with open-source tools like Kibana and Logstash for data ingestion and visualization. It also integrates seamlessly with other AWS services such as Amazon Virtual Private Cloud (Amazon VPC), AWS Key Management System (AWS KMS), Amazon Kinesis Data Firehose, AWS Lambda, AWS Identity and Access Management (IAM), Amazon Cognito, and Amazon CloudWatch, so that you can go from raw data to actionable insights quickly.

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.

Amazon Kinesis currently offers four services: Kinesis Data Firehose, Kinesis Data Analytics, Kinesis Data Streams, and Kinesis Video Streams.

Amazon Kinesis Data Firehose

Amazon Kinesis Data Firehose is the easiest way to reliably load streaming data into data stores and analytics tools. It can capture, transform, and load streaming data into Amazon S3, Amazon Redshift, Amazon Elasticsearch 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, transform, and encrypt the data before loading it, minimizing the amount of storage used at the destination and increasing security.

You can easily create a Firehose delivery stream from the AWS Management Console, configure it with a few clicks, and start sending data to the stream from hundreds of thousands of data sources to be loaded continuously to

AWS—all in just a few minutes. You can also configure your delivery stream to automatically convert the incoming data to columnar formats like Apache Parquet and Apache ORC, before the data is delivered to Amazon S3, for cost-effective storage and analytics.

Amazon Kinesis Data Analytics

Amazon Kinesis Data Analytics is the easiest way to analyze streaming data, gain actionable insights, and respond to your business and customer needs in real time. Amazon Kinesis Data Analytics reduces the complexity of building, managing, and integrating streaming applications with other AWS services. SQL users can easily query streaming data or build entire streaming applications using templates and an interactive SQL editor. Java developers can quickly build sophisticated streaming applications using open source Java libraries and AWS integrations to transform and analyze data in real-time.

Amazon Kinesis Data Analytics takes care of everything required to run your queries continuously and scales automatically to match the volume and throughput rate of your incoming data.

Amazon Kinesis Data Streams

Amazon Kinesis Data Streams (KDS) is a massively scalable and durable real-time data streaming service. KDS can continuously capture gigabytes of data per second from hundreds of thousands of sources such as website clickstreams, database event streams, financial transactions, social media feeds, IT logs, and location-tracking events. The data collected is available in milliseconds to enable real-time analytics use cases such as real-time dashboards, real-time anomaly detection, dynamic pricing, and more.

Amazon Kinesis Video Streams

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 also 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 Recognition Video, and libraries for ML frameworks such as Apache MxNet, TensorFlow, and OpenCV.

Amazon Redshift

Amazon Redshift is a fast, scalable data warehouse that makes it simple and cost-effective to analyze all your data across your data warehouse and data lake. Redshift delivers ten times faster performance than other data warehouses by using machine learning, massively parallel query execution, and columnar storage on high-performance disk. You can setup and deploy a new data warehouse in minutes, and run queries across petabytes of data in your Redshift data warehouse, and exabytes of data in your data lake built on Amazon S3. You can start small for just $0.25 per hour and scale to $250 per terabyte per year, less than one-tenth the cost of other solutions.

Amazon QuickSight

Amazon QuickSight is a fast, cloud-powered business intelligence (BI) service that makes it easy for you to deliver insights to everyone in your organization. QuickSight lets you create and publish interactive dashboards that can be accessed from browsers or mobile devices. You can embed dashboards into your applications, providing your customers with powerful self-service analytics. QuickSight easily scales to tens of thousands of users without any software to install, servers to deploy, or infrastructure to manage.

AWS Data Pipeline

AWS Data Pipeline is a web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premises data sources, at specified intervals. With AWS Data Pipeline, you can regularly access your data where it's stored, transform and process it at scale, and efficiently transfer the results to AWS services such as Amazon S3, Amazon RDS, Amazon DynamoDB, and Amazon EMR.

AWS Data Pipeline helps you easily create complex data processing workloads that are fault tolerant, repeatable, and highly available. You don't have to worry about ensuring resource availability, managing inter-task dependencies, retrying transient failures or timeouts in individual tasks, or creating a failure notification system. AWS Data Pipeline also allows you to move and process data that was previously locked up in on-premises data silos.

AWS Glue

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics. You can create and run an ETL job with a few clicks in the AWS Management Console. You simply point AWS Glue to your data stored on AWS, and AWS Glue discovers your data and stores the associated metadata (e.g. table definition and schema) in the AWS Glue Data Catalog. Once cataloged, your data is immediately searchable, queryable, and available for ETL.

AWS Lake Formation

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.

However, setting up and managing data lakes today involves a lot of manual, complicated, and time-consuming tasks. This work includes loading data from diverse sources, monitoring those data flows, setting up partitions, turning on encryption and managing keys, defining transformation jobs and monitoring their operation, re-organizing data into a columnar format, configuring access control settings, deduplicating redundant data, matching linked records, granting access to data sets, and auditing access over time.

Creating a data lake with Lake Formation is as simple as defining where your data resides and what data access and security policies you want to apply. Lake Formation then collects and catalogs data from databases and object storage, moves the data into your new Amazon S3 data lake, cleans and classifies data using machine learning algorithms, and secures access to your sensitive data. Your users can then access a centralized catalog of data which 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 EMR for Apache Spark, Amazon Redshift, Amazon Athena, Amazon SageMaker, and Amazon QuickSight.

Amazon Managed Streaming for Kafka (MSK)

Amazon Managed Streaming for Kafka (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 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 Managed Streaming for Kafka 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.

Application Integration

AWS Step Functions

AWS Step Functions lets you coordinate multiple AWS services into serverless workflows so you can build and update apps quickly. Using Step Functions, you can design and run workflows that stitch together services such as AWS Lambda and Amazon ECS into feature-rich applications. Workflows are made up of a series of steps, with the output of one step acting as input into the next. Application development is simpler and more intuitive using Step Functions, because it translates your workflow into a state machine diagram that is easy to understand, easy to explain to others, and easy to change. You can monitor each step of execution as it happens, which means you can identify and fix problems quickly. Step Functions automatically triggers and tracks each step, and retries when there are errors, so your application executes in order and as expected.

Amazon MQ

Amazon MQ is a managed message broker service for Apache ActiveMQ that makes it easy to set up and operate message brokers in the cloud. Message brokers allow different software systems–often using different programming languages, and on different platforms–to communicate and exchange information. Amazon MQ reduces your operational load by managing the provisioning, setup, and maintenance of ActiveMQ, a popular open-source message broker. Connecting your current applications to Amazon MQ is easy because it uses industry-standard APIs and protocols for messaging, including JMS, NMS, AMQP, STOMP, MQTT, and WebSocket. Using standards means that in most cases, there's no need to rewrite any messaging code when you migrate to AWS.

Amazon SQS

Amazon Simple Queue Service (Amazon SQS) is a fully managed message queuing service that enables you to decouple and scale microservices, distributed systems, and serverless applications. SQS eliminates the complexity and overhead associated with managing and operating message oriented middleware, and empowers developers to focus on differentiating work. Using SQS, you can send, store, and receive messages between software components at any volume, without losing messages or requiring other services to be available. Get started with SQS in minutes using the AWS console, Command Line Interface or SDK of your choice, and three simple commands.

SQS offers two types of message queues. Standard queues offer maximum throughput, best-effort ordering, and at-least-once delivery. SQS FIFO queues are designed to guarantee that messages are processed exactly once, in the exact order that they are sent.

Amazon SNS

Amazon Simple Notification Service (Amazon SNS) is a highly available, durable, secure, fully managed pub/sub messaging service that enables you to decouple microservices, distributed systems, and serverless applications. Amazon SNS provides topics for high-throughput, push-based, many-to-many messaging. Using Amazon SNS topics, your publisher systems can fan out messages to a large number of subscriber endpoints for parallel processing, including Amazon SQS queues, AWS Lambda functions, and HTTP/S

webhooks. Additionally, SNS can be used to fan out notifications to end users using mobile push, SMS, and email.

Amazon SWF

Amazon Simple Workflow (Amazon SWF) helps developers build, run, and scale background jobs that have parallel or sequential steps. You can think of Amazon SWF as a fully-managed state tracker and task coordinator in the cloud. If your application's steps take more than 500 milliseconds to complete, you need to track the state of processing. If you need to recover or retry if a task fails, Amazon SWF can help you.

AR and VR

Amazon Sumerian

Amazon Sumerian lets you create and run virtual reality (VR), augmented reality (AR), and 3D applications quickly and easily without requiring any specialized programming or 3D graphics expertise. With Sumerian, you can build highly immersive and interactive scenes that run on popular hardware such as Oculus Go, Oculus Rift, HTC Vive, HTC Vive Pro, Google Daydream, and Lenovo Mirage as well as Android and iOS mobile devices. For example, you can build a virtual classroom that lets you train new employees around the world, or you can build a virtual environment that enables people to tour a building remotely. Sumerian makes it easy to create all the building blocks needed to build highly immersive and interactive 3D experiences including adding objects (e.g. characters, furniture, and landscape), and designing, animating, and scripting environments. Sumerian does not require specialized expertise and you can design scenes directly from your browser.

AWS Cost Management

AWS Cost Explorer

AWS Cost Explorer has an easy-to-use interface that lets you visualize, understand, and manage your AWS costs and usage over time. Get started quickly by creating custom reports (including charts and tabular data) that analyze cost and usage data, both at a high level (e.g., total costs and usage across all accounts) and for highly-specific requests (e.g., m2.2xlarge costs within account Y that are tagged "project: secretProject").

AWS Budgets

AWS Budgets gives you the ability to set custom budgets that alert you when your costs or usage exceed (or are forecasted to exceed) your budgeted amount. You can also use AWS Budgets to set RI utilization or coverage targets and receive alerts when your utilization drops below the threshold you define. RI alerts support Amazon EC2, Amazon RDS, Amazon Redshift, and Amazon ElastiCache reservations.

Budgets can be tracked at the monthly, quarterly, or yearly level, and you can customize the start and end dates. You can further refine your budget to track costs associated with multiple dimensions, such as AWS service, linked account, tag, and others. Budget alerts can be sent via email and/or Amazon Simple Notification Service (SNS) topic.

Budgets can be created and tracked from the AWS Budgets dashboard or via the Budgets API.

AWS Cost & Usage Report

The AWS Cost & Usage Report is a single location for accessing comprehensive information about your AWS costs and usage.

The AWS Cost & Usage Report lists AWS usage for each service category used by an account and its IAM users in hourly or daily line items, as well as any tags that you have activated for cost allocation purposes. You can also customize the AWS Cost & Usage Report to aggregate your usage data to the daily or monthly level.

Reserved Instance (RI) Reporting

AWS provides a number of RI-specific cost management solutions out-of-the-box to help you better understand and manage your RIs. Using the RI Utilization and Coverage reports available in AWS Cost Explorer, you can visualize your RI data at an aggregate level or inspect a particular RI subscription. To access the most detailed RI information available, you can leverage the AWS Cost & Usage Report. You can also set a custom RI utilization target via AWS Budgets and receive alerts when your utilization drops below the threshold you define.

Blockchain

Amazon Managed Blockchain

Amazon Managed Blockchain is a fully managed service that makes it easy to create and manage scalable blockchain networks using the popular open source frameworks Hyperledger Fabric and Ethereum.

Blockchain makes it possible to build applications where multiple parties can execute transactions without the need for a trusted, central authority. Today, building a scalable blockchain network with existing technologies is complex to set up and hard to manage. To create a blockchain network, each network member needs to manually provision hardware, install software, create and manage certificates for access control, and configure networking components. Once the blockchain network is running, you need to continuously monitor the infrastructure and adapt to changes, such as an increase in transaction requests, or new members joining or leaving the network.

Amazon Managed Blockchain is a fully managed service that allows you to set up and manage a scalable blockchain network with just a few clicks. Amazon Managed Blockchain eliminates the overhead required to create the network, and automatically scales to meet the demands of thousands of applications running millions of transactions. Once your network is up and running, Managed Blockchain makes it easy to manage and maintain your blockchain network. It manages your certificates, lets you easily invite new members to join the network, and tracks operational metrics such as usage of compute, memory, and storage resources. In addition, Managed Blockchain can replicate an immutable copy of your blockchain network activity into Amazon Quantum Ledger Database (QLDB), a fully managed ledger database. This allows you to easily analyze the network activity outside the network and gain insights into trends.

Business Applications

Alexa for Business

Alexa for Business is a service that enables organizations and employees to use Alexa to get more work done. With Alexa for Business, employees can use Alexa as their intelligent assistant to be more productive in meeting rooms, at their desks, and even with the Alexa devices they already have at home.

Amazon WorkDocs

Amazon WorkDocs is a fully managed, secure enterprise storage and sharing service with strong administrative controls and feedback capabilities that improve user productivity.

Users can comment on files, send them to others for feedback, and upload new versions without having to resort to emailing multiple versions of their files as attachments. Users can take advantage of these capabilities wherever they are, using the device of their choice, including PCs, Macs, tablets, and phones. Amazon WorkDocs offers IT administrators the option of integrating with existing corporate directories, flexible sharing policies and control of the location where data is stored. You can get started using Amazon WorkDocs with a 30-day free trial providing 1 TB of storage per user for up to 50 users.

Amazon WorkMail

Amazon WorkMail is a secure, managed business email and calendar service with support for existing desktop and mobile email client applications. Amazon WorkMail gives users the ability to seamlessly access their email, contacts, and calendars using the client application of their choice, including Microsoft Outlook, native iOS and Android email applications, any client application supporting the IMAP protocol, or directly through a web browser. You can integrate Amazon WorkMail with your existing corporate directory, use email journaling to meet compliance requirements, and control both the keys that encrypt your data and the location in which your data is stored. You can also set up interoperability with Microsoft Exchange Server, and programmatically manage users, groups, and resources using the Amazon WorkMail SDK.

Amazon Chime

Amazon Chime is a communications service that transforms online meetings with a secure, easy-to-use application that you can trust. Amazon Chime works seamlessly across your devices so that you can stay connected. You can use Amazon Chime for online meetings, video conferencing, calls, chat, and to share content, both inside and outside your organization.

Amazon Chime works with Alexa for Business, which means you can use Alexa to start your meetings with your voice. Alexa can start your video meetings in large conference rooms, and automatically dial into online meetings in smaller huddle rooms and from your desk.

Compute

Amazon EC2

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

The Amazon EC2 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 (called Amazon EC2 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. Amazon EC2 provides developers and system administrators the tools to build failure resilient applications and isolate themselves from common failure scenarios.

Instance types

Amazon EC2 passes on to you the financial benefits of Amazon's scale. You pay a very low rate for the compute capacity you actually consume. See Amazon EC2 Instance Purchasing Options for a more detailed description.

- **On-Demand Instances**—With On-Demand instances, you pay for compute capacity by the hour with no long-term commitments. You can increase or decrease your compute capacity depending on the demands of your application and only pay the specified hourly rate for the instances you use. The use of On-Demand instances frees you from the costs and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs. On-Demand instances also remove the need to buy "safety net" capacity to handle periodic traffic spikes.
- **Reserved Instances**—Reserved Instances provide you with a significant discount (up to 75%) compared to On-Demand instance pricing. You have the flexibility to change families, operating system types, and tenancies while benefitting from Reserved Instance pricing when you use Convertible Reserved Instances.

- **Spot Instances**—Spot Instances are available at up to a 90% discount compared to On-Demand prices and let you take advantage of unused EC2 capacity in the AWS Cloud. You can significantly reduce the cost of running your applications, grow your application's compute capacity and throughput for the same budget, and enable new types of cloud computing applications.

Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling helps you maintain application availability and allows you to automatically add or remove EC2 instances according to conditions you define. You can use the fleet management features of Amazon EC2 Auto Scaling to maintain the health and availability of your fleet. You can also use the dynamic and predictive scaling features of Amazon EC2 Auto Scaling to add or remove EC2 instances. Dynamic scaling responds to changing demand and predictive scaling automatically schedules the right number of EC2 instances based on predicted demand. Dynamic scaling and predictive scaling can be used together to scale faster.

Amazon Elastic Container Registry

Amazon Elastic Container Registry (Amazon ECR) is a fully-managed Docker container registry that makes it easy for developers to store, manage, and deploy Docker container images. Amazon ECR is integrated with Amazon Elastic Container Service (Amazon ECS), 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 reliably deploy containers for your applications. Integration with AWS Identity and Access Management (IAM) provides resource-level control of each repository. With Amazon ECR, there are no upfront fees or commitments. You pay only for the amount of data you store in your repositories and data transferred to the Internet.

Amazon Elastic Container Service

Amazon Elastic Container Service (Amazon ECS) is a highly scalable, high-performance container orchestration service that supports Docker containers and allows you to easily run and scale containerized applications on AWS. Amazon ECS eliminates the need for you to install and operate your own

container orchestration software, manage and scale a cluster of virtual machines, or schedule containers on those virtual machines.

With simple API calls, you can launch and stop Docker-enabled applications, query the complete state of your application, and access many familiar features such as IAM roles, security groups, load balancers, Amazon CloudWatch Events, AWS CloudFormation templates, and AWS CloudTrail logs.

Amazon Elastic Container Service for Kubernetes

Amazon Elastic Container Service for Kubernetes (Amazon EKS) makes it easy to deploy, manage, and scale containerized applications using Kubernetes on AWS.

Amazon EKS runs the Kubernetes management infrastructure for you across multiple AWS availability zones to eliminate a single point of failure. Amazon EKS is certified Kubernetes conformant so you can use existing tooling and plugins from partners and the Kubernetes community. Applications running on any standard Kubernetes environment are fully compatible and can be easily migrated to Amazon EKS.

Amazon Lightsail

Amazon Lightsail is designed to be the easiest way to launch and manage a virtual private server with AWS. Lightsail plans include everything you need to jumpstart your project – a virtual machine, SSD- based storage, data transfer, DNS management, and a static IP address – for a low, predictable price.

AWS Batch

AWS Batch enables developers, scientists, and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS. AWS Batch dynamically provisions the optimal quantity and type of compute resources (e.g., CPU or memory-optimized instances) based on the volume and specific resource requirements of the batch jobs submitted. With AWS Batch, there is no need to install and manage batch computing software or server clusters that you use to run your jobs, allowing you to focus on analyzing results and solving problems. AWS Batch plans, schedules, and executes your batch computing workloads across the full range of AWS compute services and features, such as Amazon EC2 and Spot Instances.

AWS Elastic Beanstalk

AWS Elastic Beanstalk is an easy-to-use service for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker on familiar servers such as Apache, Nginx, Passenger, and Internet Information Services (IIS).

You can simply upload your code, and AWS Elastic Beanstalk automatically handles the deployment, from capacity provisioning, load balancing, and auto scaling to application health monitoring. At the same time, you retain full control over the AWS resources powering your application and can access the underlying resources at any time.

AWS Fargate

AWS Fargate is a compute engine for Amazon ECS that allows you to run containers without having to manage servers or clusters. With AWS Fargate, you no longer have to provision, configure, and scale clusters of virtual machines to run containers. This removes the need to choose server types, decide when to scale your clusters, or optimize cluster packing. AWS Fargate removes the need for you to interact with or think about servers or clusters. Fargate lets you focus on designing and building your applications instead of managing the infrastructure that runs them.

Amazon ECS has two modes: Fargate launch type and EC2 launch type. With Fargate launch type, all you have to do is package your application in containers, specify the CPU and memory requirements, define networking and IAM policies, and launch the application. EC2 launch type allows you to have server-level, more granular control over the infrastructure that runs your container applications. With EC2 launch type, you can use Amazon ECS to manage a cluster of servers and schedule placement of containers on the servers. Amazon ECS keeps track of all the CPU, memory and other resources in your cluster, and also finds the best server for a container to run on based on your specified resource requirements. You are responsible for provisioning, patching, and scaling clusters of servers. You can decide which type of server to use, which applications and how many containers to run in a cluster to optimize utilization, and when you should add or remove servers from a cluster. EC2 launch type gives you more control of your server clusters and provides a broader range of customization options, which might be required to support some specific applications or possible compliance and government requirements.

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 you can call it directly from any web or mobile app.

AWS Serverless Application Repository

The AWS Serverless Application Repository enables you to quickly deploy code samples, components, and complete applications for common use cases such as web and mobile back-ends, event and data processing, logging, monitoring, IoT, and more. 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.

You can also use the Serverless Application Repository to publish your own applications and share them within your team, across your organization, or with the community at large. To share an application you've built, publish it to the AWS Serverless Application Repository.

AWS Outposts

AWS Outposts bring native AWS services, infrastructure, and operating models to virtually any data center, co-location space, or on-premises facility. You can use the same APIs, the same tools, the same hardware, and the same functionality across on-premises and the cloud to deliver a truly consistent hybrid experience. Outposts can be used to support workloads that need to remain on-premises due to low latency or local data processing needs.

AWS Outposts come in two variants: 1) VMware Cloud on AWS Outposts allows you to use the same VMware control plane and APIs you use to run your infrastructure, 2) AWS native variant of AWS Outposts allows you to use the same exact APIs and control plane you use to run in the AWS cloud, but on-premises.

AWS Outposts infrastructure is fully managed, maintained, and supported by AWS to deliver access to the latest AWS services. Getting started is easy, you simply log into the AWS Management Console to order your Outposts servers, choosing from a wide range of compute and storage options. You can order one or more servers, or quarter, half, and full rack units.

VMware Cloud on AWS

VMware Cloud on AWS is an integrated cloud offering jointly developed by AWS and VMware delivering a highly scalable, secure and innovative service that allows organizations to seamlessly migrate and extend their on-premises VMware vSphere-based environments to the AWS Cloud running on next-generation Amazon Elastic Compute Cloud (Amazon EC2) bare metal infrastructure. VMware Cloud on AWS is ideal for enterprise IT infrastructure and operations organizations looking to migrate their on-premises vSphere-based workloads to the public cloud, consolidate and extend their data center capacities, and optimize, simplify and modernize their disaster recovery solutions. VMware Cloud on AWS is delivered, sold, and supported globally by VMware and its partners with availability in the following AWS Regions: US East (N. Virginia), US West (Oregon), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), and Europe (London). With each release, VMware Cloud on AWS availability will expand into additional global regions.

VMware Cloud on AWS brings the broad, diverse and rich innovations of AWS services natively to the enterprise applications running on VMware's compute, storage and network virtualization platforms. This allows organizations to easily and rapidly add new innovations to their enterprise applications by natively integrating AWS infrastructure and platform capabilities such as AWS Lambda, Amazon Simple Queue Service (SQS), Amazon S3, Elastic Load Balancing, Amazon RDS, Amazon DynamoDB, Amazon Kinesis and Amazon Redshift, among many others.

With VMware Cloud on AWS, organizations can simplify their Hybrid IT operations by using the same VMware Cloud Foundation technologies including vSphere, vSAN, NSX, and vCenter Server across their on-premises data centers and on the AWS Cloud without having to purchase any new or custom hardware, rewrite applications, or modify their operating models. The service automatically provisions infrastructure and provides full VM compatibility and workload portability between your on-premises environments and the AWS

Cloud. With VMware Cloud on AWS, you can leverage AWS's breadth of services, including compute, databases, analytics, Internet of Things (IoT), security, mobile, deployment, application services, and more.

Customer Engagement

Amazon Connect

Amazon Connect is a self-service, cloud-based contact center service that makes it easy for any business to deliver better customer service at lower cost. Amazon Connect is based on the same contact center technology used by Amazon customer service associates around the world to power millions of customer conversations. The self-service graphical interface in Amazon Connect makes it easy for non-technical users to design contact flows, manage agents, and track performance metrics – no specialized skills required. There are no up-front payments or long-term commitments and no infrastructure to manage with Amazon Connect; customers pay by the minute for Amazon Connect usage plus any associated telephony services.

Amazon SES

Amazon Simple Email Service (Amazon SES) is a cloud-based email sending service designed to help digital marketers and application developers send marketing, notification, and transactional emails. It is a reliable, cost-effective service for businesses of all sizes that use email to keep in contact with their customers.

You can use our SMTP interface or one of the AWS SDKs to integrate Amazon SES directly into your existing applications. You can also integrate the email sending capabilities of Amazon SES into the software you already use, such as ticketing systems and email clients.

See also Amazon Pinpoint.

Database

Amazon Aurora

Amazon Aurora is a MySQL and PostgreSQL compatible relational database engine that combines the speed and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases.

Amazon Aurora is up to five times faster than standard MySQL databases and three times faster than standard PostgreSQL databases. It provides the security, availability, and reliability of commercial databases at 1/10th the cost. Amazon Aurora is fully managed by Amazon Relational Database Service (RDS), which automates time-consuming administration tasks like hardware provisioning, database setup, patching, and backups.

Amazon Aurora features a distributed, fault-tolerant, self-healing storage system that auto-scales up to 64TB 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 S3, and replication across three Availability Zones (AZs).

Amazon RDS

Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while automating time-consuming administration tasks such as hardware provisioning, database setup, patching and backups. It frees you to focus on your applications so you can give them the fast performance, high availability, security and compatibility they need.

Amazon RDS is available on several database instance types - optimized for memory, performance or I/O - and provides you with six familiar database engines to choose from, including Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle Database, and SQL Server. You can use the AWS Database Migration Service to easily migrate or replicate your existing databases to Amazon RDS.

Amazon RDS on VMware

Amazon Relational Database Service (RDS) on VMware lets you deploy managed databases in on-premises VMware environments using the Amazon RDS technology enjoyed by hundreds of thousands of AWS customers. Amazon RDS provides cost-efficient and resizable capacity while automating time-consuming administration tasks including hardware provisioning, database setup, patching, and backups, freeing you to focus on your applications. RDS on VMware brings these same benefits to your on-premises deployments, making it easy to set up, operate, and scale databases in VMware vSphere private data centers, or to migrate them to AWS.

RDS on VMware allows you to utilize the same simple interface for managing databases in on-premises VMware environments as you would use in AWS. You can easily replicate RDS on VMware databases to RDS instances in AWS, enabling low-cost hybrid deployments for disaster recovery, read replica bursting, and optional long-term backup retention in Amazon Simple Storage Service (Amazon S3).

Amazon DynamoDB

Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale. It's a fully managed, multiregion, multimaster database with built-in security, backup and restore, and in-memory caching for internet-scale applications. DynamoDB can handle more than 10 trillion requests per day and support peaks of more than 20 million requests per second.

Many of the world's fastest growing businesses such as Lyft, Airbnb, and Redfin as well as enterprises such as Samsung, Toyota, and Capital One depend on the scale and performance of DynamoDB to support their mission-critical workloads.

More than 100,000 AWS customers have chosen DynamoDB as their key-value and document database for mobile, web, gaming, ad tech, IoT, and other applications that need low-latency data access at any scale. Create a new table for your application and let DynamoDB handle the rest.

Amazon ElastiCache

Amazon ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud. The service improves the performance of web applications by allowing you to retrieve information from fast, managed, in-memory caches, instead of relying entirely on slower disk-based databases.

Amazon ElastiCache supports two open-source in-memory caching engines:

- Redis - a fast, open source, in-memory data store and cache. Amazon ElastiCache for Redis is a Redis- compatible in-memory service that delivers the ease-of-use and power of Redis along with the availability, reliability, and performance suitable for the most demanding applications. Both single- node and up to 15-shard clusters are available, enabling scalability to up to 3.55 TiB of in-memory data. ElastiCache for Redis is fully managed, scalable, and secure. This makes it an ideal candidate to power high-performance use cases such as web, mobile apps, gaming, ad-tech, and IoT.
- Memcached - a widely adopted memory object caching system. ElastiCache for Memcached is protocol compliant with Memcached, so popular tools that you use today with existing Memcached environments will work seamlessly with the service.

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 Amazon Neptune is a purpose-built, high-performance graph database engine optimized for storing billions of relationships and querying the graph with milliseconds latency. Amazon Neptune supports popular graph models Property Graph and W3C's RDF, and their respective query languages Apache TinkerPop Gremlin and SPARQL, allowing you to easily 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.

Amazon Neptune is highly available, with read replicas, point-in-time recovery, continuous backup to Amazon S3, and replication across Availability Zones. Neptune is secure with support for encryption at rest. Neptune is fully-managed, so you no longer need to worry about database management tasks such as hardware provisioning, software patching, setup, configuration, or backups.

Amazon Quantum Ledger Database (QLDB)

Amazon QLDB is a fully managed ledger database that provides a transparent, immutable, and cryptographically verifiable transaction log owned by a central trusted authority. Amazon QLDB tracks each and every application data change and maintains 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, tracking the history of credits and debits in banking transactions, verifying the data lineage of an insurance claim, or tracing 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. However, building audit functionality with relational databases is time-consuming and prone to human error. It requires custom development, and since relational databases are not inherently immutable, any unintended changes to the data are hard to track and verify. Alternatively, blockchain frameworks, such as Hyperledger Fabric and Ethereum, can also be used as a ledger. However, this adds complexity as you need to set-up an entire blockchain network with multiple nodes, manage its infrastructure, and require the nodes to validate each transaction before it can be added to the ledger.

Amazon QLDB is a new class of database that eliminates the need to engage in the complex development effort of building your own ledger-like applications. With QLDB, your data's change history is immutable – it cannot be altered or deleted – and using cryptography, you can easily verify that there have been no unintended modifications to your application's data. QLDB uses an immutable transactional log, known as a journal, that tracks each application data change and maintains a complete and verifiable history of changes over time. QLDB is easy to use because it provides developers with a familiar SQL-like API, a flexible document data model, and full support for transactions. QLDB is also serverless, so it automatically scales to support the demands of your application. There are no servers to manage and no read or write limits to configure. With QLDB, you only pay for what you use.

Amazon Timestream

Amazon Timestream is a fast, scalable, fully managed time series database service for IoT and operational applications that makes it easy to store and analyze trillions of events per day at 1/10th the cost of relational databases. Driven by the rise of IoT devices, IT systems, and smart industrial machines, time-series data — data that measures how things change over time — is one of the fastest growing data types. Time-series data has specific characteristics such as typically arriving in time order form, data is append-only, and queries are always over a time interval. While relational databases can store this data, they are inefficient at processing this data as they lack optimizations such as

storing and retrieving data by time intervals. Timestream is a purpose-built time series database that efficiently stores and processes this data by time intervals. With Timestream, you can easily store and analyze log data for DevOps, sensor data for IoT applications, and industrial telemetry data for equipment maintenance. As your data grows over time, Timestream's adaptive query processing engine understands its location and format, making your data simpler and faster to analyze. Timestream also automates rollups, retention, tiering, and compression of data, so you can manage your data at the lowest possible cost. Timestream is serverless, so there are no servers to manage. It manages time-consuming tasks such as server provisioning, software patching, setup, configuration, or data retention and tiering, freeing you to focus on building your applications.

Amazon DocumentDB

Amazon DocumentDB (with MongoDB compatibility) is a fast, scalable, highly available, and fully managed document database service that supports MongoDB workloads.

Amazon DocumentDB is designed from the ground-up to give you the performance, scalability, and availability you need when operating mission-critical MongoDB workloads at scale. Amazon DocumentDB implements the Apache 2.0 open source MongoDB 3.6 API by emulating the responses that a MongoDB client expects from a MongoDB server, allowing you to use your existing MongoDB drivers and tools with Amazon DocumentDB.

Desktop and App Streaming

Amazon WorkSpaces

Amazon WorkSpaces is a fully managed, secure cloud desktop service. You can use Amazon WorkSpaces to provision either Windows or Linux desktops in just a few minutes and quickly scale to provide thousands of desktops to workers across the globe. You can pay either monthly or hourly, just for the WorkSpaces you launch, which helps you save money when compared to traditional desktops and on-premises VDI solutions. Amazon WorkSpaces helps you eliminate the complexity in managing hardware inventory, OS versions and patches, and Virtual Desktop Infrastructure (VDI), which helps simplify your desktop delivery strategy. With Amazon WorkSpaces, your users get a fast, responsive desktop of their choice that they can access anywhere, anytime, from any supported device.

Amazon AppStream 2.0

Amazon AppStream 2.0 is a fully managed application streaming service. You centrally manage your desktop applications on AppStream 2.0 and securely deliver them to any computer. You can easily scale to any number of users across the globe without acquiring, provisioning, and operating hardware or infrastructure. AppStream 2.0 is built on AWS, so you benefit from a data center and network architecture designed for the most security-sensitive organizations. Each user has a fluid and responsive experience with your applications, including GPU-intensive 3D design and engineering ones, because your applications run on virtual machines (VMs) optimized for specific use cases and each streaming session automatically adjusts to network conditions.

Enterprises can use AppStream 2.0 to simplify application delivery and complete their migration to the cloud. Educational institutions can provide every student access to the applications they need for class on any computer. Software vendors can use AppStream 2.0 to deliver trials, demos, and training for their applications with no downloads or installations. They can also develop a full software-as-a-service (SaaS) solution without rewriting their application.

Developer Tools

AWS CodeCommit

AWS CodeCommit is a fully-managed source control service that hosts secure Git-based repositories. It makes it easy for teams to collaborate on code in a secure and highly scalable ecosystem. CodeCommit eliminates the need to operate your own source control system or worry about scaling its infrastructure. You can use CodeCommit to securely store anything from source code to binaries, and it works seamlessly with your existing Git tools.

AWS CodeBuild

AWS CodeBuild is a fully managed build service that compiles source code, runs tests, and produces software packages that are ready to deploy. With CodeBuild, you don’t need to provision, manage, and scale your own build servers. CodeBuild scales continuously and processes multiple builds concurrently, so your builds are not left waiting in a queue. You can get started quickly by using prepackaged build environments, or you can create custom build environments that use your own build tools.

AWS CodeDeploy

AWS CodeDeploy is a service that automates code deployments to any instance, including EC2 instances and instances running on premises. AWS CodeDeploy makes it easier for you to rapidly release new features, helps you avoid downtime during application deployment, and handles the complexity of updating your applications. You can use AWS CodeDeploy to automate software deployments, eliminating the need for error-prone manual operations. The service scales with your infrastructure so you can easily deploy to one instance or thousands.

AWS CodePipeline

AWS CodePipeline is a fully managed continuous delivery service that helps you automate your release pipelines for fast and reliable application and infrastructure updates. CodePipeline automates the build, test, and deploy phases of your release process every time there is a code change, based on the release model you define. This enables you to rapidly and reliably deliver features and updates. You can easily integrate AWS CodePipeline with third-party services such as GitHub or with your own custom plugin. With AWS CodePipeline, you only pay for what you use. There are no upfront fees or long-term commitments.

AWS CodeStar

AWS CodeStar enables you to quickly develop, build, and deploy applications on AWS. AWS CodeStar provides a unified user interface, enabling you to easily manage your software development activities in one place. With AWS CodeStar, you can set up your entire continuous delivery toolchain in minutes, allowing you to start releasing code faster. AWS CodeStar makes it easy for your whole team to work together securely, allowing you to easily manage access and add owners, contributors, and viewers to your projects. Each AWS CodeStar project comes with a project management dashboard, including an integrated issue tracking capability powered by Atlassian JIRA Software. With the AWS CodeStar project dashboard, you can easily track progress across your entire software development process, from your backlog of work items to teams' recent code deployments. For more information, see AWS CodeStar features.

Amazon Corretto

Amazon Corretto is a no-cost, multiplatform, production-ready distribution of the Open Java Development Kit (OpenJDK). Corretto comes with long-term support that will include performance enhancements and security fixes. Amazon runs Corretto internally on thousands of production services and Corretto is certified as compatible with the Java SE standard. With Corretto, you can develop and run Java applications on popular operating systems, including Amazon Linux 2, Windows, and macOS. Amazon Corretto 8 is in Preview.

AWS Cloud9

AWS Cloud9 is a cloud-based integrated development environment (IDE) that lets you write, run, and debug your code with just a browser. It includes a code editor, debugger, and terminal. Cloud9 comes prepackaged with essential tools for popular programming languages, including JavaScript, Python, PHP, and more, so you don't need to install files or configure your development machine to start new projects. Since your Cloud9 IDE is cloud-based, you can work on your projects from your office, home, or anywhere using an internet-connected machine. Cloud9 also provides a seamless experience for developing serverless applications enabling you to easily define resources, debug, and switch between local and remote execution of serverless applications. With Cloud9, you can quickly share your development environment with your team, enabling you to pair program and track each other's inputs in real time.

AWS X-Ray

AWS X-Ray helps developers analyze and debug distributed applications in production or under development, such as those built using a microservices architecture. With X-Ray, you can understand how your application and its underlying services are performing so you can identify and troubleshoot the root cause of performance issues and errors. X-Ray provides an end-to-end view of requests as they travel through your application, and shows a map of your application's underlying components. You can use X- Ray to analyze both applications in development and in production, from simple three-tier applications to complex microservices applications consisting of thousands of services.

Game Tech

Amazon GameLift

Amazon GameLift is a managed service for deploying, operating, and scaling dedicated game servers for session-based multiplayer games. Amazon GameLift makes it easy to manage server infrastructure, scale capacity to lower latency and cost, match players into available game sessions, and defend from distributed denial-of-service (DDoS) attacks. You pay for the compute resources and bandwidth your games actually use, without monthly or annual contracts.

Amazon Lumberyard

Amazon Lumberyard is a free, cross-platform, 3D game engine for you to create the highest-quality games, connect your games to the vast compute and storage of the AWS Cloud, and engage fans on Twitch. By starting game projects with Lumberyard, you can spend more of your time creating great gameplay and building communities of fans, and less time on the undifferentiated heavy lifting of building a game engine and managing server infrastructure.

Internet of Things (IoT)

AWS IoT Core

AWS IoT Core is a managed cloud service 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.

AWS IoT Core makes it easy to use AWS services like AWS Lambda, Amazon Kinesis, Amazon S3, Amazon SageMaker, Amazon DynamoDB, Amazon CloudWatch, AWS CloudTrail, and Amazon QuickSight to build Internet of Things (IoT) applications that gather, process, analyze and act on data generated by connected devices, without having to manage any infrastructure.

Amazon FreeRTOS

Amazon FreeRTOS (a:FreeRTOS) is an operating system for microcontrollers that makes small, low-power edge devices easy to program, deploy, secure, connect, and manage. Amazon FreeRTOS extends the FreeRTOS kernel, a popular open source operating system for microcontrollers, with software libraries that make it easy to securely connect your small, low-power devices to AWS cloud services like AWS IoT Core or to more powerful edge devices running AWS IoT Greengrass.

A microcontroller (MCU) is a single chip containing a simple processor that can be found in many devices, including appliances, sensors, fitness trackers, industrial automation, and automobiles. Many of these small devices could benefit from connecting to the cloud or locally to other devices. For example, smart electricity meters need to connect to the cloud to report on usage, and building security systems need to communicate locally so that a door will unlock when you badge in. Microcontrollers have limited compute power and memory capacity and typically perform simple, functional tasks. Microcontrollers frequently run operating systems that do not have built-in functionality to connect to local networks or the cloud, making IoT applications a challenge. Amazon FreeRTOS helps solve this problem by providing both the core operating system (to run the edge device) as well as software libraries that make it easy to securely connect to the cloud (or other edge devices) so you can collect data from them for IoT applications and take action.

AWS IoT Greengrass

AWS IoT Greengrass seamlessly extends AWS to devices so they can act locally on the data they generate, while still using the cloud for management, analytics, and durable storage. With AWS IoT Greengrass, connected devices can run AWS Lambda functions, execute predictions based on machine learning models, keep device data in sync, and communicate with other devices securely – even when not connected to the Internet.

With AWS IoT Greengrass, you can use familiar languages and programming models to create and test your device software in the cloud, and then deploy it to your devices. AWS IoT Greengrass can be programmed to filter device data and only transmit necessary information back to the cloud. You can also connect to third-party applications, on-premises software, and AWS services out-of-the-box with AWS IoT Greengrass Connectors. Connectors also jumpstart device onboarding with pre-built protocol adapter integrations and

allow you to streamline authentication via integration with AWS Secrets Manager.

AWS IoT 1-Click

AWS IoT 1-Click is a service that enables simple devices to trigger AWS Lambda functions that can execute an action. AWS IoT 1-Click supported devices enable you to easily perform actions such as notifying technical support, tracking assets, and replenishing goods or services. AWS IoT 1-Click supported devices are ready for use right out of the box and eliminate the need for writing your own firmware or configuring them for secure connectivity. AWS IoT 1-Click supported devices can be easily managed. You can easily create device groups and associate them with a Lambda function that executes your desired action when triggered. You can also track device health and activity with the pre-built reports.

AWS IoT Analytics

AWS IoT Analytics is a fully-managed service that makes it easy to run and operationalize sophisticated analytics on massive volumes of IoT data without having to worry about the cost and complexity typically required to build an IoT analytics platform. It is the easiest way to run analytics on IoT data and get insights to make better and more accurate decisions for IoT applications and machine learning use cases.

IoT data is highly unstructured which makes it difficult to analyze with traditional analytics and business intelligence tools that are designed to process structured data. IoT data comes from devices that often record fairly noisy processes (such as temperature, motion, or sound). The data from these devices can frequently have significant gaps, corrupted messages, and false readings that must be cleaned up before analysis can occur. Also, IoT data is often only meaningful in the context of additional, third party data inputs. For example, to help farmers determine when to water their crops, vineyard irrigation systems often enrich moisture sensor data with rainfall data from the vineyard, allowing for more efficient water usage while maximizing harvest yield.

AWS IoT Analytics automates each of the difficult steps that are required to analyze data from IoT devices. AWS IoT Analytics filters, transforms, and enriches IoT data before storing it in a time-series data store for analysis. You can setup the service to collect only the data you need from your devices, apply mathematical transforms to process the data, and enrich the data with device-

specific metadata such as device type and location before storing the processed data. Then, you can analyze your data by running ad hoc or scheduled queries using the built-in SQL query engine, or perform more complex analytics and machine learning inference. AWS IoT Analytics makes it easy to get started with machine learning by including pre-built models for common IoT use cases.

You can also use your own custom analysis, packaged in a container, to execute on AWS IoT Analytics. AWS IoT Analytics automates the execution of your custom analyses created in Jupyter Notebook or your own tools (such as Matlab, Octave, etc.) to be executed on your schedule.

AWS IoT Analytics is a fully managed service that operationalizes analyses and scales automatically to support up to petabytes of IoT data. With AWS IoT Analytics, you can analyze data from millions of devices and build fast, responsive IoT applications without managing hardware or infrastructure.

AWS IoT Button

The AWS IoT Button is a programmable button based on the Amazon Dash Button hardware. This simple Wi-Fi device is easy to configure, and it's designed for developers to get started with AWS IoT Core, AWS Lambda, Amazon DynamoDB, Amazon SNS, and many other Amazon Web Services without writing device-specific code.

You can code the button's logic in the cloud to configure button clicks to count or track items, call or alert someone, start or stop something, order services, or even provide feedback. For example, you can click the button to unlock or start a car, open your garage door, call a cab, call your spouse or a customer service representative, track the use of common household chores, medications or products, or remotely control your home appliances.

The button can be used as a remote control for Netflix, a switch for your Philips Hue light bulb, a check-in/check-out device for Airbnb guests, or a way to order your favorite pizza for delivery. You can integrate it with third-party APIs like Twitter, Facebook, Twilio, Slack or even your own company's applications. Connect it to things we haven't even thought of yet.

AWS IoT Device Defender

AWS IoT Device Defender is a fully managed service that helps you secure your fleet of IoT devices. AWS IoT Device Defender continuously audits your IoT configurations to make sure that they aren't deviating from security best practices. A configuration is a set of technical controls you set to help keep information secure when devices are communicating with each other and the cloud. AWS IoT Device Defender makes it easy to maintain and enforce IoT configurations, such as ensuring device identity, authenticating and authorizing devices, and encrypting device data. AWS IoT Device Defender continuously audits the IoT configurations on your devices against a set of predefined security best practices. AWS IoT Device Defender sends an alert if there are any gaps in your IoT configuration that might create a security risk, such as identity certificates being shared across multiple devices or a device with a revoked identity certificate trying to connect to AWS IoT Core.

AWS IoT Device Defender also lets you continuously monitor security metrics from devices and AWS IoT Core for deviations from what you have defined as appropriate behavior for each device. If something doesn't look right, AWS IoT Device Defender sends out an alert so you can take action to remediate the issue. For example, traffic spikes in outbound traffic might indicate that a device is participating in a DDoS attack. AWS IoT Greengrass and Amazon FreeRTOS automatically integrate with AWS IoT Device Defender to provide security metrics from the devices for evaluation.

AWS IoT Device Defender can send alerts to the AWS IoT Console, Amazon CloudWatch, and Amazon SNS. If you determine that you need to take an action based on an alert, you can use AWS IoT Device Management to take mitigating actions such as pushing security fixes.

AWS IoT Device Management

As many IoT deployments consist of hundreds of thousands to millions of devices, it is essential to track, monitor, and manage connected device fleets. You need to ensure your IoT devices work properly and securely after they have been deployed. You also need to secure access to your devices, monitor health, detect and remotely troubleshoot problems, and manage software and firmware updates.

AWS IoT Device Management makes it easy to securely onboard, organize, monitor, and remotely manage IoT devices at scale. With AWS IoT Device

Management, you can register your connected devices individually or in bulk, and easily manage permissions so that devices remain secure. You can also organize your devices, monitor and troubleshoot device functionality, query the state of any IoT device in your fleet, and send firmware updates over-the-air (OTA). AWS IoT Device Management is agnostic to device type and OS, so you can manage devices from constrained microcontrollers to connected cars all with the same service. AWS IoT Device Management allows you to scale your fleets and reduce the cost and effort of managing large and diverse IoT device deployments.

AWS IoT Events

AWS IoT Events is a fully managed IoT service that makes it easy to detect and respond to events from IoT sensors and applications. Events are patterns of data identifying more complicated circumstances than expected, such as changes in equipment when a belt is stuck or connected motion detectors using movement signals to activate lights and security cameras. To detect events before AWS IoT Events, you had to build costly, custom applications to collect data, apply decision logic to detect an event, and then trigger another application to react to the event. Using AWS IoT Events, it's simple to detect events across thousands of IoT sensors sending different telemetry data, such as temperature from a freezer, humidity from respiratory equipment, and belt speed on a motor, and hundreds of equipment management applications. You simply select the relevant data sources to ingest, define the logic for each event using simple 'if-then-else' statements, and select the alert or custom action to trigger when an event occurs. AWS IoT Events continuously monitors data from multiple IoT sensors and applications, and it integrates with other services, such as AWS IoT Core and AWS IoT Analytics, to enable early detection and unique insights into events. AWS IoT Events automatically triggers alerts and actions in response to events based on the logic you define. This helps resolve issues quickly, reduce maintenance costs, and increase operational efficiency.

AWS IoT SiteWise

AWS IoT SiteWise is a managed service that makes it easy to collect and organize data from industrial equipment at scale. You can easily monitor equipment across your industrial facilities to identify waste, such as breakdown of equipment and processes, production inefficiencies, and defects in products. Today, getting performance metrics from industrial equipment is tough because data is often locked into proprietary on-premises data stores and typically

requires specialized expertise to retrieve it and put it in a format that is useful for searching and analysis. IoT SiteWise simplifies this process by providing software running on a gateway that resides in your facilities and automates the process of collecting and organizing industrial equipment data. This gateway securely connects to your on-premises data servers, collects data, and sends the data to the AWS Cloud. You can run the IoT SiteWise software on an AWS Snowball Edge gateway or install the IoT SiteWise software on popular third-party industrial gateways. These gateways are specifically designed for industrial environments that are likely already in your facilities connecting your industrial equipment.

You can use IoT SiteWise to monitor operations across facilities, quickly compute common industrial performance metrics, and build applications to analyze industrial equipment data, prevent costly equipment issues, and reduce production inefficiencies. With IoT SiteWise, you can focus on understanding and optimizing your operations, rather than building costly in-house data collection and management applications.

AWS IoT Things Graph

AWS IoT Things Graph is a service that makes it easy to visually connect different devices and web services to build IoT applications.

IoT applications are being built today using a variety of devices and web services to automate tasks for a wide range of use cases, such as smart homes, industrial automation, and energy management. Because there aren't any widely adopted standards, it's difficult today for developers to get devices from multiple manufacturers to connect to each other as well as with web services. This forces developers to write lots of code to wire together all of the devices and web services they need for their IoT application. AWS IoT Things Graph provides a visual drag-and-drop interface for connecting and coordinating devices and web services, so you can build IoT applications quickly. For example, in a commercial agriculture application, you can define interactions between humidity, temperature, and sprinkler sensors with weather data services in the cloud to automate watering. You represent devices and services using pre-built reusable components, called models, that hide low-level details, such as protocols and interfaces, and are easy to integrate to create sophisticated workflows.

You can get started with AWS IoT Things Graph using these pre-built models for popular device types, such as switches and programmable logic controllers (PLCs), or create your own custom model using a GraphQL-based schema modeling language, and deploy your IoT application to AWS IoT Greengrass-enabled devices such as cameras, cable set-top boxes, or robotic arms in just a few clicks. IoT Greengrass is software that provides local compute and secure cloud connectivity so devices can respond quickly to local events even without internet connectivity, and runs on a huge range of devices from a Raspberry Pi to a server-level appliance. IoT Things Graph applications run on IoT Greengrass-enabled devices.

AWS Partner Device Catalog

The AWS Partner Device Catalog helps you find devices and hardware to help you explore, build, and go to market with your IoT solutions. Search for and find hardware that works with AWS, including development kits and embedded systems to build new devices, as well as off-the-shelf-devices such as gateways, edge servers, sensors, and cameras for immediate IoT project integration. The choice of AWS enabled hardware from our curated catalog of devices from APN partners can help make the rollout of your IoT projects easier. All devices listed in the AWS Partner Device Catalog are also available for purchase from our partners to get you started quickly.

Machine Learning

Amazon SageMaker

Amazon SageMaker is a fully-managed platform that enables developers and data scientists to quickly and easily build, train, and deploy machine learning models at any scale. Amazon SageMaker removes all the barriers that typically slow down developers who want to use machine learning.

Machine learning often feels a lot harder than it should be to most developers because the process to build and train models, and then deploy them into production is too complicated and too slow. First, you need to collect and prepare your training data to discover which elements of your data set are important. Then, you need to select which algorithm and framework you'll use. After deciding on your approach, you need to teach the model how to make predictions by training, which requires a lot of compute. Then, you need to tune the model so it delivers the best possible predictions, which is often a tedious

and manual effort. After you've developed a fully trained model, you need to integrate the model with your application and deploy this application on infrastructure that will scale. All of this takes a lot of specialized expertise, access to large amounts of compute and storage, and a lot of time to experiment and optimize every part of the process. In the end, it's not a surprise that the whole thing feels out of reach for most developers.

Amazon SageMaker removes the complexity that holds back developer success with each of these steps. Amazon SageMaker includes modules that can be used together or independently to build, train, and deploy your machine learning models.

Amazon SageMaker Ground Truth

Amazon SageMaker Ground Truth helps you build highly accurate training datasets for machine learning quickly. SageMaker Ground Truth offers easy access to public and private human labelers and provides them with built-in workflows and interfaces for common labeling tasks. Additionally, SageMaker Ground Truth can lower your labeling costs by up to 70% using automatic labeling, which works by training Ground Truth from data labeled by humans so that the service learns to label data independently.

Successful machine learning models are built on the shoulders of large volumes of high-quality training data. But, the process to create the training data necessary to build these models is often expensive, complicated, and time-consuming. The majority of models created today require a human to manually label data in a way that allows the model to learn how to make correct decisions. For example, building a computer vision system that is reliable enough to identify objects - such as traffic lights, stop signs, and pedestrians - requires thousands of hours of video recordings that consist of hundreds of millions of video frames. Each one of these frames needs all of the important elements like the road, other cars, and signage to be labeled by a human before any work can begin on the model you want to develop.

Amazon SageMaker Ground Truth significantly reduces the time and effort required to create datasets for training to reduce costs. These savings are achieved by using machine learning to automatically label data. The model is able to get progressively better over time by continuously learning from labels created by human labelers.

Where the labeling model has high confidence in its results based on what it has learned so far, it will automatically apply labels to the raw data. Where the labeling model has lower confidence in its results, it will pass the data to humans to do the labeling. The human-generated labels are provided back to the labeling model for it to learn from and improve. Over time, SageMaker Ground Truth can label more and more data automatically and substantially speed up the creation of training datasets.

Amazon Comprehend

Amazon Comprehend is a natural language processing (NLP) service that uses machine learning to find insights and relationships in text. No machine learning experience required.

There is a treasure trove of potential sitting in your unstructured data. Customer emails, support tickets, product reviews, social media, even advertising copy represents insights into customer sentiment that can be put to work for your business. The question is how to get at it? As it turns out, Machine learning is particularly good at accurately identifying specific items of interest inside vast swathes of text (such as finding company names in analyst reports), and can learn the sentiment hidden inside language (identifying negative reviews, or positive customer interactions with customer service agents), at almost limitless scale.

Amazon Comprehend uses machine learning to help you uncover the insights and relationships in your unstructured data. The service identifies the language of the text; extracts key phrases, places, people, brands, or events; understands how positive or negative the text is; analyzes text using tokenization and parts of speech; and automatically organizes a collection of text files by topic. You can also use AutoML capabilities in Amazon Comprehend to build a custom set of entities or text classification models that are tailored uniquely to your organization’s needs.

For extracting complex medical information from unstructured text, you can use Amazon Comprehend Medical. The service can identify medical information, such as medical conditions, medications, dosages, strengths, and frequencies from a variety of sources like doctor’s notes, clinical trial reports, and patient health records. Amazon Comprehend Medical also identifies the relationship among the extracted medication and test, treatment and procedure information for easier analysis. For example, the service identifies a particular dosage,

strength, and frequency related to a specific medication from unstructured clinical notes.

Amazon Lex

Amazon Lex is a service for building conversational interfaces into any application using voice and text. Lex provides the advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, to enable you to build applications with highly engaging user experiences and lifelike conversational interactions. With Amazon Lex, the same deep learning technologies that power Amazon Alexa are now available to any developer, enabling you to quickly and easily build sophisticated, natural language, conversational bots ("chatbots").

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 democratizes these deep learning technologies by putting the power of Alexa within reach of all developers. Harnessing these technologies, Amazon Lex enables you to define entirely new categories of products made possible through conversational interfaces.

Amazon Polly

Amazon Polly is a service that turns text into lifelike speech. Polly lets you create applications that talk, enabling you to build entirely new categories of speech-enabled products. Polly is an Amazon artificial intelligence (AI) service that uses advanced deep learning technologies to synthesize speech that sounds like a human voice. Polly includes 47 lifelike voices spread across 24 languages, so you can select the ideal voice and build speech-enabled applications that work in many different countries.

Amazon Polly delivers the consistently fast response times required to support real-time, interactive dialog. You can cache and save Polly's speech audio to replay offline or redistribute. And Polly is easy to use. You simply send the text you want converted into speech to the Polly API, and Polly immediately returns the audio stream to your application so your application can play it directly or store it in a standard audio file format, such as MP3.

With Polly, you only pay for the number of characters you convert to speech, and you can save and replay Polly's generated speech. Polly's low cost per character converted, and lack of restrictions on storage and reuse of voice output, make it a cost-effective way to enable Text-to-Speech everywhere.

Amazon Rekognition

Amazon Rekognition is a service that makes it easy to add image analysis to your applications. With Rekognition, you can detect objects, scenes, and faces in images. You can also search and compare faces. The Amazon Rekognition API enables you to quickly add sophisticated deep-learning-based visual search and image classification to your applications.

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 daily for Prime Photos. Amazon Rekognition uses deep neural network models to detect and label thousands of objects and scenes in your images, and we are continually adding new labels and facial recognition features to the service.

The Amazon Rekognition API lets you easily build powerful visual search and discovery into your applications. With Amazon Rekognition, you only pay for the images you analyze and the face metadata you store. There are no minimum fees, and there are no upfront commitments.

Amazon Translate

Amazon Translate is a neural machine translation service that delivers fast, high-quality, and affordable language translation. Neural machine translation is a form of language translation automation that uses deep learning models to deliver more accurate and more natural sounding translation than traditional statistical and rule-based translation algorithms. Amazon Translate allows you to localize content - such as websites and applications - for international users, and to easily translate large volumes of text efficiently.

Amazon Transcribe

Amazon Transcribe is an automatic speech recognition (ASR) service that makes it easy for developers to add speech-to-text capability to their applications. Using the Amazon Transcribe API, you can analyze audio files stored in Amazon S3 and have the service return a text file of the transcribed

speech. You can also send a live audio stream to Amazon Transcribe and receive a stream of transcripts in real time.

Amazon Transcribe can be used for lots of common applications, including the transcription of customer service calls and generating subtitles on audio and video content. The service can transcribe audio files stored in common formats, like WAV and MP3, with time stamps for every word so that you can easily locate the audio in the original source by searching for the text. Amazon Transcribe is continually learning and improving to keep pace with the evolution of language.

Amazon Elastic Inference

Amazon Elastic Inference allows you to attach low-cost GPU-powered acceleration to Amazon EC2 and Amazon SageMaker instances to reduce the cost of running deep learning inference by up to 75%. Amazon Elastic Inference supports TensorFlow, Apache MXNet, and ONNX models, with more frameworks coming soon.

In most deep learning applications, making predictions using a trained model—a process called inference—can drive as much as 90% of the compute costs of the application due to two factors. First, standalone GPU instances are designed for model training and are typically oversized for inference. While training jobs batch process hundreds of data samples in parallel, most inference happens on a single input in real time that consumes only a small amount of GPU compute. Even at peak load, a GPU's compute capacity may not be fully utilized, which is wasteful and costly. Second, different models need different amounts of GPU, CPU, and memory resources. Selecting a GPU instance type that is big enough to satisfy the requirements of the least used resource often results in under-utilization of the other resources and high costs.

Amazon Elastic Inference solves these problems by allowing you to attach just the right amount of GPU-powered inference acceleration to any EC2 or SageMaker instance type with no code changes. With Amazon Elastic Inference, you can now choose the instance type that is best suited to the overall CPU and memory needs of your application, and then separately configure the amount of inference acceleration that you need to use resources efficiently and to reduce the cost of running inference.

Amazon Forecast

Amazon Forecast is a fully managed service that uses machine learning to deliver highly accurate forecasts.

Companies today use everything from simple spreadsheets to complex financial planning software to attempt to accurately forecast future business outcomes such as product demand, resource needs, or financial performance. These tools build forecasts by looking at a historical series of data, which is called time series data. For example, such tools may try to predict the future sales of a raincoat by looking only at its previous sales data with the underlying assumption that the future is determined by the past. This approach can struggle to produce accurate forecasts for large sets of data that have irregular trends. Also, it fails to easily combine data series that change over time (such as price, discounts, web traffic, and number of employees) with relevant independent variables like product features and store locations.

Based on the same technology used at Amazon.com, Amazon Forecast uses machine learning to combine time series data with additional variables to build forecasts. Amazon Forecast requires no machine learning experience to get started. You only need to provide historical data, plus any additional data that you believe may impact your forecasts. For example, the demand for a particular color of a shirt may change with the seasons and store location. This complex relationship is hard to determine on its own, but machine learning is ideally suited to recognize it. Once you provide your data, Amazon Forecast will automatically examine it, identify what is meaningful, and produce a forecasting model capable of making predictions that are up to 50% more accurate than looking at time series data alone.

Amazon Forecast is a fully managed service, so there are no servers to provision, and no machine learning models to build, train, or deploy. You pay only for what you use, and there are no minimum fees and no upfront commitments.

Amazon Textract

Amazon Textract is a service that automatically extracts text and data from scanned documents. Amazon Textract goes beyond simple optical character recognition (OCR) to also identify the contents of fields in forms and information stored in tables.

Many companies today extract data from documents and forms through manual data entry that's slow and expensive or through simple optical character recognition (OCR) software that is difficult to customize. Rules and workflows for each document and form often need to be hard-coded and updated with each change to the form or when dealing with multiple forms. If the form deviates from the rules, the output is often scrambled and unusable.

Amazon Textract overcomes these challenges by using machine learning to instantly "read" virtually any type of document to accurately extract text and data without the need for any manual effort or custom code. With Textract you can quickly automate document workflows, enabling you to process millions of document pages in hours. Once the information is captured, you can take action on it within your business applications to initiate next steps for a loan application or medical claims processing. Additionally, you can create smart search indexes, build automated approval workflows, and better maintain compliance with document archival rules by flagging data that may require redaction.

Amazon Personalize

Amazon Personalize is a machine learning service that makes it easy for developers to create individualized recommendations for customers using their applications.

Machine learning is being increasingly used to improve customer engagement by powering personalized product and content recommendations, tailored search results, and targeted marketing promotions. However, developing the machine-learning capabilities necessary to produce these sophisticated recommendation systems has been beyond the reach of most organizations today due to the complexity of developing machine learning functionality. Amazon Personalize allows developers with no prior machine learning experience to easily build sophisticated personalization capabilities into their applications, using machine learning technology perfected from years of use on Amazon.com.

With Amazon Personalize, you provide an activity stream from your application – page views, signups, purchases, and so forth – as well as an inventory of the items you want to recommend, such as articles, products, videos, or music. You can also choose to provide Amazon Personalize with additional demographic information from your users such as age, or geographic location. Amazon

Personalize will process and examine the data, identify what is meaningful, select the right algorithms, and train and optimize a personalization model that is customized for your data.

All data analyzed by Amazon Personalize is kept private and secure, and only used for your customized recommendations. You can start serving your personalized predictions via a simple API call from inside the virtual private cloud that the service maintains. You pay only for what you use, and there are no minimum fees and no upfront commitments.

Amazon Personalize is like having your own Amazon.com machine learning personalization team at your disposal, 24 hours a day.

Amazon Deep Learning AMIs

The AWS Deep Learning AMIs provide machine learning practitioners and researchers with the infrastructure and tools to accelerate deep learning in the cloud, at any scale. You can quickly launch Amazon EC2 instances pre-installed with popular deep learning frameworks such as Apache MXNet and Gluon, TensorFlow, Microsoft Cognitive Toolkit, Caffe, Caffe2, Theano, Torch, PyTorch, Chainer, and Keras to train sophisticated, custom AI models, experiment with new algorithms, or to learn new skills and techniques

AWS DeepLens

AWS DeepLens helps put deep learning in the hands of developers, literally, with a fully programmable video camera, tutorials, code, and pre-trained models designed to expand deep learning skills.

AWS DeepRacer

AWS DeepRacer is a 1/18th scale race car which gives you an interesting and fun way to get started with reinforcement learning (RL). RL is an advanced machine learning (ML) technique which 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.

With AWS DeepRacer, you now have a way to get hands-on with RL, experiment, and learn through autonomous driving. You can get started with the virtual car and tracks in the cloud-based 3D racing simulator, and for a real-

world experience, you can deploy your trained models onto AWS DeepRacer and race your friends, or take part in the global AWS DeepRacer League. Developers, the race is on.

Apache MXNet on AWS

Apache MXNet on AWS is a fast and scalable training and inference framework with an easy-to-use, concise API for machine learning.

MXNet includes the Gluon interface that allows developers of all skill levels to get started with deep learning on the cloud, on edge devices, and on mobile apps. In just a few lines of Gluon code, you can build linear regression, convolutional networks and recurrent LSTMs for object detection, speech recognition, recommendation, and personalization.

You can get started with MxNet on AWS with a fully-managed experience using Amazon SageMaker, a platform to build, train, and deploy machine learning models at scale. Or, you can use the AWS Deep Learning AMIs to build custom environments and workflows with MxNet as well as other frameworks including TensorFlow, PyTorch, Chainer, Keras, Caffe, Caffe2, and Microsoft Cognitive Toolkit.

TensorFlow on AWS

TensorFlow™ enables developers to quickly and easily get started with deep learning in the cloud. The framework has broad support in the industry and has become a popular choice for deep learning research and application development, particularly in areas such as computer vision, natural language understanding and speech translation.

You can get started on AWS with a fully-managed TensorFlow experience with Amazon SageMaker, a platform to build, train, and deploy machine learning models at scale. Or, you can use the AWS Deep Learning AMIs to build custom environments and workflows with TensorFlow and other popular frameworks including Apache MXNet, PyTorch, Caffe, Caffe2, Chainer, Gluon, Keras, and Microsoft Cognitive Toolkit.

AWS Inferentia

AWS Inferentia is a machine learning inference chip designed to deliver high performance at low cost. AWS Inferentia will support the TensorFlow, Apache

MXNet, and PyTorch deep learning frameworks, as well as models that use the ONNX format.

Making predictions using a trained machine learning model–a process called inference–can drive as much as 90% of the compute costs of the application. Using Amazon Elastic Inference, developers can reduce inference costs by up to 75% by attaching GPU-powered inference acceleration to Amazon EC2 and Amazon SageMaker instances. However, some inference workloads require an entire GPU or have extremely low latency requirements. Solving this challenge at low cost requires a dedicated inference chip.

AWS Inferentia provides high throughput, low latency inference performance at an extremely low cost. Each chip provides hundreds of TOPS (tera operations per second) of inference throughput to allow complex models to make fast predictions. For even more performance, multiple AWS Inferentia chips can be used together to drive thousands of TOPS of throughput. AWS Inferentia will be available for use with Amazon SageMaker, Amazon EC2, and Amazon Elastic Inference.

Management and Governance

Amazon CloudWatch

Amazon CloudWatch is a monitoring and management service built for developers, system operators, site reliability engineers (SRE), and IT managers. CloudWatch provides you with data and actionable insights to monitor your applications, understand and respond to system-wide performance changes, optimize resource utilization, and get a unified view of operational health. CloudWatch collects monitoring and operational data in the form of logs, metrics, and events, providing you with a unified view of AWS resources, applications and services that run on AWS, and on-premises servers. You can use CloudWatch to set high resolution alarms, visualize logs and metrics side by side, take automated actions, troubleshoot issues, and discover insights to optimize your applications, and ensure they are running smoothly.

AWS Auto Scaling

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.

AWS Control Tower

AWS Control Tower automates the set-up of a baseline environment, or landing zone, that is a secure, well-architected multi-account AWS environment. The configuration of the landing zone is based on best practices that have been established by working with thousands of enterprise customers to create a secure environment that makes it easier to govern AWS workloads with rules for security, operations, and compliance.

As enterprises migrate to AWS, they typically have a large number of applications and distributed teams. They often want to create multiple accounts to allow their teams to work independently, while still maintaining a consistent level of security and compliance. In addition, they use AWS's management and security services, like AWS Organizations, AWS Service Catalog and AWS Config, that provide very granular controls over their workloads. They want to maintain this control, but they also want a way to centrally govern and enforce the best use of AWS services across all the accounts in their environment.

Control Tower automates the set-up of their landing zone and configures AWS management and security services based on established best practices in a secure, compliant, multi-account environment. Distributed teams are able to provision new AWS accounts quickly, while central teams have the peace of mind knowing that new accounts are aligned with centrally established, company-wide compliance policies. This gives you control over your environment, without sacrificing the speed and agility AWS provides your development teams.

AWS Systems Manager

AWS Systems Manager gives you visibility and control of your infrastructure on AWS. Systems Manager provides a unified user interface so you can view operational data from multiple AWS services and allows you to automate operational tasks across your AWS resources. With Systems Manager, you can group resources, like Amazon EC2 instances, Amazon S3 buckets, or Amazon RDS instances, by application, view operational data for monitoring and troubleshooting, and take action on your groups of resources. Systems Manager simplifies resource and application management, shortens the time to detect and resolve operational problems, and makes it easy to operate and manage your infrastructure securely at scale.

AWS Systems Manager contains the following tools:

- **Resource groups:** Lets you create a logical group of resources associated with a particular workload such as different layers of an application stack, or production versus development environments. For example, you can group different layers of an application, such as the frontend web layer and the backend data layer. Resource groups can be created, updated, or removed programmatically through the API.
- **Insights Dashboard:** Displays operational data that the AWS Systems Manager automatically aggregates for each resource group. Systems Manager eliminates the need for you to navigate across multiple AWS consoles to view your operational data. With Systems Manager you can view API call logs from AWS CloudTrail, resource configuration changes from AWS Config, software inventory, and patch compliance status by resource group. You can also easily integrate your AWS CloudWatch Dashboards, AWS Trusted Advisor notifications, and AWS Personal Health Dashboard performance and availability alerts into your Systems Manager dashboard. Systems Manager centralizes all relevant operational data, so you can have a clear view of your infrastructure compliance and performance.
- **Run Command:** Provides a simple way of automating common administrative tasks like remotely executing shell scripts or PowerShell commands, installing software updates, or making changes to the configuration of OS, software, EC2 and instances and servers in your on-premises data center.

- **State Manager:** Helps you define and maintain consistent OS configurations such as firewall settings and anti-malware definitions to comply with your policies. You can monitor the configuration of a large set of instances, specify a configuration policy for the instances, and automatically apply updates or configuration changes.
- **Inventory:** Helps you collect and query configuration and inventory information about your instances and the software installed on them. You can gather details about your instances such as installed applications, DHCP settings, agent detail, and custom items. You can run queries to track and audit your system configurations.
- **Maintenance Window:** Lets you define a recurring window of time to run administrative and maintenance tasks across your instances. This ensures that installing patches and updates, or making other configuration changes does not disrupt business-critical operations. This helps improve your application availability.
- **Patch Manager:** Helps you select and deploy operating system and software patches automatically across large groups of instances. You can define a maintenance window so that patches are applied only during set times that fit your needs. These capabilities help ensure that your software is always up to date and meets your compliance policies.
- **Automation:** Simplifies common maintenance and deployment tasks, such as updating Amazon Machine Images (AMIs). Use the Automation feature to apply patches, update drivers and agents, or bake applications into your AMI using a streamlined, repeatable, and auditable process.
- **Parameter Store:** Provides an encrypted location to store important administrative information such as passwords and database strings. The Parameter Store integrates with AWS KMS to make it easy to encrypt the information you keep in the Parameter Store.

- **Distributor:** Helps you securely distribute and install software packages, such as software agents. Systems Manager Distributor allows you to centrally store and systematically distribute software packages while you maintain control over versioning. You can use Distributor to create and distribute software packages and then install them using Systems Manager Run Command and State Manager. Distributor can also use Identity and Access Management (IAM) policies to control who can create or update packages in your account. You can use the existing IAM policy support for Systems Manager Run Command and State Manager to define who can install packages on your hosts.
- **Session Manager:** Provides a browser-based interactive shell and CLI for managing Windows and Linux EC2 instances, without the need to open inbound ports, manage SSH keys, or use bastion hosts. Administrators can grant and revoke access to instances through a central location by using AWS Identity and Access Management (IAM) policies. This allows you to control which users can access each instance, including the option to provide non-root access to specified users. Once access is provided, you can audit which user accessed an instance and log each command to Amazon S3 or Amazon CloudWatch Logs using AWS CloudTrail.

AWS CloudFormation

AWS CloudFormation gives developers and systems administrators an easy way to create and manage a collection of related AWS resources, provisioning and updating them in an orderly and predictable fashion.

You can use the AWS CloudFormation sample templates or create your own templates to describe your AWS resources, and any associated dependencies or runtime parameters, required to run your application. You don't need to figure out the order for provisioning AWS services or the subtleties of making those dependencies work. CloudFormation takes care of this for you. After the AWS resources are deployed, you can modify and update them in a controlled and predictable way, in effect applying version control to your AWS infrastructure the same way you do with your software. You can also visualize your templates as diagrams and edit them using a drag-and-drop interface with the AWS CloudFormation Designer.

AWS CloudTrail

AWS CloudTrail is a web service that records AWS API calls for your account and delivers log files to you. The recorded information includes 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 the AWS service.

With CloudTrail, you can get a history of AWS API calls for your account, including API calls made using the AWS Management Console, AWS SDKs, command line tools, and higher-level AWS services (such as AWS CloudFormation). The AWS API call history produced by CloudTrail enables security analysis, resource change tracking, and compliance auditing.

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. The Config Rules feature enables you to create rules that automatically check the configuration of AWS resources recorded by AWS Config.

With AWS Config, you can discover existing and deleted AWS resources, determine your overall compliance against rules, and dive into configuration details of a resource at any point in time. These capabilities enable compliance auditing, security analysis, resource change tracking, and troubleshooting.

AWS OpsWorks

AWS OpsWorks is a configuration management service that provides managed instances of Chef and Puppet. Chef and Puppet are automation platforms that allow you to use code to automate the configurations of your servers. OpsWorks lets you use Chef and Puppet to automate how servers are configured, deployed, and managed across your Amazon EC2 instances or on-premises compute environments. OpsWorks has three offerings, AWS Opsworks for Chef Automate, AWS OpsWorks for Puppet Enterprise, and AWS OpsWorks Stacks.

AWS Service Catalog

AWS Service Catalog allows organizations to create and manage catalogs of IT services that are approved for use on AWS. These IT services can include

everything from virtual machine images, servers, software, and databases to complete multi-tier application architectures. AWS Service Catalog allows you to centrally manage commonly deployed IT services and helps you achieve consistent governance and meet your compliance requirements, while enabling users to quickly deploy only the approved IT services they need.

AWS Trusted Advisor

AWS Trusted Advisor is an online resource to help you reduce cost, increase performance, and improve security by optimizing your AWS environment. Trusted Advisor provides real-time guidance to help you provision your resources following AWS best practices.

AWS Personal Health Dashboard

AWS Personal Health Dashboard provides alerts and remediation guidance when AWS is experiencing events that might affect you. While the Service Health Dashboard displays the general status of AWS services, Personal Health Dashboard gives you a personalized view into the performance and availability of the AWS services underlying your AWS resources. The dashboard displays relevant and timely information to help you manage events in progress, and provides proactive notification to help you plan for scheduled activities. With Personal Health Dashboard, alerts are automatically triggered by changes in the health of AWS resources, giving you event visibility and guidance to help quickly diagnose and resolve issues.

AWS Managed Services

AWS Managed Services provides ongoing management of your AWS infrastructure so you can focus on your applications. By implementing best practices to maintain your infrastructure, AWS Managed Services helps to reduce your operational overhead and risk. AWS Managed Services automates common activities such as change requests, monitoring, patch management, security, and backup services, and provides full-lifecycle services to provision, run, and support your infrastructure. Our rigor and controls help to enforce your corporate and security infrastructure policies, and enables you to develop solutions and applications using your preferred development approach. AWS Managed Services improves agility, reduces cost, and unburdens you from infrastructure operations so you can direct resources toward differentiating your business.

AWS Console Mobile Application

The AWS Console Mobile Application lets customers view and manage a select set of resources to support incident response while on-the-go.

The Console Mobile Application allows AWS customers to monitor resources through a dedicated dashboard and view configuration details, metrics, and alarms for select AWS services. The Dashboard provides permitted users with a single view a resource's status, with real-time data on Amazon CloudWatch, Personal Health Dashboard, and AWS Billing and Cost Management. Customers can view ongoing issues and follow through to the relevant CloudWatch alarm screen for a detailed view with graphs and configuration options. In addition, customers can check on the status of specific AWS services, view detailed resource screens, and perform select actions.

AWS License Manager

AWS License Manager makes it easier to manage licenses in AWS and on-premises servers from software vendors such as Microsoft, SAP, Oracle, and IBM. AWS License Manager lets administrators create customized licensing rules that emulate the terms of their licensing agreements, and then enforces these rules when an instance of EC2 gets launched. Administrators can use these rules to limit licensing violations, such as using more licenses than an agreement stipulates or reassigning licenses to different servers on a short-term basis. The rules in AWS License Manager enable you to limit a licensing breach by physically stopping the instance from launching or by notifying administrators about the infringement. Administrators gain control and visibility of all their licenses with the AWS License Manager dashboard and reduce the risk of non-compliance, misreporting, and additional costs due to licensing overages.

AWS License Manager integrates with AWS services to simplify the management of licenses across multiple AWS accounts, IT catalogs, and on-premises, through a single AWS account. License administrators can add rules in AWS Service Catalog, which allows them to create and manage catalogs of IT services that are approved for use on all their AWS accounts. Through seamless integration with AWS Systems Manager and AWS Organizations, administrators can manage licenses across all the AWS accounts in an organization and on-premises environments. AWS Marketplace buyers can also use AWS License Manager to track bring your own license (BYOL) software obtained from the Marketplace and keep a consolidated view of all their licenses.

AWS Well-Architected Tool

The AWS Well-Architected Tool helps you review the state of your workloads and compares them to the latest AWS architectural best practices. The tool is based on the AWS Well-Architected Framework, developed to help cloud architects build secure, high-performing, resilient, and efficient application infrastructure. This Framework provides a consistent approach for customers and partners to evaluate architectures, has been used in tens of thousands of workload reviews conducted by the AWS solutions architecture team, and provides guidance to help implement designs that scale with application needs over time.

To use this free tool, available in the AWS Management Console, just define your workload and answer a set of questions regarding operational excellence, security, reliability, performance efficiency, and cost optimization. The AWS Well-Architected Tool then provides a plan on how to architect for the cloud using established best practices.

Media Services

Amazon Elastic Transcoder

Amazon Elastic Transcoder is media transcoding in the cloud. It is designed to be a highly scalable, easy- to-use, and cost-effective way for developers and businesses to convert (or transcode) media files from their source format into versions that will play back on devices like smartphones, tablets, and PCs.

AWS Elemental MediaConnect

AWS Elemental MediaConnect is a high-quality transport service for live video. Today, broadcasters and content owners rely on satellite networks or fiber connections to send their high-value content into the cloud or to transmit it to partners for distribution. Both satellite and fiber approaches are expensive, require long lead times to set up, and lack the flexibility to adapt to changing requirements. To be more nimble, some customers have tried to use solutions that transmit live video on top of IP infrastructure, but have struggled with reliability and security.

Now you can get the reliability and security of satellite and fiber combined with the flexibility, agility, and economics of IP-based networks using AWS Elemental MediaConnect. MediaConnect enables you to build mission-critical

live video workflows in a fraction of the time and cost of satellite or fiber services. You can use MediaConnect to ingest live video from a remote event site (like a stadium), share video with a partner (like a cable TV distributor), or replicate a video stream for processing (like an over-the-top service). MediaConnect combines reliable video transport, highly secure stream sharing, and real-time network traffic and video monitoring that allow you to focus on your content, not your transport infrastructure.

AWS Elemental MediaConvert

AWS Elemental MediaConvert is a file-based video transcoding service with broadcast-grade features. It allows you to easily create video-on-demand (VOD) content for broadcast and multiscreen delivery at scale. The service combines advanced video and audio capabilities with a simple web services interface and pay-as-you-go pricing. With AWS Elemental MediaConvert, you can focus on delivering compelling media experiences without having to worry about the complexity of building and operating your own video processing infrastructure.

AWS Elemental MediaLive

AWS Elemental MediaLive is a broadcast-grade live video processing 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 works by encoding your live video streams in real-time, taking a larger-sized live video source and compressing it into smaller versions for distribution to your viewers. With AWS Elemental MediaLive, you can easily set up streams for both live events and 24x7 channels with advanced broadcasting features, high availability, and pay-as-you-go pricing. AWS Elemental MediaLive lets you focus on creating compelling live video experiences for your viewers without the complexity of building and operating broadcast-grade video processing infrastructure.

AWS Elemental Media Package

AWS Elemental MediaPackage reliably prepares and protects your video for delivery over the Internet. From a single video input, AWS Elemental MediaPackage creates video streams formatted to play on connected TVs, mobile phones, computers, tablets, and game consoles. It makes it easy to implement popular video features for viewers (start-over, pause, rewind, etc.),

like those commonly found on DVRs. AWS Elemental MediaPackage can also protect your content using Digital Rights Management (DRM). AWS Elemental MediaPackage scales automatically in response to load, so your viewers will always get a great experience without you having to accurately predict in advance the capacity you'll need.

AWS Elemental MediaStore

AWS Elemental MediaStore is an AWS storage service optimized for media. It gives you the performance, consistency, and low latency required to deliver live streaming video content. AWS Elemental MediaStore acts as the origin store in your video workflow. Its high performance capabilities meet the needs of the most demanding media delivery workloads, combined with long-term, cost-effective storage.

AWS Elemental MediaTailor

AWS Elemental MediaTailor lets video providers insert individually targeted advertising into their video streams without sacrificing broadcast-level quality-of-service. With AWS Elemental MediaTailor, viewers of your live or on-demand video each receive a stream that combines your content with ads personalized to them. But unlike other personalized ad solutions, with AWS Elemental MediaTailor your entire stream – video and ads – is delivered with broadcast-grade video quality to improve the experience for your viewers. AWS Elemental MediaTailor delivers automated reporting based on both client and server-side ad delivery metrics, making it easy to accurately measure ad impressions and viewer behavior. You can easily monetize unexpected high-demand viewing events with no up-front costs using AWS Elemental MediaTailor. It also improves ad delivery rates, helping you make more money from every video, and it works with a wider variety of content delivery networks, ad decision servers, and client devices.

See also Amazon Kinesis Video Streams.

Migration and Transfer

AWS Migration Hub

AWS Migration Hub provides a single location to track the progress of application migrations across multiple AWS and partner solutions. Using

Migration Hub allows you to choose the AWS and partner migration tools that best fit your needs, while providing visibility into the status of migrations across your portfolio of applications. Migration Hub also provides key metrics and progress for individual applications, regardless of which tools are being used to migrate them. For example, you might use AWS Database Migration Service, AWS Server Migration Service, and partner migration tools such as ATADATA ATAmotion, CloudEndure Live Migration, or RiverMeadow Server Migration Saas to migrate an application comprised of a database, virtualized web servers, and a bare metal server. Using Migration Hub, you can view the migration progress of all the resources in the application. This allows you to quickly get progress updates across all of your migrations, easily identify and troubleshoot any issues, and reduce the overall time and effort spent on your migration projects.

AWS Application Discovery Service

AWS Application Discovery Service helps enterprise customers plan migration projects by gathering information about their on-premises data centers.

Planning data center migrations can involve thousands of workloads that are often deeply interdependent. Server utilization data and dependency mapping are important early first steps in the migration process. AWS Application Discovery Service collects and presents configuration, usage, and behavior data from your servers to help you better understand your workloads.

The collected data is retained in encrypted format in an AWS Application Discovery Service data store. You can export this data as a CSV file and use it to estimate the Total Cost of Ownership (TCO) of running on AWS and to plan your migration to AWS. In addition, this data is also available in AWS Migration Hub, where you can migrate the discovered servers and track their progress as they get migrated to AWS.

AWS Database Migration Service

AWS Database Migration Service helps you migrate databases to AWS easily and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to and from most widely used commercial and open-source databases. The service supports homogeneous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle to Amazon

Aurora or Microsoft SQL Server to MySQL. It also allows you to stream data to Amazon Redshift from any of the supported sources including Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle, SAP ASE, and SQL Server, enabling consolidation and easy analysis of data in the petabyte-scale data warehouse. AWS Database Migration Service can also be used for continuous data replication with high availability.

AWS Server Migration Service

AWS Server Migration Service (SMS) is an agentless service which makes it easier and faster for you to migrate thousands of on-premises workloads to AWS. AWS SMS allows you to automate, schedule, and track incremental replications of live server volumes, making it easier for you to coordinate large-scale server migrations.

AWS Snowball

AWS Snowball is a petabyte-scale data transport solution that uses secure appliances to transfer large amounts of data into and out of AWS. The use of Snowball addresses common challenges with large- scale data transfers including high network costs, long transfer times, and security concerns. Transferring data with Snowball is simple, fast, secure, and can be as little as one-fifth the cost of high-speed Internet.

With Snowball, you don't need to write any code or purchase any hardware to transfer your data. Simply create a job in the AWS Management Console and a Snowball appliance will be automatically shipped to you. Once it arrives, attach the appliance to your local network, download and run the Snowball client to establish a connection, and then use the client to select the file directories that you want to transfer to the appliance. The client will then encrypt and transfer the files to the appliance at high speed. Once the transfer is complete and the appliance is ready to be returned, the E Ink shipping label will automatically update and you can track the job status using the Amazon Simple Notification Service (Amazon SNS), text messages, or directly in the console.

Snowball uses multiple layers of security designed to protect your data including tamper-resistant enclosures, 256-bit encryption, and an industry-standard Trusted Platform Module (TPM) designed to ensure both security and full chain of custody of your data. Once the data transfer job has been processed and verified, AWS performs a software erasure of the Snowball appliance.

AWS Snowball Edge

AWS Snowball Edge is a data migration and edge computing device that comes in two options. Snowball Edge Storage Optimized provides 100 TB of capacity and 24 vCPUs and is well suited for local storage and large scale data transfer. Snowball Edge Compute Optimized provides 52 vCPUs and an optional GPU for use cases such as advanced machine learning and full motion video analysis in disconnected environments. Customers can use these two options for data collection, machine learning and processing, and storage in environments with intermittent connectivity (such as manufacturing, industrial, and transportation) or in extremely remote locations (such as military or maritime operations) before shipping it back to AWS. These devices may also be rack mounted and clustered together to build larger, temporary installations.

Snowball Edge supports specific Amazon EC2 instance types as well as AWS Lambda functions, so customers may develop and test in AWS then deploy applications on devices in remote locations to collect, pre-process, and return the data. Common use cases include data migration, data transport, image collation, IoT sensor stream capture, and machine learning.

AWS Snowmobile

AWS Snowmobile is an exabyte-scale data transfer service used to move extremely large amounts of data to AWS. You can transfer up to 100 PB per Snowmobile, a 45-foot long ruggedized shipping container, pulled by a semi-trailer truck. Snowmobile makes it easy to move massive volumes of data to the cloud, including video libraries, image repositories, or even a complete data center migration. Transferring data with Snowmobile is secure, fast, and cost effective.

After an initial assessment, a Snowmobile will be transported to your data center, and AWS personnel will configure it for you so it can be accessed as a network storage target. When your Snowmobile is on site, AWS personnel will work with your team to connect a removable, high-speed network switch from the Snowmobile to your local network. Then you can begin your high-speed data transfer from any number of sources within your data center to the Snowmobile. After your data is loaded, the Snowmobile is driven back to AWS where your data is imported into Amazon S3 or Amazon Glacier.

AWS Snowmobile uses multiple layers of security designed to protect your data including dedicated security personnel, GPS tracking, alarm monitoring, 24/7

video surveillance, and an optional escort security vehicle while in transit. All data is encrypted with 256-bit encryption keys managed through AWS KMS and designed to ensure both security and full chain of custody of your data.

AWS DataSync

AWS DataSync is a data transfer service that makes it easy for you to automate moving data between on-premises storage and Amazon S3 or Amazon Elastic File System (Amazon EFS). DataSync automatically handles many of the tasks related to data transfers that can slow down migrations or burden your IT operations, including running your own instances, handling encryption, managing scripts, network optimization, and data integrity validation. You can use DataSync to transfer data at speeds up to 10 times faster than open-source tools. DataSync uses an on-premises software agent to connect to your existing storage or file systems using the Network File System (NFS) protocol, so you don't have write scripts or modify your applications to work with AWS APIs. You can use DataSync to copy data over AWS Direct Connect or internet links to AWS. The service enables one-time data migrations, recurring data processing workflows, and automated replication for data protection and recovery. Getting started with DataSync is easy: Deploy the DataSync agent on premises, connect it to a file system or storage array, select Amazon EFS or S3 as your AWS storage, and start moving data. You pay only for the data you copy.

AWS Transfer for SFTP

AWS Transfer for SFTP is a fully managed service that enables the transfer of files directly into and out of Amazon S3 using the Secure File Transfer Protocol (SFTP)—also known as Secure Shell (SSH) File Transfer Protocol. AWS helps you seamlessly migrate your file transfer workflows to AWS Transfer for SFTP—by integrating with existing authentication systems, and providing DNS routing with Amazon Route 53—so nothing changes for your customers and partners, or their applications. With your data in S3, you can use it with AWS services for processing, analytics, machine learning, and archiving. Getting started with AWS Transfer for SFTP (AWS SFTP) is easy; there is no infrastructure to buy and setup.

Mobile

AWS Amplify

AWS Amplify makes it easy to create, configure, and implement scalable mobile applications powered by AWS. Amplify seamlessly provisions and manages your mobile backend and provides a simple framework to easily integrate your backend with your iOS, Android, Web, and React Native frontends. Amplify also automates the application release process of both your frontend and backend allowing you to deliver features faster.

Mobile applications require cloud services for actions that can't be done directly on the device, such as offline data synchronization, storage, or data sharing across multiple users. You often have to configure, set up, and manage multiple services to power the backend. You also have to integrate each of those services into your application by writing multiple lines of code. However, as the number of application features grow, your code and release process becomes more complex and managing the backend requires more time.

Amplify provisions and manages backends for your mobile applications. You just select the capabilities you need such as authentication, analytics, or offline data sync and Amplify will automatically provision and manage the AWS service that powers each of the capabilities. You can then integrate those capabilities into your application through the Amplify libraries and UI components.

Amazon Cognito

Amazon Cognito lets you add user sign-up, sign-in, and access control to your web and mobile apps quickly and easily. With Amazon Cognito, you also have the option to authenticate users through social identity providers such as Facebook, Twitter, or Amazon, with SAML identity solutions, or by using your own identity system. In addition, Amazon Cognito enables you to save data locally on users' devices, allowing your applications to work even when the devices are offline. You can then synchronize data across users' devices so that their app experience remains consistent regardless of the device they use.

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 devices.

Amazon Pinpoint

Amazon Pinpoint makes it easy to send targeted messages to your customers through multiple engagement channels. Examples of targeted campaigns are promotional alerts and customer retention campaigns, and transactional messages are messages such as order confirmations and password reset messages.

You can integrate Amazon Pinpoint into your mobile and web apps to capture usage data to provide you with insight into how customers interact with your apps. Amazon Pinpoint also tracks the ways that your customers respond to the messages you send—for example, by showing you the number of messages that were delivered, opened, or clicked.

You can develop custom audience segments and send them pre-scheduled targeted campaigns via email, SMS, and push notifications. Targeted campaigns are useful for sending promotional or educational content to re-engage and retain your users.

You can send transactional messages using the console or the Amazon Pinpoint REST API. Transactional campaigns can be sent via email, SMS, push notifications, and voice messages. You can also use the API to build custom applications that deliver campaign and transactional messages.

AWS Device Farm

AWS Device Farm is an app testing service that lets you test and interact with your Android, iOS, and web apps on many devices at once, or reproduce issues on a device in real time. View video, screenshots, logs, and performance data to pinpoint and fix issues before shipping your app.

AWS AppSync

AWS AppSync is a serverless back-end for mobile, web, and enterprise applications.

AWS AppSync makes it easy to build data driven mobile and web applications by handling securely all the application data management tasks like online and offline data access, data synchronization, and data manipulation across multiple data sources. AWS AppSync uses GraphQL, an API query language designed to build client applications by providing an intuitive and flexible syntax for describing their data requirement.

Networking and Content Delivery

Amazon VPC

Amazon Virtual Private Cloud (Amazon VPC) lets you provision a logically isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways. You can use both IPv4 and IPv6 in your VPC for secure and easy access to resources and applications.

You can easily customize the network configuration for your VPC. For example, you can create a public- facing subnet for your web servers that has access to the Internet, and place your backend systems, such as databases or application servers, in a private-facing subnet with no Internet access. You can leverage multiple layers of security (including security groups and network access control lists) to help control access to EC2 instances in each subnet.

Additionally, you can create a hardware virtual private network (VPN) connection between your corporate data center and your VPC and leverage the AWS Cloud as an extension of your corporate data center.

Amazon CloudFront

Amazon CloudFront is a fast content delivery network (CDN) service that securely delivers data, videos, applications, and APIs to customers globally with low latency, high transfer speeds, all within a developer-friendly environment. CloudFront is integrated with AWS – both physical locations that are directly connected to the AWS global infrastructure, as well as other AWS services. CloudFront works seamlessly with services including AWS Shield for DDoS mitigation, Amazon S3, Elastic Load Balancing or Amazon EC2 as origins for your applications, and Lambda@Edge to run custom code closer to customers' users and to customize the user experience.

You can get started with the Content Delivery Network in minutes, using the same AWS tools that you're already familiar with: APIs, AWS Management Console, AWS CloudFormation, CLIs, and SDKs. Amazon's CDN offers a simple, pay-as-you-go pricing model with no upfront fees or required long-term

contracts, and support for the CDN is included in your existing AWS Support subscription.

Amazon Route 53

Amazon Route 53 is a highly available and scalable cloud Domain Name System (DNS) web service. It is designed to give developers and businesses an extremely reliable and cost-effective way to route end users to Internet applications by translating human readable names, such as www.example.com, into the numeric IP addresses, such as 192.0.2.1, that computers use to connect to each other. Amazon Route 53 is fully compliant with IPv6 as well.

Amazon Route 53 effectively connects user requests to infrastructure running in AWS—such as EC2 instances, Elastic Load Balancing load balancers, or Amazon S3 buckets—and can also be used to route users to infrastructure outside of AWS. You can use Amazon Route 53 to configure DNS health checks to route traffic to healthy endpoints or to independently monitor the health of your application and its endpoints. Amazon Route 53 traffic flow makes it easy for you to manage traffic globally through a variety of routing types, including latency-based routing, Geo DNS, and weighted round robin—all of which can be combined with DNS Failover in order to enable a variety of low-latency, fault-tolerant architectures. Using Amazon Route 53 traffic flow's simple visual editor, you can easily manage how your end users are routed to your application's endpoints—whether in a single AWS Region or distributed around the globe. Amazon Route 53 also offers Domain Name Registration—you can purchase and manage domain names such as example.com and Amazon Route 53 will automatically configure DNS settings for your domains.

AWS PrivateLink

AWS PrivateLink simplifies the security of data shared with cloud-based applications by eliminating the exposure of data to the public Internet. AWS PrivateLink provides private connectivity between VPCs, AWS services, and on-premises applications, securely on the Amazon network. AWS PrivateLink makes it easy to connect services across different accounts and VPCs to significantly simplify the network architecture.

AWS Direct Connect

AWS Direct Connect makes it easy to establish a dedicated network connection from your premises to AWS. Using AWS Direct Connect, you can establish

private connectivity between AWS and your data center, office, or co-location environment, which in many cases can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

AWS Direct Connect lets you establish a dedicated network connection between your network and one of the AWS Direct Connect locations. Using industry standard 802.1Q virtual LANS (VLANs), this dedicated connection can be partitioned into multiple virtual interfaces. This allows you to use the same connection to access public resources, such as objects stored in Amazon S3 using public IP address space, and private resources such as EC2 instances running within a VPC using private IP address space, while maintaining network separation between the public and private environments. Virtual interfaces can be reconfigured at any time to meet your changing needs.

AWS Global Accelerator

AWS Global Accelerator is a networking service that improves the availability and performance of the applications that you offer to your global users.

Today, if you deliver applications to your global users over the public internet, your users might face inconsistent availability and performance as they traverse through multiple public networks to reach your application. These public networks are often congested and each hop can introduce availability and performance risk. AWS Global Accelerator uses the highly available and congestion-free AWS global network to direct internet traffic from your users to your applications on AWS, making your users' experience more consistent.

To improve the availability of your application, you must monitor the health of your application endpoints and route traffic only to healthy endpoints. AWS Global Accelerator improves application availability by continuously monitoring the health of your application endpoints and routing traffic to the closest healthy endpoints.

AWS Global Accelerator also makes it easier to manage your global applications by providing static IP addresses that act as a fixed entry point to your application hosted on AWS which eliminates the complexity of managing specific IP addresses for different AWS Regions and Availability Zones. AWS Global Accelerator is easy to set up, configure and manage.

Amazon API Gateway

Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, and secure APIs at any scale. With a few clicks in the AWS Management Console, you can create an API that acts as a "front door" for applications to access data, business logic, or functionality from your back-end services, such as workloads running on Amazon EC2, code running on AWS Lambda, or any web application. Amazon API Gateway handles all the tasks involved in accepting and processing up to hundreds of thousands of concurrent API calls, including traffic management, authorization and access control, monitoring, and API version management.

AWS Transit Gateway

AWS Transit Gateway is a service that enables customers to connect their Amazon Virtual Private Clouds (VPCs) and their on-premises networks to a single gateway. As you grow the number of workloads running on AWS, you need to be able to scale your networks across multiple accounts and Amazon VPCs to keep up with the growth. Today, you can connect pairs of Amazon VPCs using peering. However, managing point-to-point connectivity across many Amazon VPCs, without the ability to centrally manage the connectivity policies, can be operationally costly and cumbersome. For on-premises connectivity, you need to attach your AWS VPN to each individual Amazon VPC. This solution can be time consuming to build and hard to manage when the number of VPCs grows into the hundreds.

With AWS Transit Gateway, you only have to create and manage a single connection from the central gateway in to each Amazon VPC, on-premises data center, or remote office across your network. Transit Gateway acts as a hub that controls how traffic is routed among all the connected networks which act like spokes. This hub and spoke model significantly simplifies management and reduces operational costs because each network only has to connect to the Transit Gateway and not to every other network. Any new VPC is simply connected to the Transit Gateway and is then automatically available to every other network that is connected to the Transit Gateway. This ease of connectivity makes it easy to scale your network as you grow.

AWS App Mesh

AWS App Mesh makes it easy to monitor and control microservices running on AWS. App Mesh standardizes how your microservices communicate, giving you end-to-end visibility and helping to ensure high-availability for your applications.

Modern applications are often composed of multiple microservices that each perform a specific function. This architecture helps to increase the availability and scalability of the application by allowing each component to scale independently based on demand, and automatically degrading functionality when a component fails instead of going offline. Each microservice interacts with all the other microservices through an API. As the number of microservices grows within an application, it becomes increasingly difficult to pinpoint the exact location of errors, re-route traffic after failures, and safely deploy code changes. Previously, this has required you to build monitoring and control logic directly into your code and redeploy your microservices every time there are changes.

AWS App Mesh makes it easy to run microservices by providing consistent visibility and network traffic controls for every microservice in an application. App Mesh removes the need to update application code to change how monitoring data is collected or traffic is routed between microservices. App Mesh configures each microservice to export monitoring data and implements consistent communications control logic across your application. This makes it easy to quickly pinpoint the exact location of errors and automatically re-route network traffic when there are failures or when code changes need to be deployed.

You can use App Mesh with Amazon ECS and Amazon EKS to better run containerized microservices at scale. App Mesh uses the open source Envoy proxy, making it compatible with a wide range of AWS partner and open source tools for monitoring microservices.

AWS Cloud Map

AWS Cloud Map is a cloud resource discovery service. With Cloud Map, you can define custom names for your application resources, and it maintains the updated location of these dynamically changing resources. This increases your application availability because your web service always discovers the most up-to-date locations of its resources.

Modern applications are typically composed of multiple services that are accessible over an API and perform a specific function. Each service interacts with a variety of other resources such as databases, queues, object stores, and customer-defined microservices, and they also need to be able to find the location of all the infrastructure resources on which it depends, in order to function. You typically manually manage all these resource names and their locations within the application code. However, manual resource management becomes time consuming and error-prone as the number of dependent infrastructure resources increases or the number of microservices dynamically scale up and down based on traffic. You can also use third-party service discovery products, but this requires installing and managing additional software and infrastructure.

Cloud Map allows you to register any application resources such as databases, queues, microservices, and other cloud resources with custom names. Cloud Map then constantly checks the health of resources to make sure the location is up-to-date. The application can then query the registry for the location of the resources needed based on the application version and deployment environment.

Elastic Load Balancing

Elastic Load Balancing (ELB) automatically distributes incoming application traffic across multiple targets, such as Amazon EC2 instances, containers, and IP addresses. It can handle the varying load of your application traffic in a single Availability Zone or across multiple Availability Zones. Elastic Load Balancing offers three types of load balancers that all feature the high availability, automatic scaling, and robust security necessary to make your applications fault tolerant.

- Application Load Balancer is best suited for load balancing of HTTP and HTTPS traffic and provides advanced request routing targeted at the delivery of modern application architectures, including microservices and containers. Operating at the individual request level (Layer 7), Application Load Balancer routes traffic to targets within Amazon Virtual Private Cloud (Amazon VPC) based on the content of the request.

- Network Load Balancer is best suited for load balancing of TCP traffic where extreme performance is required. Operating at the connection level (Layer 4), Network Load Balancer routes traffic to targets within Amazon Virtual Private Cloud (Amazon VPC) and is capable of handling millions of requests per second while maintaining ultra-low latencies. Network Load Balancer is also optimized to handle sudden and volatile traffic patterns.
- Classic Load Balancer provides basic load balancing across multiple Amazon EC2 instances and operates at both the request level and connection level. Classic Load Balancer is intended for applications that were built within the EC2-Classic network.

Robotics

AWS RoboMaker

AWS RoboMaker is a service that makes it easy to develop, test, and deploy intelligent robotics applications at scale. RoboMaker extends the most widely used open-source robotics software framework, Robot Operating System (ROS), with connectivity to cloud services. This includes AWS machine learning services, monitoring services, and analytics services that enable a robot to stream data, navigate, communicate, comprehend, and learn. RoboMaker provides a robotics development environment for application development, a robotics simulation service to accelerate application testing, and a robotics fleet management service for remote application deployment, update, and management.

Robots are machines that sense, compute, and take action. Robots need instructions to accomplish tasks, and these instructions come in the form of applications that developers code to determine how the robot will behave. Receiving and processing sensor data, controlling actuators for movement, and performing a specific task are all functions that are typically automated by these intelligent robotics applications. Intelligent robots are being increasingly used in warehouses to distribute inventory, in homes to carry out tedious housework, and in retail stores to provide customer service. Robotics applications use machine learning in order to perform more complex tasks like recognizing an object or face, having a conversation with a person, following a spoken command, or navigating autonomously. Until now, developing, testing, and deploying intelligent robotics applications was difficult and time consuming.

Building intelligent robotics functionality using machine learning is complex and requires specialized skills. Setting up a development environment can take each developer days and building a realistic simulation system to test an application can take months due to the underlying infrastructure needed. Once an application has been developed and tested, a developer needs to build a deployment system to deploy the application into the robot and later update the application while the robot is in use.

AWS RoboMaker provides the tools to make building intelligent robotics applications more accessible, a fully managed simulation service for quick and easy testing, and a deployment service for lifecycle management. AWS RoboMaker removes the heavy lifting from each step of robotics development so you can focus on creating innovative robotics applications.

Satellite

AWS Ground Station

AWS Ground Station is a fully managed service that lets you control satellite communications, downlink and process satellite data, and scale your satellite operations quickly, easily and cost-effectively 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 are at the core of global satellite networks, which are facilities that provide communications between the ground and the satellites by using antennas to receive data and control systems to send radio signals to command and control the satellite. Today, you must either build your own ground stations and antennas, or obtain long-term leases with ground station providers, often in multiple countries to provide enough opportunities to contact the satellites as they orbit the globe. Once all this data is downloaded, you need servers, storage, and networking in close proximity to the antennas to process, store, and transport the data from the satellites.

AWS Ground Station eliminates these problems by delivering a global Ground Station as a Service. We provide direct access to AWS services and the AWS Global Infrastructure including our low-latency global fiber network right where your data is downloaded into our AWS Ground Station. This enables you to easily control satellite communications, quickly ingest and process your satellite data, and rapidly integrate that data with your applications and other services

running in the AWS Cloud. For example, you can use Amazon S3 to store the downloaded data, Amazon Kinesis Data Streams for managing data ingestion from satellites, Amazon SageMaker for building custom machine learning applications that apply to your data sets, and Amazon EC2 to command and download data from satellites. AWS Ground Station can help you save up to 80% on the cost of your ground station operations by allowing you to pay only for the actual antenna time used, and relying on our global footprint of ground stations to download data when and where you need it, instead of building and operating your own global ground station infrastructure. There are no long-term commitments, and you gain the ability to rapidly scale your satellite communications on-demand when your business needs it.

Security, Identity, and Compliance

AWS Security Hub

AWS Security Hub gives you a comprehensive view of your high-priority security alerts and compliance status across AWS accounts. There are a range of powerful security tools at your disposal, from firewalls and endpoint protection to vulnerability and compliance scanners. But oftentimes this leaves your team switching back-and-forth between these tools to deal with hundreds, and sometimes thousands, of security alerts every day. With Security Hub, you now have a single place that aggregates, organizes, and prioritizes your security alerts, or findings, from multiple AWS services, such as Amazon GuardDuty, Amazon Inspector, and Amazon Macie, as well as from AWS Partner solutions. Your findings are visually summarized on integrated dashboards with actionable graphs and tables. You can also continuously monitor your environment using automated compliance checks based on the AWS best practices and industry standards your organization follows. Get started with AWS Security Hub just a few clicks in the Management Console and once enabled, Security Hub will begin aggregating and prioritizing findings.

Amazon Cloud Directory

Amazon Cloud Directory enables you to build flexible, cloud-native directories for organizing hierarchies of data along multiple dimensions. With Cloud Directory, you can create directories for a variety of use cases, such as organizational charts, course catalogs, and device registries. While traditional directory solutions, such as Active Directory Lightweight Directory Services (AD LDS) and other LDAP-based directories, limit you to a single hierarchy, Cloud

Directory offers you the flexibility to create directories with hierarchies that span multiple dimensions. For example, you can create an organizational chart that can be navigated through separate hierarchies for reporting structure, location, and cost center.

Amazon Cloud Directory automatically scales to hundreds of millions of objects and provides an extensible schema that can be shared with multiple applications. As a fully-managed service, Cloud Directory eliminates time-consuming and expensive administrative tasks, such as scaling infrastructure and managing servers. You simply define the schema, create a directory, and then populate your directory by making calls to the Cloud Directory API.

AWS Identity and Access Management

AWS Identity and Access Management (IAM) enables you to securely control access to AWS services and resources for your users. Using IAM, you can create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources. IAM allows you to do the following:

- Manage IAM users and their access: You can create users in IAM, assign them individual security credentials (access keys, passwords, and multi-factor authentication devices), or request temporary security credentials to provide users access to AWS services and resources. You can manage permissions in order to control which operations a user can perform.
- Manage IAM roles and their permissions: You can create roles in IAM and manage permissions to control which operations can be performed by the entity, or AWS service, that assumes the role. You can also define which entity is allowed to assume the role.
- Manage federated users and their permissions: You can enable identity federation to allow existing identities (users, groups, and roles) in your enterprise to access the AWS Management Console, call AWS APIs, and access resources, without the need to create an IAM user for each identity.

Amazon GuardDuty

Amazon GuardDuty is a threat detection service that continuously monitors for malicious or unauthorized behavior to help you protect your AWS accounts and workloads. It monitors for activity such as unusual API calls or potentially

unauthorized deployments that indicate a possible account compromise. GuardDuty also detects potentially compromised instances or reconnaissance by attackers.

Enabled with a few clicks in the AWS Management Console, Amazon GuardDuty can immediately begin analyzing billions of events across your AWS accounts for signs of risk. GuardDuty identifies suspected attackers through integrated threat intelligence feeds and uses machine learning to detect anomalies in account and workload activity. When a potential threat is detected, the service delivers a detailed security alert to the GuardDuty console and AWS CloudWatch Events. This makes alerts actionable and easy to integrate into existing event management and workflow systems.

Amazon GuardDuty is cost effective and easy. It does not require you to deploy and maintain software or security infrastructure, meaning it can be enabled quickly with no risk of negatively impacting existing application workloads. There are no upfront costs with GuardDuty, no software to deploy, and no threat intelligence feeds required. Customers pay for the events analyzed by GuardDuty and there is a 30-day free trial available for every new account to the service.

Amazon Inspector

Amazon Inspector is an automated security assessment service that helps improve the security and compliance of applications deployed on AWS. Amazon Inspector automatically assesses applications for exposure, vulnerabilities, and deviations from best practices. After performing an assessment, Amazon Inspector produces a detailed list of security findings prioritized by level of severity. These findings can be reviewed directly or as part of detailed assessment reports which are available via the Amazon Inspector console or API.

Amazon Inspector security assessments help you check for unintended network accessibility of your Amazon EC2 instances and for vulnerabilities on those EC2 instances. Amazon Inspector assessments are offered to you as pre-defined rules packages mapped to common security best practices and vulnerability definitions. Examples of built-in rules include checking for access to your EC2 instances from the internet, remote root login being enabled, or vulnerable software versions installed. These rules are regularly updated by AWS security researchers.

Amazon Macie

Amazon Macie is a security service that uses machine learning to automatically discover, classify, and protect sensitive data in AWS. Amazon Macie recognizes sensitive data such as personally identifiable information (PII) or intellectual property, and provides you with dashboards and alerts that give visibility into how this data is being accessed or moved. The fully managed service continuously monitors data access activity for anomalies, and generates detailed alerts when it detects risk of unauthorized access or inadvertent data leaks. Today, Amazon Macie is available to protect data stored in Amazon S3, with support for additional AWS data stores coming later this year.

AWS Artifact

AWS Artifact is your go-to, central resource for compliance-related information that matters to you. It provides on-demand access to AWS' security and compliance reports and select online agreements. Reports available in AWS Artifact include our Service Organization Control (SOC) reports, Payment Card Industry (PCI) reports, and certifications from accreditation bodies across geographies and compliance verticals that validate the implementation and operating effectiveness of AWS security controls. Agreements available in AWS Artifact include the Business Associate Addendum (BAA) and the Nondisclosure Agreement (NDA).

AWS Certificate Manager

AWS Certificate Manager is a service that lets you easily provision, manage, and deploy Secure Sockets Layer/Transport Layer Security (SSL/TLS) certificates for use with AWS services and your internal connected resources. SSL/TLS certificates are used to secure network communications and establish the identity of websites over the Internet as well as resources on private networks. AWS Certificate Manager removes the time-consuming manual process of purchasing, uploading, and renewing SSL/TLS certificates.

With AWS Certificate Manager, you can quickly request a certificate, deploy it on ACM-integrated AWS resources, such as Elastic Load Balancers, Amazon CloudFront distributions, and APIs on API Gateway, and let AWS Certificate Manager handle certificate renewals. It also enables you to create private certificates for your internal resources and manage the certificate lifecycle centrally. Public and private certificates provisioned through AWS Certificate Manager for use with ACM-integrated services are free. You pay only for the

AWS resources you create to run your application. With AWS Certificate Manager Private Certificate Authority, you pay monthly for the operation of the private CA and for the private certificates you issue.

AWS CloudHSM

AWS CloudHSM is a cloud-based hardware security module (HSM) that enables you to easily generate and use your own encryption keys on the AWS Cloud. With CloudHSM, you can manage your own encryption keys using FIPS 140-2 Level 3 validated HSMs. CloudHSM offers you the flexibility to integrate with your applications using industry-standard APIs, such as PKCS#11, Java Cryptography Extensions (JCE), and Microsoft CryptoNG (CNG) libraries.

CloudHSM is standards-compliant and enables you to export all of your keys to most other commercially-available HSMs, subject to your configurations. It is a fully-managed service that automates time-consuming administrative tasks for you, such as hardware provisioning, software patching, high-availability, and backups. CloudHSM also enables you to scale quickly by adding and removing HSM capacity on-demand, with no up-front costs.

AWS Directory Service

AWS Directory Service for Microsoft Active Directory, also known as AWS Managed Microsoft AD, enables your directory-aware workloads and AWS resources to use managed Active Directory in the AWS Cloud. AWS Managed Microsoft AD is built on actual Microsoft Active Directory and does not require you to synchronize or replicate data from your existing Active Directory to the cloud. You can use standard Active Directory administration tools and take advantage of built-in Active Directory features such as Group Policy and single sign-on (SSO). With AWS Managed Microsoft AD, you can easily join Amazon EC2 and Amazon RDS for SQL Server instances to a domain, and use AWS Enterprise IT applications such as Amazon WorkSpaces with Active Directory users and groups.

AWS Firewall Manager

AWS Firewall Manager is a security management service that makes it easier to centrally configure and manage AWS WAF rules across your accounts and applications. Using Firewall Manager, you can easily roll out AWS WAF rules for your Application Load Balancers and Amazon CloudFront distributions across accounts in AWS Organizations. As new applications are created,

Firewall Manager also makes it easy to bring new applications and resources into compliance with a common set of security rules from day one. Now you have a single service to build firewall rules, create security policies, and enforce them in a consistent, hierarchical manner across your entire Application Load Balancers and Amazon CloudFront infrastructure.

AWS Key Management Service

AWS Key Management Service (KMS) makes it easy for you to create and manage keys and control the use of encryption across a wide range of AWS services and in your applications. AWS KMS is a secure and resilient service that uses FIPS 140-2 validated hardware security modules to protect your keys. AWS KMS is integrated with AWS CloudTrail to provide you with logs of all key usage to help meet your regulatory and compliance needs.

AWS Organizations

AWS Organizations offers policy-based management for multiple AWS accounts. With Organizations, you can create groups of accounts, automate account creation, apply and manage policies for those groups. Organizations enables you to centrally manage policies across multiple accounts, without requiring custom scripts and manual processes.

Using AWS Organizations, you can create Service Control Policies (SCPs) that centrally control AWS service use across multiple AWS accounts. You can also use Organizations to help automate the creation of new accounts through APIs. Organizations helps simplify the billing for multiple accounts by enabling you to setup a single payment method for all the accounts in your organization through consolidated billing. AWS Organizations is available to all AWS customers at no additional charge.

AWS Secrets Manager

AWS Secrets Manager helps you protect secrets needed to access your applications, services, and IT resources. The service enables you to easily rotate, manage, and retrieve database credentials, API keys, and other secrets throughout their lifecycle. Users and applications retrieve secrets with a call to Secrets Manager APIs, eliminating the need to hardcode sensitive information in plain text. Secrets Manager offers secret rotation with built-in integration for Amazon RDS for MySQL, PostgreSQL, and Amazon Aurora. Also, the service is extensible to other types of secrets, including API keys and OAuth tokens. In

addition, Secrets Manager enables you to control access to secrets using fine-grained permissions and audit secret rotation centrally for resources in the AWS Cloud, third-party services, and on-premises.

AWS Shield

AWS Shield is a managed Distributed Denial of Service (DDoS) protection service that safeguards web applications running on AWS. AWS Shield provides always-on detection and automatic inline mitigations that minimize application downtime and latency, so there is no need to engage AWS Support to benefit from DDoS protection. There are two tiers of AWS Shield: Standard and Advanced.

All AWS customers benefit from the automatic protections of AWS Shield Standard, at no additional charge. AWS Shield Standard defends against most common, frequently occurring network and transport layer DDoS attacks that target your website or applications. When you use AWS Shield Standard with Amazon CloudFront and Amazon Route 53, you receive comprehensive availability protection against all known infrastructure (Layer 3 and 4) attacks.

For higher levels of protection against attacks targeting your applications running on Amazon Elastic Compute Cloud (EC2), Elastic Load Balancing (ELB), Amazon CloudFront, and Amazon Route 53 resources, you can subscribe to AWS Shield Advanced. In addition to the network and transport layer protections that come with Standard, AWS Shield Advanced provides additional detection and mitigation against large and sophisticated DDoS attacks, near real-time visibility into attacks, and integration with AWS WAF, a web application firewall. AWS Shield Advanced also gives you 24x7 access to the AWS DDoS Response Team (DRT) and protection against DDoS related spikes in your Amazon Elastic Compute Cloud (EC2), Elastic Load Balancing (ELB), Amazon CloudFront, and Amazon Route 53 charges.

AWS Shield Advanced is available globally on all Amazon CloudFront and Amazon Route 53 edge locations. You can protect your web applications hosted anywhere in the world by deploying Amazon CloudFront in front of your application. Your origin servers can be Amazon S3, Amazon Elastic Compute Cloud (EC2), Elastic Load Balancing (ELB), or a custom server outside of AWS. You can also enable AWS Shield Advanced directly on an Elastic IP or Elastic Load Balancing (ELB) in the following AWS Regions - Northern Virginia, Oregon, Ireland, Tokyo, and Northern California.

AWS Single Sign-On (SSO)

AWS Single Sign-On (SSO) is a cloud SSO service that makes it easy to centrally manage SSO access to multiple AWS accounts and business applications. With just a few clicks, you can enable a highly available SSO service without the upfront investment and on-going maintenance costs of operating your own SSO infrastructure. With AWS SSO, you can easily manage SSO access and user permissions to all of your accounts in AWS Organizations centrally. AWS SSO also includes built-in SAML integrations to many business applications, such as Salesforce, Box, and Office 365. Further, by using the AWS SSO application configuration wizard, you can create Security Assertion Markup Language (SAML) 2.0 integrations and extend SSO access to any of your SAML-enabled applications. Your users simply sign in to a user portal with credentials they configure in AWS SSO or using their existing corporate credentials to access all their assigned accounts and applications from one place.

AWS WAF

AWS WAF is a web application firewall that helps protect your web applications from common web exploits that could affect application availability, compromise security, or consume excessive resources. AWS WAF gives you control over which traffic to allow or block to your web application by defining customizable web security rules. You can use AWS WAF to create custom rules that block common attack patterns, such as SQL injection or cross-site scripting, and rules that are designed for your specific application. New rules can be deployed within minutes, letting you respond quickly to changing traffic patterns. Also, AWS WAF includes a full-featured API that you can use to automate the creation, deployment, and maintenance of web security rules.

Storage

Amazon S3

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. This means customers of all sizes and industries can use it to store and protect any amount of data for a range of use cases, such as websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides easy-to-use management features

so you can organize your data and configure finely-tuned access controls to meet your specific business, organizational, and compliance requirements. Amazon S3 is designed for 99.999999999% (11 9's) of durability, and stores data for millions of applications for companies all around the world.

Amazon Elastic Block Store

Amazon Elastic Block Store (Amazon EBS) provides persistent block storage volumes for use with Amazon EC2 instances in the AWS Cloud. Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability. Amazon EBS volumes offer the consistent and low-latency performance needed to run your workloads. With Amazon EBS, you can scale your usage up or down within minutes—all while paying a low price for only what you provision.

Amazon Elastic File System

Amazon Elastic File System (Amazon EFS) provides a simple, scalable, elastic file system for Linux-based workloads 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, so your applications have the storage they need – when they need it. It is designed to provide massively parallel shared access to thousands of Amazon EC2 instances, enabling your applications to achieve high levels of aggregate throughput and IOPS with consistent low latencies. Amazon EFS is a fully managed service that requires no changes to your existing applications and tools, providing access through a standard file system interface for seamless integration. Amazon EFS is a regional service storing data within and across multiple Availability Zones (AZs) for high availability and durability. You can access your file systems across AZs and regions and share files between thousands of Amazon EC2 instances and on-premises servers via AWS Direct Connect or AWS VPN.

Amazon EFS is well suited to support a broad spectrum of use cases from highly parallelized, scale-out workloads that require the highest possible throughput to single-threaded, latency-sensitive workloads. Use cases such as lift-and-shift enterprise applications, big data analytics, web serving and content management, application development and testing, media and entertainment workflows, database backups, and container storage.

Amazon FSx for Lustre

Amazon FSx for Lustre is a fully managed file system that is optimized for compute-intensive workloads, such as high performance computing, machine learning, and media data processing workflows. Many of these applications require the high-performance and low latencies of scale-out, parallel file systems. Operating these file systems typically requires specialized expertise and administrative overhead, requiring you to provision storage servers and tune complex performance parameters. With Amazon FSx, you can launch and run a Lustre file system that can process massive data sets at up to hundreds of gigabytes per second of throughput, millions of IOPS, and sub-millisecond latencies.

Amazon FSx for Lustre is seamlessly integrated with Amazon S3, making it easy to link your long-term data sets with your high performance file systems to run compute-intensive workloads. You can automatically copy data from S3 to FSx for Lustre, run your workloads, and then write results back to S3. FSx for Lustre also enables you to burst your compute-intensive workloads from on-premises to AWS by allowing you to access your FSx file system over Amazon Direct Connect or VPN. FSx for Lustre helps you cost-optimize your storage for compute-intensive workloads: It provides cheap and performant non-replicated storage for processing data, with your long-term data stored durably in Amazon S3 or other low-cost data stores. With Amazon FSx, you pay for only the resources you use. There are no minimum commitments, upfront hardware or software costs, or additional fees.

Amazon FSx for Windows File Server

Amazon FSx for Windows File Server provides a fully managed native Microsoft Windows file system so you can easily move your Windows-based applications that require file storage to AWS. Built on Windows Server, Amazon FSx provides shared file storage with the compatibility and features that your Windows-based applications rely on, including full support for the SMB protocol and Windows NTFS, Active Directory (AD) integration, and Distributed File System (DFS). Amazon FSx uses SSD storage to provide the fast performance your Windows applications and users expect, with high levels of throughput and IOPS, and consistent sub-millisecond latencies. This compatibility and performance is particularly important when moving workloads that require Windows shared file storage, like CRM, ERP, and .NET applications, as well as home directories.

With Amazon FSx, you can launch highly durable and available Windows file systems that can be accessed from up to thousands of compute instances using the industry-standard SMB protocol. Amazon FSx eliminates the typical administrative overhead of managing Windows file servers. You pay for only the resources used, with no upfront costs, minimum commitments, or additional fees.

Amazon S3 Glacier

Amazon S3 Glacier is a secure, durable, and extremely low-cost storage service for data archiving and long-term backup. It is designed to deliver 99.999999999% durability, and provides 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 archive data at rest. You can store data for as little as $0.004 per gigabyte per month, a significant savings compared to on-premises solutions. To keep costs low yet suitable for varying retrieval needs, Amazon S3 Glacier provides three options for access to archives, from a few minutes to several hours.

AWS Storage Gateway

AWS Storage Gateway is a hybrid storage service that enables your on-premises applications to seamlessly use AWS cloud storage. You can use the service for backup and archiving, disaster recovery, cloud data processing, storage tiering, and migration. Your applications connect to the service through a virtual machine or hardware gateway appliance using standard storage protocols, such as NFS, SMB and iSCSI. The gateway connects to AWS storage services, such as Amazon S3, Amazon Glacier, and Amazon EBS, providing storage for files, volumes, and virtual tapes in AWS. The service includes a highly-optimized data transfer mechanism, with bandwidth management, automated network resilience, and efficient data transfer, along with a local cache for low-latency on-premises access to your most active data.

Next Steps

Reinvent how you work with IT by signing up for the AWS Free Tier, which enables you to gain hands-on experience with a broad selection of AWS products and services. Within the AWS Free Tier, you can test workloads and

run applications to learn more and build the right solution for your organization. You can also contact AWS Sales and Business Development.

By signing up for AWS, you have access to Amazon's cloud computing services. Note: The sign-up process requires a credit card, which will not be charged until you start using services. There are no long-term commitments and you can stop using AWS at any time.

To help familiarize you with AWS, view these short videos that cover topics like creating an account, launching a virtual server, storing media and more. Learn about the breadth and depth of AWS on our general AWS Channel and AWS Online Tech Talks. Get hands on experience from our self-paced labs.

Conclusion

AWS provides building blocks that you can assemble quickly to support virtually any workload. With AWS, you'll find a complete set of highly available services that are designed to work together to build sophisticated scalable applications.

You have access to highly durable storage, low-cost compute, high-performance databases, management tools, and more. All this is available without up-front cost, and you pay for only what you use. These services help organizations move faster, lower IT costs, and scale. AWS is trusted by the largest enterprises and the hottest start-ups to power a wide variety of workloads, including web and mobile applications, game development, data processing and warehousing, storage, archive, and many others.

Contributors

The following individuals and organizations contributed to this document:

- Sajee Mathew, Principal Solutions Architect, Amazon Web Services

Further Reading

For additional information, see the following:

- AWS Architecture Center[1]

Contents

Introduction 1

Key principles 1

- Understand the fundamentals of pricing 1
- Start early with cost optimization 2
- Maximize the power of flexibility 2
- Use the right pricing model for the job 2

Get started with the AWS Free Tier 3

- 12 Months Free 3
- Always Free 4
- Trials 4

AWS Pricing/TCO Tools 4

- AWS Pricing Calculator 5
- Migration Evaluator 5

Pricing details for individual services 6

- Amazon Elastic Compute Cloud (Amazon EC2) 6
- AWS Lambda 10
- Amazon Elastic Block Store (Amazon EBS) 11
- Amazon Simple Storage Service (Amazon S3) 12
- Amazon S3 Glacier 13
- AWS Outposts 14
- AWS Snow Family 16
- Amazon RDS 18
- Amazon DynamoDB 19
- Amazon CloudFront 23
- Amazon Kendra 23
- Amazon Kinesis 25

AWS IoT Events 27
AWS Cost Optimization 28
Choose the right pricing models 28
Match Capacity with Demand 28
Implement processes to identify resource waste 29
AWS Support Plan Pricing 30
Cost calculation examples 30
AWS Cloud cost calculation example 30
Hybrid cloud cost calculation example 33
Conclusion 37
Contributors 38
Further Reading 38
Document Revisions 39

Introduction

AWS has the services to help you build sophisticated applications with increased flexibility, scalability and reliability. Whether you're looking for compute power, database storage, content delivery, or other functionality, with AWS you pay only for the individual services you need, for as long as you use them, without complex licensing. AWS offers you a variety of pricing models for over 160 cloud services. You only pay for the services you consume, and once you stop using them, there are no additional costs or termination fees. This whitepaper provides an overview of how AWS pricing works across some of the most widely used services. The latest pricing information for each AWS service is available at http://aws.amazon.com/pricing/.

Key principles

Although pricing models vary across services, it's worthwhile to review key principles and best practices that are broadly applicable.

Understand the fundamentals of pricing

There are three fundamental drivers of cost with AWS: compute, storage, and outbound data transfer. These characteristics vary somewhat, depending on the AWS product and pricing model you choose.

In most cases, there is no charge for inbound data transfer or for data transfer between other AWS services within the same Region. There are some exceptions, so be sure to verify data transfer rates before beginning. Outbound data transfer is aggregated across services and then charged at the outbound data transfer rate. This charge appears on the monthly statement as *AWS Data Transfer Out*. The more data you transfer, the less you pay per GB. For compute resources, you pay hourly from the time you launch a resource until the time you terminate it, unless you have made a reservation for which the cost is agreed upon beforehand. For data storage and transfer, you typically pay per GB.

Except as otherwise noted, AWS prices are exclusive of applicable taxes and duties, including VAT and sales tax. For customers with a Japanese billing address, use of AWS is subject to Japanese Consumption Tax. For more information, see Amazon Web Services Consumption Tax FAQ.

Start early with cost optimization

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 started 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. 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. To learn how you can optimize and save costs today, visit AWS Cost Optimization.

Maximize the power of flexibility

AWS services are priced independently, transparently, and available on-demand, so you can choose and pay for exactly what you need. You may also choose to save money through a reservation model. By paying for services on an as-needed basis, you can redirect your focus to innovation and invention, reducing procurement complexity and enabling your business to be fully elastic.

One of the key advantages of cloud-based resources is that you don't pay for them when they're not running. By turning off instances you don't use, you can reduce costs by 70 percent or more compared to using them 24/7. This enables you to be cost efficient and, at the same time, have all the power you need when workloads are active.

Use the right pricing model for the job

AWS offers several pricing models depending on product. These include:

- **On-Demand Instances** let you pay for compute or database capacity by the hour or second (minimum of 60 seconds) depending on which instances you run with no long-term commitments or upfront payments.
- **Savings Plans** are a flexible pricing model that offer low prices on Amazon EC2, AWS Lambda and AWS Fargate usage, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a one- or three-year term.

- **Spot Instances** are an Amazon EC2 pricing mechanism that let you request spare computing capacity with no upfront commitment and at discounted hourly rate (up to 90% off the on-demand price).
- **Reservations** provide you with the ability to receive a greater discount, up to 75 percent, by paying for capacity ahead of time. For more details, see the Optimizing costs with reservations section.

Get started with the AWS Free Tier

The AWS Free Tier enables you to gain free, hands-on experience with more than 60 products on AWS platform. AWS Free Tier includes the following free offer types:

- **12 Months Free** – These tier offers include 12 months free usage following your initial sign-up date to AWS. When your 12 month free usage term expires, or if your application use exceeds the tiers, you simply pay standard, pay-as-you-go service rates.
- **Always Free** – These free tier offers do not expire and are available to all AWS customers.
- **Trials** – These offers are short term free trials starting from date you activate a particular service. Once the trial period expires, you simply pay standard, pay-as-you-go service rates.

This section lists some of the most commonly used AWS Free Tier services. Terms and conditions apply. For the full list of AWS Free Tier services, see AWS Free Tier.

12 Months Free

- Amazon Elastic Compute Cloud (Amazon EC2): 750 hours per month of Linux, RHEL, or SLES t2.micro/t3.micro instance usage or 750 hours per month of Windows t2.micro/t3.micro instance usage dependent on Region.
- Amazon Simple Storage Service (Amazon S3): 5 GB of Amazon S3 standard storage, 20,000 Get Requests, and 2,000 Put Requests.
- Amazon Relational Database Service (Amazon RDS): 750 hours of Amazon RDS Single-AZ db.t2.micro database usage for running MySQL, PostgreSQL, MariaDB, Oracle BYOL, or SQL Server (running SQL Server Express Edition); 20 GB of general purpose SSD database storage and 20 GB of storage for database backup and DB snapshots.

- Amazon CloudFront: 50 GB Data Transfer Out and 2,000,000 HTTP and HTTPS Requests each month.

Always Free

- Amazon DynamoDB: Up to 200 million requests per month (25 Write Capacity units and 25 Read Capacity units); 25 GB of storage.
- Amazon S3 Glacier: Retrieve up to 10 GB of your Amazon S3 Glacier data per month for free (applies to standard retrievals using the Glacier API only).
- AWS Lambda: 1 million free requests per month; up to 3.2 million seconds of compute time per month.

Trials

- Amazon SageMaker: 250 hours per month of t2.medium notebook,50 hours per month of m4.xlarge for training, 125 hours per month of m4.xlarge for hosting for the first two months.
- Amazon Redshift: 750 hours per month for free, enough hours to continuously run one DC2.Large node with 160GB of compressed SSD storage. You can also build clusters with multiple nodes to test larger data sets, which will consume your free hours more quickly. Once your two month free trial expires or your usage exceeds 750 hours per month, you can shut down your cluster to avoid any charges, or keep it running at the standard On-Demand Rate.

The AWS Free Tier is not available in the AWS GovCloud (US) Regions or the China (Beijing) Region at this time. The Lambda Free Tier is available in the AWS GovCloud (US) Region.

AWS Pricing/TCO Tools

To get the most out of your estimates, you should have a good idea of your basic requirements. For example, if you're going to try Amazon Elastic Compute Cloud (Amazon EC2), it might help if you know what kind of operating system you need, what your memory requirements are, and how much I/O you need. You should also decide whether you need storage, such as if you're going to run a database and how long you intend to use the servers. You don't need to make these decisions before generating an estimate, though. You can play around with the service configuration and parameters to

see which options fit your use case and budget best. For more information about AWS service pricing, see AWS Services Pricing.

AWS offers couple of tools (free of cost) for you to use. If the workload details and services to be used are identified, AWS pricing calculator can help with calculating the total cost of ownership. Migration Evaluator helps with inventorying your existing environment, identifying workload information, and designing and planning your AWS migration.

AWS Pricing Calculator

AWS Pricing Calculator is a web based service that you can use to create cost estimates to suit your AWS use cases. AWS Pricing Calculator is useful both for people who have never used AWS and for those who want to reorganize or expand their usage.

AWS Pricing Calculator allows you to explore AWS services based on your use cases and create a cost estimate. You can model your solutions before building them, explore the price points and calculations behind your estimate, and find the available instance types and contract terms that meet your needs. This enables you to make informed decisions about using AWS. You can plan your AWS costs and usage or price out setting up a new set of instances and services.

AWS Pricing Calculator is free for use. It provides an estimate of your AWS fees and charges. The estimate doesn't include any taxes that might apply to the fees and charges. AWS Pricing Calculator provides pricing details for your information only. AWS Pricing Calculator provides a console interface at https://calculator.aws/#/.

Migration Evaluator

Migration Evaluator (Formerly TSO Logic) is a complimentary service to create data-driven business cases for AWS Cloud planning and migration.

Creating business cases on your own can be a time-consuming process and does not always identify the most cost-effective deployment and purchasing options. Migration Evaluator quickly provides a business case to make sound AWS planning and migration decisions. With Migration Evaluator, your organization can build a data-driven business case for AWS, gets access to AWS expertise, visibility into the costs associated with multiple migration strategies, and insights on how reusing existing software licensing reduces costs further.

A business case is the first step in the AWS migration journey. Beginning with on-premises inventory discovery, you can choose to upload exports from 3rd party tools or install a complimentary agentless collector to monitor Windows, Linux and SQL Server footprints. As part of a white-gloved experience, Migration Evaluator includes a team of program managers and solution architects to capture your migration objective and use analytics to narrow down the subset of migration patterns best suited to your business needs. The results are captured in a transparent business case which aligns business and technology stakeholders to provide a prescriptive next step in your migration journey.

Migration Evaluator service analyzes an enterprise's compute footprint, including server configuration, utilization, annual costs to operate, eligibility for bring-your-own-license, and hundreds of other parameters. It then statistically models utilization patterns, matching each workload with optimized placements in the AWS Amazon Elastic Cloud Compute and Amazon Elastic Block Store. Finally, it outputs a business case with a comparison of the current-state against multiple future-state configurations showing the flexibility of AWS.

For more information, see Migration Evaluator.

Pricing details for individual services

Different types of services lend themselves to different pricing models. For example, Amazon EC2 pricing varies by instance type, whereas the Amazon Aurora database service includes charges for data input/output (I/O) and storage. This section provides an overview of pricing concepts and examples for few AWS services. You can always find current price information for each AWS service at AWS Pricing.

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers. The simple web service interface of Amazon EC2 allows you to obtain and configure capacity with minimal friction with complete control of your computing resources.

Amazon EC2 reduces the time required to obtain and boot new server instances in minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change.

Pricing models for Amazon EC2

There are five ways to pay for Amazon EC2 instances: On-Demand Instances, Savings Plans, Reserved Instances, and Spot Instances.

On-Demand Instances

With On-Demand Instances, you pay for compute capacity per hour or per second, depending on which instances you run. No long-term commitments or upfront payments are required. You can increase or decrease your compute capacity to meet the demands of your application and only pay the specified hourly rates for the instance you use. On-Demand Instances are recommended for the following use cases:

- Users who prefer the low cost and flexibility of Amazon EC2 without upfront payment or long-term commitments
- Applications with short-term, spiky, or unpredictable workloads that cannot be interrupted
- Applications being developed or tested on Amazon EC2 for the first time

Savings Plans

Savings Plans are a flexible pricing model that offer low prices on Amazon EC2, AWS Lambda, and AWS Fargate usage, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a 1 or 3 year term. Savings Plans is a flexible pricing model that provides savings of up to 72% on your AWS compute usage. This pricing model offers lower prices on Amazon EC2 instances usage, regardless of instance family, size, OS, tenancy or AWS Region, and also applies to AWS Fargate and AWS Lambda usage.

For workloads that have predictable and consistent usage, Savings Plans can provide significant savings compared to On-Demand Instances. it is recommended for:

- Workloads with a consistent and steady-state usage
- Customers who want to use different instance types and compute solutions across different locations
- Customers who can make monetary commitment to use EC2 over a one-or three-year term

Spot Instances

Amazon EC2 Spot Instances allow you to request spare Amazon EC2 computing capacity for up to 90 percent off the On-Demand price. Spot Instances are recommended for:

- Applications that have flexible start and end times
- Applications that are only feasible at very low compute prices
- Users with fault-tolerant and/or stateless workloads

> Spot Instance prices are set by Amazon EC2 and adjust gradually based on long-term trends in supply and demand for Spot Instance capacity.

Reserved Instances

Amazon EC2 Reserved Instances provide you with a significant discount (up to 75 percent) compared to On-Demand Instance pricing. In addition, when Reserved Instances are assigned to a specific Availability Zone, they provide a capacity reservation, giving you additional confidence in your ability to launch instances when you need them.

Per-second billing

Per-second billing saves money and has a minimum of 60 seconds billing. It is particularly effective for resources that have periods of low and high usage such as development and testing, data processing, analytics, batch processing, and gaming applications. Learn more about per-second billing.

Estimating Amazon EC2 costs

When you begin to estimate the cost of using Amazon EC2, consider the following:

- **Clock hours of server time:** Resources incur charges when they are running—for example, from the time Amazon EC2 instances are launched until they are terminated, or from the time Elastic IP addresses are allocated until the time they are de-allocated.

- **Instance type:** Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instance types comprise varying combinations of CPU, memory, storage, and networking capacity and give you the flexibility to choose the appropriate mix of resources for your applications. Each instance type includes at least one instance size, allowing you to scale your resources to the requirements of your target workload.
- **Pricing model**: With On-Demand Instances, you pay for compute capacity by the hour with no required minimum commitments.
- **Number of instances**: You can provision multiple instances of your Amazon EC2 and Amazon EBS resources to handle peak loads.
- **Load balancing:** You can use Elastic Load Balancing to distribute traffic among Amazon EC2 Instances. The number of hours Elastic Load Balancing runs and the amount of data it processes contribute to the monthly cost.
- **Detailed monitoring**: You can use Amazon CloudWatch to monitor your EC2 instances. By default, basic monitoring is enabled. For a fixed monthly rate, you can opt for detailed monitoring, which includes seven preselected metrics recorded once a minute. Partial months are charged on an hourly pro rata basis, at a per instance-hour rate.
- **Amazon EC2 Auto Scaling**: Amazon EC2 Auto Scaling automatically adjusts the number of Amazon EC2 instances in your deployment according to the scaling policies you define. This service is available at no additional charge beyond Amazon CloudWatch fees.
- **Elastic IP addresses**: You can have one Elastic IP address associated with a running instance at no charge.
- **Licensing:** To run operating systems and applications on AWS, you can obtain variety of software licenses from AWS on a pay-as-you-go basis that are fully-compliant and do not require you to manage complex licensing terms and conditions. However, if you have existing licensing agreements with software vendors, you can bring your eligible licenses to the cloud to reduce total cost of ownership (TCO). AWS offers License Manager which makes it easier to manage your software licenses from vendors such as Microsoft, SAP, Oracle, and IBM across AWS and on-premises environments.

For more information, see Amazon EC2 pricing.

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.

AWS Lambda pricing

With AWS Lambda, you pay only for what you use. You are charged based on the number of requests for your functions and the time it takes for your code to execute. Lambda registers a request each time it starts executing in response to an event notification or invoke call, including test invokes from the console. You are charged for the total number of requests across all your functions.

Duration is calculated from the time your code begins executing until it returns or otherwise terminates, rounded up to the nearest 100 milliseconds. The price depends on the amount of memory you allocate to your function.

AWS Lambda participates in Compute Savings Plans, a flexible pricing model that offers low prices on Amazon EC2, AWS Fargate, and AWS Lambda usage, in exchange for a commitment to a consistent amount of usage (measured in $/hour) for a 1 or 3 year term. With Compute Savings Plans, you can save up to 17% on AWS Lambda. Savings apply to Duration, Provisioned Concurrency, and Duration (Provisioned Concurrency).

Request pricing

- Free Tier: 1 million requests per month, 400,000 GB-seconds of compute time per month
- $0.20 per 1 million requests thereafter, or $0.0000002 per request

Duration pricing

- 400,000 GB-seconds per month free, up to 3.2 million seconds of compute time
- $0.00001667 for every GB-second used thereafter

Additional charges

You may incur additional charges if your Lambda function uses other AWS services or transfers data. For example, if your Lambda function reads and writes data to or from Amazon S3, you will be billed for the read/write requests and the data stored in Amazon S3. Data transferred into and out of your AWS Lambda functions from outside the Region the function executed in will be charged at the EC2 data transfer rates as listed on Amazon EC2 On-Demand Pricing under *Data Transfer*.

Amazon Elastic Block Store (Amazon EBS)

Amazon Elastic Block Store (Amazon EBS) an easy to use, high performance block storage service designed for use with Amazon EC2 instances. Amazon EBS volumes are off-instance storage that persists independently from the life of an instance. They are analogous to virtual disks in the cloud. Amazon EBS provides two volume types:

- **SSD-backed volumes** are optimized for transactional workloads involving frequent read/write operations with small I/O size, where the dominant performance attribute is IOPS.
- **HDD-backed volumes** are optimized for large streaming workloads where throughput (measured in megabits per second) is a better performance measure than IOPS.

How Amazon EBS is priced

Amazon EBS pricing includes three factors:

- **Volumes**: Volume storage for all EBS volume types is charged by the amount of GB you provision per month, until you release the storage.
- **Snapshots**: Snapshot storage is based on the amount of space your data consumes in Amazon S3. Because Amazon EBS does not save empty blocks, it is likely that the snapshot size will be considerably less than your volume size. Copying EBS snapshots is charged based on the volume of data transferred across Regions. For the first snapshot of a volume, Amazon EBS saves a full copy of your data to Amazon S3. For each incremental snapshot, only the changed part of your Amazon EBS volume is saved. After the snapshot is copied, standard EBS snapshot charges apply for storage in the destination Region.

- **EBS Fast Snapshot Restore (FSR)**: This is charged in Date Services Unit-Hours (DSUs) for each Availability Zone in which it is enabled. DSUs are billed per minute with a 1 hour minimum. The price of 1 FSR DSU-hour is $0.75 per Availability Zone. (pricing based on us-east-1 (N.Virginia)).
- **EBS direct APIs for Snapshots**: EBS direct APIs for Snapshots provide access to directly read EBS snapshot data and identify differences between two snapshots. The following charges apply for these APIs.
    - ListChangedBlocks and ListSnapshotBlocks APIs are charged per request.
    - GetSnapshotBlock API is charged per SnapshotAPIUnit (block size 512 KiB)
- **Data transfer:** Consider the amount of data transferred out of your application. Inbound data transfer is free, and outbound data transfer charges are tiered. If you use external or cross-region data transfers, additional EC2 data transfer charges will apply.

For more information, see the Amazon EBS pricing page.

Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service (Amazon S3) is object storage built to store and retrieve any amount of data from anywhere: websites, mobile apps, corporate applications, and data from IoT sensors or devices. It is designed to deliver 99.999999999 percent durability, and stores data for millions of applications used by market leaders in every industry. As with other AWS services, Amazon S3 provides the simplicity and cost-effectiveness of pay-as-you-go pricing.

Estimating Amazon S3 storage costs

With Amazon S3, you pay only for the storage you use, with no minimum fee. Prices are based on the location of your Amazon S3 bucket. When you begin to estimate the cost of Amazon S3, consider the following:

- **Storage class:** Amazon S3 offers a range of storage classes designed for different use cases. These include S3 Standard for general-purpose storage of frequently accessed data; S3 Intelligent-Tiering for data with unknown or changing access patterns; S3 Standard-Infrequent Access (S3 Standard-IA) and S3 One Zone-Infrequent Access (S3 One Zone-IA) for long-lived, but less frequently accessed data; and Amazon S3 Glacier (S3 Glacier) and Amazon S3 Glacier Deep Archive (S3 Glacier Deep Archive) for long-term archive and digital preservation. Amazon S3 also offers capabilities to manage your data throughout its lifecycle. Once an S3 Lifecycle policy is set, your data will automatically transfer to a different storage class without any changes to your application.
- **Storage:** Costs vary with number and size of objects stored in your Amazon S3 buckets as well as type of storage.
- **Requests and Data retrievals:** Requests costs made against S3 buckets and objects are based on request type and quantity of requests.
- **Data transfer:** The amount of data transferred out of the Amazon S3 region. Transfers between S3 buckets or from Amazon S3 to any service(s) within the same AWS Region are free.
- **Management and replication:** You pay for the storage management features (Amazon S3 inventory, analytics, and object tagging) that are enabled on your account's buckets.

For more information, see Amazon S3 pricing. You can estimate your monthly bill using the AWS Pricing Calculator.

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.

Amazon S3 Glacier provides low-cost, long-term storage

Starting at $0.004 per GB per month, Amazon S3 Glacier allows you to archive large amounts of data at a very low cost. You pay only for what you need, with no minimum

commitments or upfront fees. Other factors determining pricing include requests and data transfers out of Amazon S3 Glacier (incoming transfers are free).

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. For details, see the Amazon S3 Glacier FAQs.

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.

Amazon S3 Glacier Select pricing

Amazon S3 Glacier Select allows queries to run directly on data stored in Amazon S3 Glacier without having to retrieve the entire archive. Pricing for this feature is based on the total amount of data scanned, the amount of data returned by Amazon S3 Glacier Select, and the number of Amazon S3 Glacier Select requests initiated.

For more information, see the Amazon S3 Glacier pricing page.

Data transfer

Data transfer in to Amazon S3 is free. Data transfer out of Amazon S3 is priced by Region. For more information on AWS Snowball pricing, see the AWS Snowball pricing page.

AWS Outposts

AWS Outposts is a fully managed service that extends AWS infrastructure, AWS services, APIs, and tools to any datacenter, co-location space, or on-premises facility. AWS Outposts is ideal for workloads that require low latency access to on-premises systems, local data processing, or local data storage.

Outposts are connected to the nearest AWS Region to provide the same management and control plane services on premises for a truly consistent operational experience across your on-premises and cloud environments. Your Outposts infrastructure and AWS services are managed, monitored, and updated by AWS just like in the cloud.

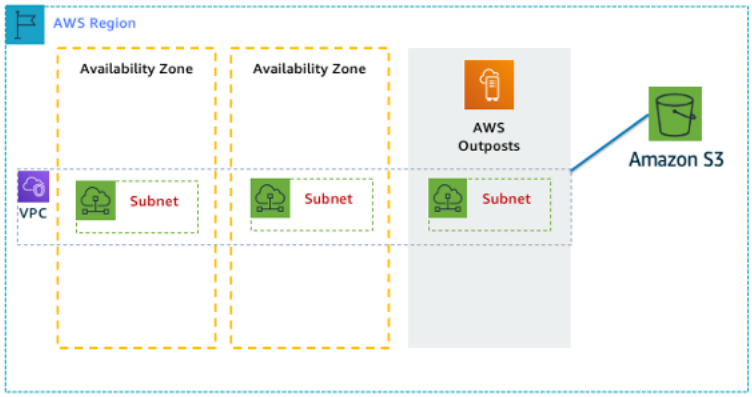

Figure 1: Example AWS Outposts architecture

Pricing of Outposts configurations

Priced for Amazon EC2 and Amazon EBS capacity in the SKU. Three-year term with partial upfront, all upfront, and no upfront options available. Price includes delivery, installation, servicing, and removal at the end of term.

AWS Services running locally on AWS Outposts will be charged on usage only. Amazon EC2 capacity and Amazon EBS storage upgrades available. Operating system charges are billed based on usage as an uplift to cover the license fee and no minimum fee required. Same AWS Region data ingress and egress charges apply. No additional data transfer charges for local network.

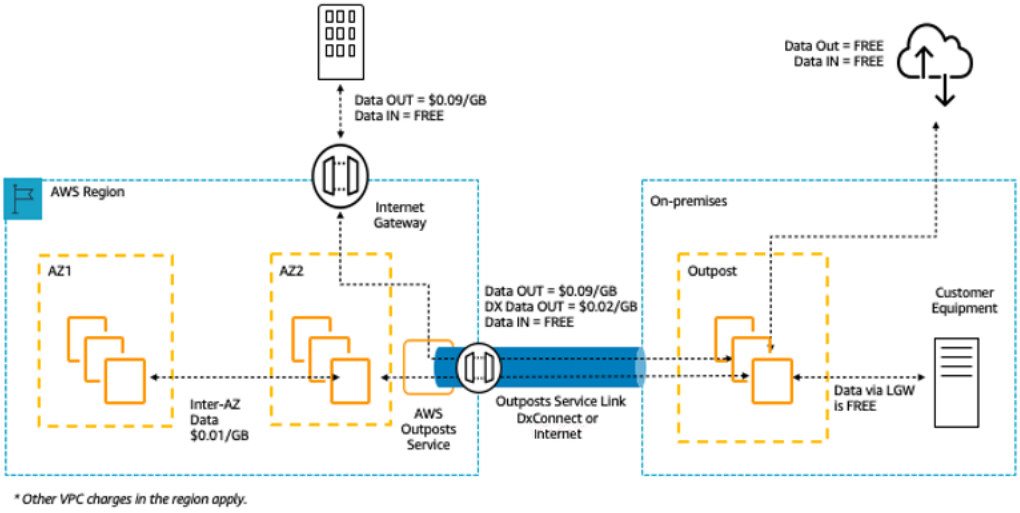

Figure 2: AWS Outposts ingress/egress charges

For more information, see the [AWS Outposts pricing page.]

AWS Snow Family

The AWS Snow Family helps customers that need to run operations in austere, non-data center environments, and in locations where there's lack of consistent network connectivity. The Snow Family, comprised of AWS Snowcone, AWS Snowball, and AWS Snowmobile, offers a number of physical devices and capacity points, most with built-in computing capabilities. These services help physically transport up to exabytes of data into and out of AWS. Snow Family devices are owned and managed by AWS and integrate with AWS security, monitoring, storage management, and computing capabilities.

AWS Snowcone

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.

With AWS Snowcone, you pay only for the use of the device and for data transfer out of AWS. Data transferred offline into AWS with Snowcone does not incur any transfer fees. For online data transfer pricing with AWS DataSync, please refer to the DataSync pricing page. Standard pricing applies once data is stored in the AWS Cloud.

For AWS Snowcone, you pay a service fee per job, which includes five days usage on-site, and for any extra days you have the device on-site. For high-volume deployments, contact your AWS sales team.

For pricing details, see AWS Snowcone Pricing.

AWS Snowball

AWS Snowball is a 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. Customers can use these two options for data collection, machine learning and processing, and storage in environments with

intermittent connectivity (such as manufacturing, industrial, and transportation) or in extremely remote locations (such as military or maritime operations) before shipping it back to AWS. These devices may also be rack mounted and clustered together to build larger, temporary installations.

AWS Snowball has three pricing elements to consider: usage, device type, and term of use.

First, understand your planned use case. Is it data transfer only, or will you be running compute on the device? You can use either device for data transfer or computing, but it is more cost-effective to use a Snowball Edge Storage Optimized for data transfer jobs.

Second, choose your device, either Snowball Edge Storage Optimized or Snowball Edge Compute Optimized. You can also select the option to run GPU instances on Snowball Edge Compute Optimized for edge applications.

For on-demand use, you pay a service fee per data transfer job, which includes 10 days of on-site Snowball Edge device usage. Shipping days, including the day the device is received and the day it is shipped back to AWS, are not counted toward the 10 days. After the 10 days, you pay a low per-day fee for each additional day you keep the device.

For 1-year or 3-year commitments, please contact your sales team; you cannot make this selection in the AWS Console.

Data transferred into AWS does not incur any data transfer fees, and standard pricing applies for data stored in the AWS Cloud.

For pricing details, see AWS Snowball Pricing.

AWS Snowmobile

AWS Snowmobile moves up to 100 PB of data in a 45-foot long ruggedized shipping container and is ideal for multi-petabyte or Exabyte-scale digital media migrations and data center shutdowns. A Snowmobile arrives at the customer site and appears as a network-attached data store for more secure, high-speed data transfer. After data is transferred to Snowmobile, it is driven back to an AWS Region where the data is loaded into Amazon S3.

Snowmobile pricing is based on the amount of data stored on the truck per month.

Snowmobile can be made available for use with AWS services in select AWS regions. Please follow up with AWS Sales to discuss data transport needs for your specific region and schedule an evaluation.

For pricing details, see AWS Snowmobile Pricing.

Amazon RDS

Amazon RDS is a web 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, so you can focus on your applications and business.

Estimating Amazon RDS costs

The factors that drive the costs of Amazon RDS include:

- **Clock hours of server time:** Resources incur charges when they are running—for example, from the time you launch a DB instance until you terminate it.
- **Database characteristics:** The physical capacity of the database you choose will affect how much you are charged. Database characteristics vary depending on the database engine, size, and memory class.
- **Database purchase type**: When you use On-Demand DB Instances, you pay for compute capacity for each hour your DB Instance runs, with no required minimum commitments. With Reserved DB Instances, you can make a low, one-time, upfront payment for each DB Instance you wish to reserve for a 1- or 3-year term.
- **Number of database instances:** With Amazon RDS, you can provision multiple DB instances to handle peak loads.
- **Provisioned storage**: There is no additional charge for backup storage of up to 100 percent of your provisioned database storage for an active DB Instance. After the DB Instance is terminated, backup storage is billed per GB per month.
- **Additional storage:** The amount of backup storage in addition to the provisioned storage amount is billed per GB per month.

- **Long Term Retention**: Long Term Retention is priced per vCPU per month for each database instance in which it is enabled. The price depends on the RDS instance type used by your database, and may vary by region. If Long Term Retention is turned off, performance data older than 7 days is deleted.
- **API Requests:** The API free tier includes all calls from the Performance Insights dashboard as well as 1 million calls outside of the Performance Insights dashboard. API requests outside of the Performance Insights free tier are charged at $0.01 per 1,000 requests.
- **Deployment type:** You can deploy your DB Instance to a single Availability Zone (analogous to a standalone data center) or multiple Availability Zones (analogous to a secondary data center for enhanced availability and durability). Storage and I/O charges vary, depending on the number of Availability Zones you deploy to.
- **Data transfer:** Inbound data transfer is free, and outbound data transfer costs are tiered.

Depending on your application's needs, it's possible to optimize your costs for Amazon RDS database instances by purchasing reserved Amazon RDS database instances. To purchase Reserved Instances, you make a low, one-time payment for each instance you want to reserve and in turn receive a significant discount on the hourly usage charge for that instance.

For more information, see Amazon RDS pricing.

Amazon DynamoDB

Amazon DynamoDB is a fast and flexible NoSQL database service for all applications that need consistent, single-digit millisecond latency at any scale. It is a fully managed cloud database and supports both document and key-value store models. Its flexible data model, reliable performance, and automatic scaling of throughput capacity make it a great fit for mobile, web, games, ad tech, IoT, and many other applications.

Amazon DynamoDB pricing at a glance

DynamoDB charges for reading, writing, and storing data in your DynamoDB tables, along with any optional features you choose to enable. DynamoDB has two capacity modes and those come with specific billing options for processing reads and writes on your tables: on-demand capacity mode and provisioned capacity mode.

DynamoDB read requests can be either strongly consistent, eventually consistent, or transactional.

On-Demand Capacity Mode

With on-demand capacity mode, you pay per request for the data reads and writes your application performs on your tables. You do not need to specify how much read and write throughput you expect your application to perform as DynamoDB instantly accommodates your workloads as they ramp up or down. DynamoDB charges for the core and optional features of DynamoDB.

Table 1: Amazon DynamoDB On-Demand Pricing

Core Feature Billing unit	Details
Read request unit (RRU)	API calls to read data from your table are billed in RRU. A strongly consistent read request of up to 4 KB requires one RRU. For items larger than 4 KB, additional RRUs are required. For items up to 4 KB, An eventually consistent read request requires one-half RRU. A transactional read request requires two RRUs
Write request unit (WRU)	Each API call to write data to your table is a WRU A *standard* WRU can write an item up to 1KB. Items larger than 1 KB require additional WRUs. *Transactional* write requires two WRUs.

Example RRU:

- A strongly consistent read request of an 8 KB item requires two read request units
- An eventually consistent read of an 8 KB item requires one read request unit.
- A transactional read of an 8 KB item requires four read request units.

Example WRU:

- A write request of a 1 KB item requires one WRU
- A write request of a 3 KB item requires three WRUs.
- A transactional write request of a 3 KB item requires six WRUs.

For details on how DynamoDB charges for the core and optional features of DynamoDB, see Pricing for On-Demand Capacity.

Provisioned Capacity Mode

With provisioned capacity mode, you specify the number of data reads and writes per second that you require for your application. You can use auto scaling to automatically adjust your table's capacity based on the specified utilization rate to ensure application performance while reducing costs.

Table 2: Amazon DynamoDB Provisioned Capacity Mode

Core Feature Billing unit	Details
Read Capacity unit (RCU)	API calls to read data from your table is an RCU. Items up to 4 KB in size, one RCU can perform one *strongly consistent* read request per second. For Items larger than 4 KB require additional RCUs For items up to 4 KB, One RCU can perform two *eventually consistent* read requests per second Transactional read requests require two RCUs to perform one read per second
Write Capacity Unit (WCU)	Each API call to write data to your table is a write request For items up to 1 KB in size, one WCU can perform one *standard* write request per second Items larger than 1 KB require additional WCUs. *Transactional* write requests require two WCUs to perform one write per second for items up to 1 KB
Data Storage	DynamoDB monitors the size of tables continuously to determine storage charges DynamoDB measures the size of your billable data by adding the raw byte size of the data you upload plus a per-item storage overhead of 100 bytes to account for indexing. First 25 GB stored per month is free

Example WCU

- A standard write request of a 1 KB item would require one WCU.
- A standard write request of a 3 KB item would require three WCUs.

- A transactional write request of a 3 KB item would require six WCUs.

Example RCU:

- A strongly consistent read of an 8 KB item would require two RCUs.
- An eventually consistent read of an 8 KB item would require one RCU.
- A transactional read of an 8 KB item would require four RCUs.

For details see Amazon DynamoDB pricing.

Data transfer

There is no additional charge for data transferred between Amazon DynamoDB and other AWS services within the same Region. Data transferred across Regions (e.g., between Amazon DynamoDB in the US East (Northern Virginia) Region and Amazon EC2 in the EU (Ireland) Region) will be charged on both sides of the transfer.

Global tables

Global tables builds on DynamoDB's global footprint to provide you with a fully managed, multi-region, and multi-master database that provides fast local read and write performance for massively scaled, global applications. Global tables replicates your Amazon DynamoDB tables automatically across your choice of AWS Regions.

DynamoDB charges for global tables usage based on the resources used on each replica table. Write requests for global tables are measured in replicated WCUs instead of standard WCUs. The number of replicated WCUs consumed for replication depends on the version of global tables you are using.

Read requests and data storage are billed consistently with standard tables (tables that are not global tables). If you add a table replica to create or extend a global table in new Regions, DynamoDB charges for a table restore in the added regions per gigabyte of data restored. Cross-Region replication and adding replicas to tables that contain data also incur charges for data transfer out.

For more information, see Best Practices and Requirements for Managing Global Tables.

Learn more about pricing for additional DynamoDB features at the Amazon DynamoDB pricing page.

Amazon CloudFront

Amazon CloudFront is a global content delivery network (CDN) service that securely delivers data, videos, applications, and APIs to your viewers with low latency and high transfer speeds.

Amazon CloudFront pricing

Amazon CloudFront charges are based on the data transfers and requests used to deliver content to your customers. There are no upfront payments or fixed platform fees, no long-term commitments, no premiums for dynamic content, and no requirements for professional services to get started. There is no charge for data transferred from AWS services such as Amazon S3 or Elastic Load Balancing. And, best of all, you can get started with CloudFront for free.

When you begin to estimate the cost of Amazon CloudFront, consider the following:

- **Data Transfer OUT (Internet/Origin):** The amount of data transferred out of your Amazon CloudFront edge locations.
- **HTTP/HTTPS Requests:** The number and type of requests (HTTP or HTTPS) made and the geographic region in which the requests are made.
- **Invalidation Requests:** No additional charge for the first 1,000 paths requested for invalidation each month. Thereafter, $0.005 per path requested for invalidation.
- **Field Level Encryption Requests**: Field-level encryption is charged based on the number of requests that need the additional encryption; you pay $0.02 for every 10,000 requests that CloudFront encrypts using field-level encryption in addition to the standard HTTPS request fee.
- **Dedicated IP Custom SSL:** $600 per month for each custom SSL certificate associated with one or more CloudFront distributions using the Dedicated IP version of custom SSL certificate support. This monthly fee is pro-rated by the hour.

For more information, see Amazon CloudFront pricing.

Amazon Kendra

Amazon Kendra is a highly accurate and easy to use enterprise search service that's powered by machine learning. Amazon Kendra enables developers to add search

capabilities to their applications so their end users can discover information stored within the vast amount of content spread across their company. When you type a question, the service uses machine learning algorithms to understand the context and return the most relevant results, whether that be a precise answer or an entire document. For example, you can ask a question like "How much is the cash reward on the corporate credit card?" and Amazon Kendra will map to the relevant documents and return a specific answer like "2%".

Amazon Kendra pricing

With the Amazon Kendra service, you pay only for what you use. There is no minimum fee or usage requirement. Once you provision Amazon Kendra by creating an index, you are charged for Amazon Kendra hours from the time an index is created until it is deleted. Partial index instance-hours are billed in one-second increments. This applies to Kendra Enterprise Edition and Kendra Developer Edition.

Amazon Kendra comes in two editions. Kendra Enterprise Edition provides a high-availability service for production workloads. Kendra Developer Edition provides developers with a lower-cost option to build a proof-of-concept; this edition is not recommended for production workloads.

You can get started for free with the Amazon Kendra Developer Edition that provides free usage of up to 750 hours for the first 30 days. Connector usage does not qualify for free usage, regular run time and scanning pricing will apply. If you exceed the free tier usage limits, you will be charged the Amazon Kendra Developer Edition rates for the additional resources you use. See Amazon Kendra Pricing for pricing details.

Amazon Macie

Amazon Macie is a fully managed data security and data privacy service that uses machine learning and pattern matching to discover and protect your sensitive data in AWS. Amazon Macie uses machine learning and pattern matching to cost efficiently discover sensitive data at scale. Macie automatically detects a large and growing list of sensitive data types, including personally identifiable information (PII) such as names, addresses, and credit card numbers. It also gives you constant visibility of the data security and data privacy of your data stored in Amazon S3. Macie is easy to set up with one click in the AWS Management Console or a single API call. Macie provides multi-account support using AWS Organizations, so you can enable Macie across all of your accounts with a few clicks.

Amazon Macie pricing

With Amazon Macie, you are charged based on the number of Amazon S3 buckets evaluated for bucket-level security and access controls and the quantity of data processed for sensitive data discovery.

When you enable Macie, the service will gather detail on all of your S3 buckets, including bucket names, size, object count, resource tags, encryption status, access controls, and region placement. Macie will then automatically and continually evaluate all of your buckets for security and access control, alerting you to any unencrypted buckets, publicly accessible buckets, or buckets shared with an AWS account outside of your organization. You are charged based on the total number of buckets in your account after the 30-day free trial and charges are pro-rated per day.

After enabling the service, you are able to configure and submit buckets for sensitive data discovery. This is done by selecting the buckets you would like scanned, configuring a one-time or periodic sensitive data discovery job, and submitting it to Macie. Macie only charges for the bytes processed in supported object types it inspects. As part of Macie sensitive data discovery jobs, you will also incur the standard Amazon S3 charges for GET and LIST requests. See *Requests and data retrievals* pricing on the Amazon S3 pricing page.

Free tier | Sensitive data discovery

For sensitive data discovery jobs, the first 1 GB processed every month in each account comes at no cost. For each GB processed beyond the first 1 GB, charges will occur. Please refer this link for pricing details. *You are only charged for jobs you configure and submit to the service for sensitive data discovery

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.

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. See Amazon Kinesis Data Streams Pricing for pricing details.

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. See Amazon Kinesis Data Firehose Pricing for pricing details.

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. See Amazon Kinesis Data Analytics Pricing for pricing details.

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.

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.

Amazon Kinesis Video Streams pricing

You pay only for the volume of data you ingest, store, and consume in your video streams.

WebRTC pricing

If you use WebRTC capabilities, you pay for the number of signaling channels that are active in a given month, number of signaling messages sent and received, and TURN streaming minutes used for relaying media. A signaling channel is considered active in

a month if at any time during the month a device or an application connects to it. TURN streaming minutes are metered in 1-minute increments.

Note: You will incur standard AWS data transfer charges when you retrieve data from your video streams to destinations outside of AWS over the internet.

See Amazon Kinesis Video Streams Pricing for pricing details.

AWS IoT Events

AWS IoT Events helps companies continuously monitor their equipment and fleets of devices for failure or changes in operation and trigger alerts to respond when events occur. AWS IoT Events recognizes events across multiple sensors to identify operational issues, such as equipment slowdowns, and generates alerts such as notifying support teams of an issue. AWS IoT Events offers a managed complex event detection service on the AWS Cloud, accessible through the AWS IoT Events console, a browser-based GUI where you can define and manage your event detectors, or direct ingest application program interfaces (APIs), code that allows two applications to communicate with each other. Understanding equipment or a process based on telemetry from a single sensor is often not possible; a complex event detection service will combine multiple sources of telemetry to gain full insight into equipment and processes. You define conditional logic and states inside AWS IoT Events to evaluate incoming telemetry data to detect events in equipment or a process. When AWS IoT Events detects an event, it can trigger pre-defined actions in another AWS service, such as sending alerts through Amazon Simple Notification Service (Amazon SNS).

AWS IoT Events pricing

With AWS IoT Events, you pay only for what you use with no minimum fees or mandatory service usage. When you create an event detector in AWS IoT Events, you apply conditional logic such as if-then-else statements to understand events, such as when a motor might be stuck. You are only charged for each message that is evaluated in AWS IoT Events.

See AWS IoT Events Pricing for pricing details.

The AWS Free Tier is available to you for 12 months starting on the date you create your AWS account. When your free usage expires or if your application use exceeds the free usage tiers, you simply pay the above rates. Your usage is calculated each month across all regions and is automatically applied to your bill. Note that free usage does not accumulate from one billing period to the next.

AWS Cost Optimization

AWS enables you to take control of cost and continuously optimize your spend, while building modern, scalable applications to meet your needs. AWS's breadth of services and pricing options offer the flexibility to effectively manage your costs and still keep the performance and capacity you require. AWS is dedicated to helping customers achieve highest saving potential. During this period of crisis, we will work with you to develop a plan that meets your financial needs. Get started with the steps below that will have an immediate impact on your bill today.

Choose the right pricing models

Use Reserved Instances (RI) to reduce Amazon RDS, Amazon Redshift, Amazon ElastiCache, and Amazom Elasticsearch costs

For certain services like Amazon EC2 and Amazon RDS, you can invest in reserved capacity. With Reserved Instances, you can save up to 72% over equivalent on-demand capacity. Reserved Instances are available in 3 options – All up-front (AURI), partial up-front (PURI) or no upfront payments (NURI). Use the recommendations provided in AWS Cost Explorer RI purchase recommendations, which is based on your Amazon RDS, Amazon Redshift, Amazon ElastiCache, and Elasticsearch usage.

Amazon EC2 Cost Savings

Use Amazon EC2 Spot Instances to reduce EC2 costs or use Compute Savings Plans to reduce EC2, Fargate and Lambda cost.

Match Capacity with Demand

Identify Amazon EC2 instances with low-utilization and reduce cost by stopping or rightsizing

Use AWS Cost Explorer Resource Optimization to get a report of EC2 instances that are either idle or have low utilization. You can reduce costs by either stopping or downsizing these instances. Use AWS Instance Scheduler to automatically stop instances. Use AWS Operations Conductor to automatically resize the EC2 instances (based on the recommendations report from Cost Explorer).

Identify Amazon RDS, Amazon Redshift instances with low utilization and reduce cost by stopping (RDS) and pausing (Redshift)

Use the Trusted Advisor Amazon RDS Idle DB instances check, to identify DB instances which have not had any connection over the last 7 days. To reduce costs, stop these DB instances using the automation steps described in this blog post. For Redshift, use the Trusted Advisor Underutilized Redshift clusters check, to identify clusters which have had no connections for the last 7 days, and less than 5% cluster wide average CPU utilization for 99% of the last 7 days. To reduce costs, pause these clusters using the steps in this blog.

Analyze Amazon DynamoDB usage and reduce cost by leveraging Autoscaling or On-demand

Analyze your DynamoDB usage by monitoring 2 metrics, ConsumedReadCapacityUnits and ConsumedWriteCapacityUnits, in CloudWatch. To automatically scale (in and out) your DynamoDB table, use the AutoScaling feature. Using the steps here, you can enable AutoScaling on your existing tables. Alternately, you can also use the on-demand option. This option allows you to pay-per-request for read and write requests so that you only pay for what you use, making it easy to balance costs and performance.

Implement processes to identify resource waste

Identify Amazon EBS volumes with low-utilization and reduce cost by snapshotting then deleting them

EBS volumes that have very low activity (less than 1 IOPS per day) over a period of 7 days indicate that they are probably not in use. Identify these volumes using the Trusted Advisor Underutilized Amazon EBS Volumes Check. To reduce costs, first snapshot the volume (in case you need it later), then delete these volumes. You can automate the creation of snapshots using the Amazon Data Lifecycle Manager. Follow the steps here to delete EBS volumes.

Analyze Amazon S3 usage and reduce cost by leveraging lower cost storage tiers

Use S3 Analytics to analyze storage access patterns on the object data set for 30 days or longer. It makes recommendations on where you can leverage S3 Infrequently Accessed (S3 IA) to reduce costs. You can automate moving these objects into lower cost storage tier using Life Cycle Policies. Alternately, you can also use S3 Intelligent-

Tiering, which automatically analyzes and moves your objects to the appropriate storage tier.

Review networking and reduce costs by deleting idle load balancers

Use the Trusted Advisor Idle Load Balancers check to get a report of load balancers that have RequestCount of less than 100 over the past 7 days. Then, use the steps here, to delete these load balancers to reduce costs. Additionally, use the steps provided in this blog, review your data transfer costs using Cost Explorer.

AWS Support Plan Pricing

AWS Support provides a mix of tools and technology, people, and programs designed to proactively help you optimize performance, lower costs, innovate faster and focused on solving some of the toughest challenges that hold you back in your cloud journey.

There are three types of support plans available: Developer, Business, and Enterprise. For more details, see Compare AWS Support Plans and AWS Support Plan Pricing.

Cost calculation examples

The following sections use the AWS Pricing Calculator to provide example cost calculations for two use cases.

AWS Cloud cost calculation example

This example is a common use case of a dynamic website hosted on AWS using Amazon EC2, AWS Auto Scaling, and Amazon RDS. The Amazon EC2 instance runs the web and application tiers, and AWS Auto Scaling matches the number of instances to the traffic load. Amazon RDS uses one DB instance for its primary storage, and this DB instance is deployed across multiple Availability Zones.

Architecture

Elastic Load Balancing balances traffic to the Amazon EC2 Instances in an AWS Auto Scaling group, which adds or subtracts Amazon EC2 Instances to match load. Deploying Amazon RDS across multiple Availability Zones enhances data durability and availability. Amazon RDS provisions and maintains a standby in a different Availability Zone for automatic failover in the event of outages, planned or unplanned. The following illustration shows the example architecture for a dynamic website using Amazon EC2,

AWS Auto Scaling, Security Groups to enforce least-privilege access to AWS infrastructure and selected architecture components, and one Amazon RDS database instance across multiple Availability Zones (Multi AZ deployment). All these components are deployed into single region and VPC. The VPC is spread out into two availability zones to support failover scenarios with and Route 53 Resolver to manage and route requests for 1 hosted zone towards Elastic Load Balancer.

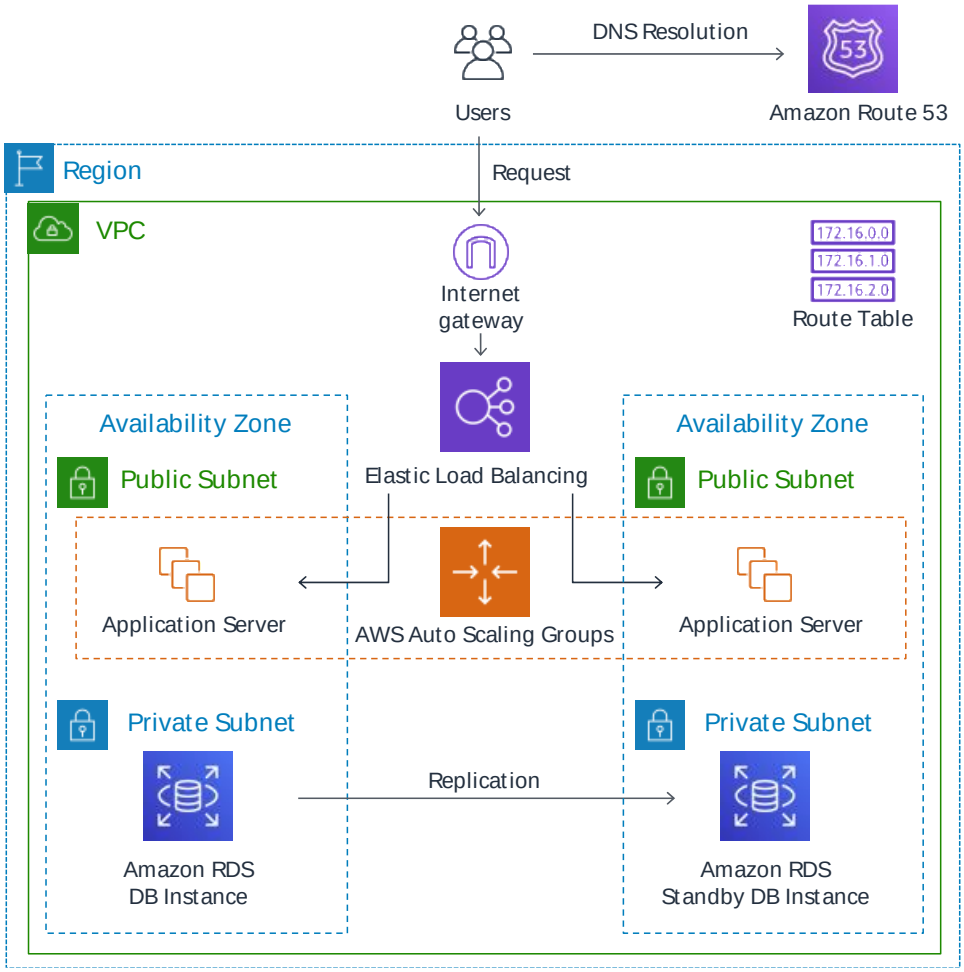

Figure 3: AWS Cloud deployment architecture

Daily usage profile

You can monitor daily usage for your application so that you can better estimate your costs. For instance, you can look at the daily pattern to figure out how your application handles traffic. For each hour, track how many hits you get on your website and how many instances are running, and then add up the total number of hits for that day.

Hourly instance pattern = (hits per hour on website) / (number of instances)

Examine the number of Amazon EC2 instances that run each hour, and then take the average. You can use the number of hits per day and the average number of instances for your calculations.

Daily profile = SUM(Hourly instance pattern) / 24

Amazon EC2 cost breakdown

The following table shows the characteristics for Amazon EC2 used for this dynamic site in the US East Region.

Characteristic	Estimated Usage	Description
Utilization	100%	All infrastructure components run 24 hour per day, 7 days per week
Instance	t3a.xlarge	16 GB memory, 4 vCPU
Storage	Amazon EBS SSD gp2	1 EBS volume per instance with 30 GB of storage per volume
Data backup	Daily EBS snapshots	1 EBS volume per instance with 30 GB of storage per volume
Data transfer	Data in: 1 Tb/month Data out: 1 Tb/month	10% incremental change per day
Instance scale	4	On average per day, there are 4 instances running
Load Balancing	20 Gb/Hour	Elastic Load Balancing is used 24 hours per day, 7 days per week. It processes a total of 20 Gb/Hour (data in + data out)
Database	MySQL, db.m5.large instance with 8 GB memory, 2 vCPUs, 100 GB storage	Multi-AZ deployment with synchronous standby replica in separate Availability Zone

The total cost for one month is the sum of the cost of the running services and data transfer out, minus the AWS Free Tier discount. We calculated the total cost using the AWS Pricing Calculator.

Table 3: Cost breakdown

Service	Monthly	Annually	Configuration
Elastic Load Balancing	$87.60	$1051.20	Number of Network Load Balancers (1), Processed bytes per NLB for TCP (20 GB per hour)
Amazon EC2	$439.16	$5269.92	Operating system (Linux), Quantity (4), Storage for each EC2 instance (General Purpose SSD (gp2)), Storage amount (30 GB), Instance type (t3a.xlarge)
Amazon Elastic IP address	$0	$0	Number of EC2 instances (1), Number of EIPs per instance (1)
Amazon RDS for MySQL	$272.66	$ 3271.92	Quantity (1) db.m5.large, Storage for each RDS instance (General Purpose SSD [gp2]), Storage amount (100 GB)
Amazon Route 53	$183.00	$2,196.00	Hosted Zones (1), Number of Elastic Network Interfaces (2), Basic Checks Within AWS (0)
Amazon Virtual Private Cloud (Amazon VPC)	$92.07	$1,104.84	Data Transfer cost, Inbound (from: Internet) 1 TB per month Outbound (to: Internet) 1 TB per month Intra-Region 0 TB per month

Hybrid cloud cost calculation example

This example is a hybrid cloud use case of AWS Outposts deployed on-premises connected to AWS Cloud using AWS Direct Connect. AWS Outposts extends the existing VPC from the selected AWS Region to the customer data center. Selected AWS services required to run on-premises (i.e. Amazon EKS) are available at AWS Outposts inside the Outpost Availability Zone, deployed inside a separate subnet.

Hybrid architecture description

The following example shows Outpost deployment with distributed Amazon EKS service extending to on-premises environments.

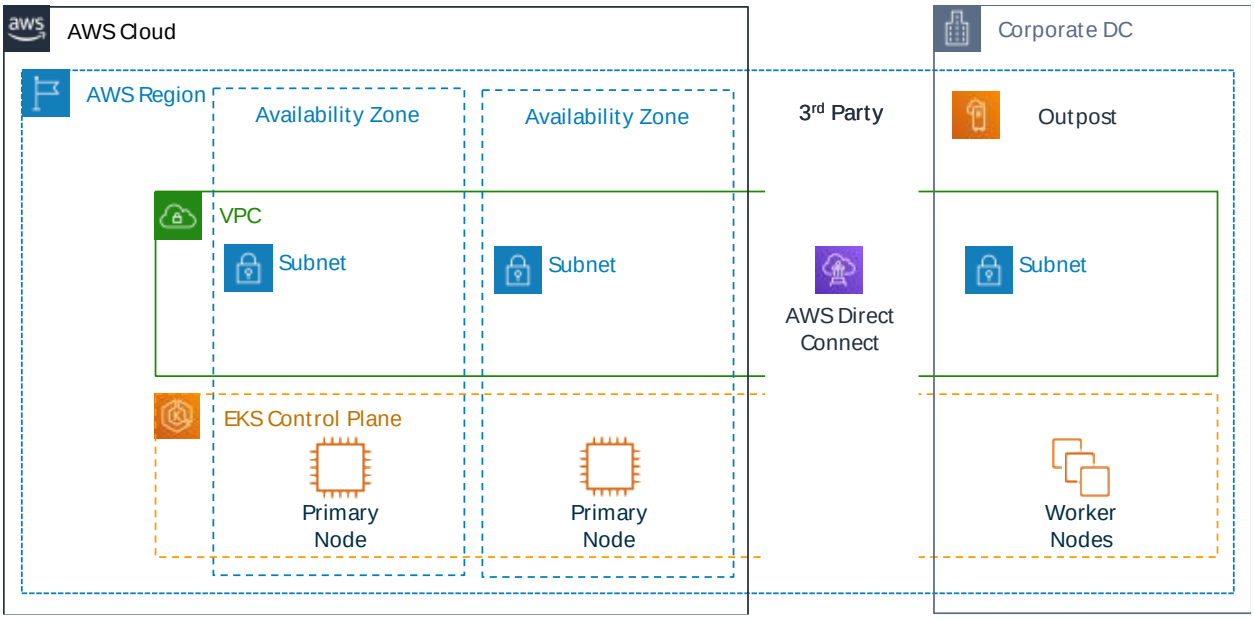

Figure 4: AWS Outpost with Amazon EKS Control Plane and Data Plane Architecture

Architecture

- The Control Plane for Amazon EKS remains in the Region, which means in the case of Amazon EKS, the Kubernetes Primary node will stay in the Availability Zone deployed to the Region (not on the Outposts).
- The Amazon EKS worker nodes are deployed on the Outpost, controlled by Primary node deployed in the Availability Zone.

Traffic Flow

- The EKS Control Plane Traffic between EKS, AWS metrics and Amazon CloudWatch transits third-party network (AWS Direct Connect/AWS Site-to-Site VPN to the AWS Region).
- The Application / Data Traffic is isolated from Control plane and distributed between Outposts and local network.
- Distribution of AMIs (deployed on Outpost) is driven by central Amazon ECR in Region, however all images are cached locally on the Outpost.

Load Balancers

- Application Load Balancer is supported on Outpost as the only local Elastic Load Balancing available

- The Network Load Balancer and Classic Load Balancer stay in the Region, but targets deployed at AWS Outposts are supported (including Application Load Balancer).
- On-premises (inside corporate DC) Load Balancers (i.e. F5 BIG IP, NetScaler) can be deployed and routed via Local Gateway (inside AWS Outpost).

Hybrid cloud components selection

Customers can choose from a range of pre-validated Outposts configurations (*Figure 2*) offering a mix of EC2 and EBS capacity designed to meet a variety of application needs. AWS can also work with customer to create a customized configuration designed for their unique application needs.

To consider correct configuration, make sure to verify deployment and operational parameters of the selected physical location for AWS Outpost rack installation. The following example represents a set of parameters highlighting facility, networking and power requirements needed for location validation (selected parameter: example value):

Purchase Option: All Upfront

Term: 3 Years

Max on premises power capacity: 20kVA

Max weight: 2,500lb

Networking uplink speed: 100Gbps

Number of Racks: 1

Average Power Draw per Rack: 9.34

Constraint (power draw/weight): Power Draw

Total Outpost vCPU: 480

Total Outpost Memory: 2,496GiB

In addition to minimum parameters, you should make deployment assumptions prior to any order to minimize performance and security impact on existing infrastructure landscape, deeply affecting existing cost of on-premises infrastructure (selected question: example assumption).

What is the speed of the uplink ports from your Outposts Network Devices (OND): 40 or 100Gbps

How many uplinks per Outpost Networking Device (OND) will you use to connect the AWS Outpost to your network: 4 uplinks

How will the Outpost service link (the Outpost control plane) access AWS services: Service link will access AWS over a Direct Connect public VIF

Is there a firewall between Outposts and the Internet: Yes

These assumptions together with selected components will further lead to an architecture with higher granularity of details influencing overall cost of a hybrid cloud architecture deployment (*Figure 5*).

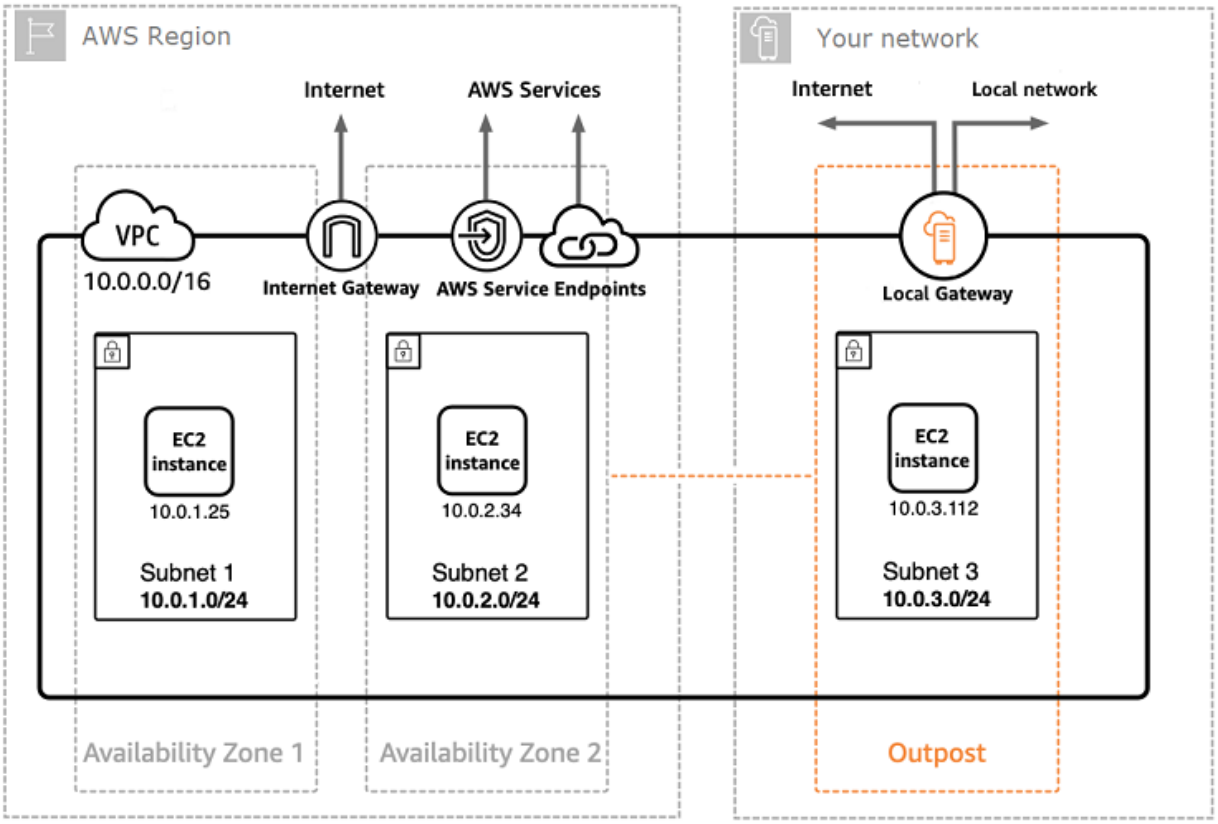

Figure 5: Hybrid cloud architecture deployment example

Hybrid cloud architecture cost breakdown

Hybrid cloud cost include multiple layers and components deployed across the AWS cloud and on-premises location. When you use AWS Managed Services on AWS Outposts, you are charged only for the services based on usage by instance-hour and excludes underlying EC2 instance and EBS storage charges.

Breakdown of these services is showcased in next sections for a 3-year term with partial upfront, all upfront, and no upfront options (EC2 and EBS capacity). Price includes delivery, installation, servicing and removal at the end of term – there is no additional charge.

Outpost rack charges (customized example)

EC2 Charges

- c5.24xlarge, 11 TB
  - $7,148.67 monthly;
  - $123,650.18 upfront, $3,434.73 monthly
  - $239,761.41 upfront
- 1 m5.24xlarge, 11 TB
  - $7,359.69 monthly
  - $127,167.06 upfront, $3,532.42 monthly
  - $246,373.14 upfront

EBS Charges

- 11 TB EBS tier is priced at $0.30/GB monthly

Conclusion

Although the number and types of services offered by AWS have increased dramatically, our philosophy on pricing has not changed. You pay as you go, pay for what you use, pay less as you use more, and pay even less when you reserve capacity. All these options are empowering AWS customers to choose they preferred pricing model and increase flexibility of their cost strategy.

What is Cloud Computing?

- On-demand delivery of compute, database storage, applications, other IT resources through cloud platform via Internet
- Pay-as-you-go pricing

6 Advantages of Cloud Computing

1. Trade capital expense for variable expense
   - Only pay for what you use
2. Benefit from massive economies of scale
   - You won't have same purchasing power as Amazon
   - They get cheaper prices to purchase servers, hardware
3. Stop guessing about capacity
   - You'll buy too much or too little. Too much = wasted money, too little
4. Increase speed and agility
   - Websites/apps can scale infinitely with demand
5. Stop spending money running/maintaining data centers
   - Focus on what you're good at, not managing infrastructures
6. Go global in minutes
   - Deploy apps in minutes
   - Provide lower latency and better experience at minimal cost

3 Types of Cloud Computing

1. Infrastructure as a Service (IAAS)
   - You manage the server (physical or virtual) as well as the operating system
   - Data center provider has no access to your server
2. Platform as a Service (PAAS)
   - Someone else manages the underlying hardware and operating systems, you just focus on your applications
   - You upload your code and it just executes
   - Think of GoDaddy
3. Software as a Service (SAAS)
   - Think of Gmail
   - All you do is interact with the application, manage the software and how you want to use it
   - Someone else takes care of the infrastructure and everything related to it

3 Types of Cloud Computing Deployments

1. Public Cloud – AWS, Azure, Google Cloud Platform
2. Hybrid – Mix of public and private
   - May want to keep some sensitive data on-premise
3. Private Cloud (or on-premise) – You manage it in your data center. Openstack or Vmware

Around the World with AWS

Region

- Geographic area consisting of 2 or more availability zones

Availability Zone

- A data center

Edge Location

- CDN Endpoints for CloudFront
- Many more edge locations than regions

Let's Log Into AWS

Support Plans

1. Basic
2. Developer

- Experimenting with AWS
- $29/month
- One person can ask technical questions through support center, 12-24 hour support rate

3. Business
   - 24/7 support by phone
   - Full access to AWS Trusted Advisor
   - $100/mo
4. Enterprise
   - $15,000/month
   - Everything in business + technical account manager
   - 15 min response time for critical support cases

Create Billing Alarm

- Click Name at top-right, click My Billing Dashboard
- Enable Billing Alert
- Add the threshold in Cloudwatch and add e-mail address

Identify Access Management

Tick off five marks on Security Status:

1. Delete root access keys
   - We can skip this as a new user
2. Enable MFA (multi-factor authentication)
   - Set up MFA with Google Authenticator
3. Create individual IAM users
   - Access types:
     - **Programmatic access:** Access via command line
     - **AWS Management Console access:** Login to AWS console and make changes
     - **AWS SDK access:**
4. Create IAM groups
   - Once you decide access, you need to add to a Group
   - Choose a policy access type (or multiple types) and add the group name
5. IAM password policy
   - A password policy is a set of rules that define the type of password an IAM user can set
   - Change upper/lowercase required, number required, min password length, etc.

Policies

- How to define permissions to users, groups, roles
- You can click on a policy to get details about what it gives people access to
- Detailed JSON allows you to define a **statement** comprised of an effect (Allow, Disallow), Action (what will happen), and Resource (to what resource)

Exam Tips

3 ways to access AWS platform:

1. Via AWS Console
2. Programatically using command line
3. Using Software Development Kit (SDK)

Root account

- Full admin access
- Should never give away
- Instead, create a user for each individual in your organization and secure root account with MFA

Groups

- Place to store your users
- Users will inherit all permissions that group has
- To set permissions for a group, need to set policies for that group using JSON

S3 (Simple Storage Service)

Overview

- Provides developers and IT teams with secure, durable, highly-scalable object storage
- Safe place to store your files
- Object-based storage (pictures, Word files, videos), not operating system or database
- Files can be anywhere from 0 bytes to 5 TB in size
- Unlimited storage, but you pay by the gig
- Files are stored in **buckets**
- Bucket is a folder in the cloud
- Buckets are universal namespace. Must be unique globally.
- URL looks like **https://s3-eu-west-1.amazonaws.com/acloudguru**
  - Region + Region # + amazonaws.com + bucket name
- You will receive HTTP 200 code if file upload is successful

Data Consistency Model

- Read after Write consistency for PUTS of new objects
  - If you read a file as soon as you upload it, you'll be able to read the file
- Eventual consistency for overwrite PUTS and DELETES (can take some time to propogate)
  - If you update a file and overwrite the old version, you may get the old file or the new file. It will eventually show up.
  - You may be updating to one availability zone, may take time to propagate to other availability zones

S3 Key-Value Store

- S3 is object based, objects consist of:
  1. Key (name of object, e.g. *hello.txt*
  2. Data in file (sequence of bytes)
  3. Version ID (important for versioning)
  4. Metadata (data about data, e.g. tags)
  5. Subresources:
     1. Access Control List
     2. Torrents

Other Points

- Built for 99.99% availability
- Amazon guarantees 99.99999999999% (11 x 9s) durability
  - If you upload x amount of files, 99.99999999999% of those files will actually be uploaded (you won't lose any files)
- Tiered storage availability
- Lifecycle management
  - If a file is over 30 days old, move it from one storage tier to another and eventually archive to Glacier
- Versioning
  - Multiple versions of a file
- Encryption
- Secure data with Access Control Lists and Bucket Policies
  - Bucket policies: Policy is for a specific bucket
  - ACL: Individual file level, control who can access a specific file

S3 Storage Tiers/Classes

1. S3 Standard
   - 99.99% availability
   - 99.99999999999% durability
   - Stored redundantly across multiple devices (multiple disks) and multiple facilities (multiple availability zones)
   - Designed to sustain loss of 2 facilities concurrently
2. S3 – IA (Infrequently Accessed)
   - Data accessed less frequently but requires rapid access when needed
   - Lower fee than Standard but charged a retrieval fee
3. S3 One Zone – IA
   - Same as S3 – IA but do not require multiple availability zone data resilience

    - Only stored in one availability zone
4. Glacier
    - Used for archival only
    - Cheapest
    - Expedited, standard, or bulk
        - Expedited: Restored within few mins, high fee
        - Standard: 3-5 hours for restore
        - Bulk: 5-12 hours
- No retrieval fee for Standard, only for the other three

S3 – Charges

Charged for:

- Storage
- Requests
- Storage management pricing
    - Tags that define who owns an object
- Data transfer pricing
    - Transferring from one region to another
- Transfer acceleration
    - Fast and easy secure transfers across long distances

S3 Transfer Acceleration

- Users upload to an edge location instead of directly to S3 bucket
- Once it goes to an edge location, it automatically gets distributed to the S3 bucket
- File goes across Amazon's backbone to transfer much faster

Read the S3 FAQ before taking the exam!

Creating an S3 Bucket

- Buckets must have unique names
- **Note:** Interface for S3 is Global (similar to IAM), but buckets created can be deployed in any region
- Bucket names must be DNS compliant (3-63 characters, no invalid characters)
- By default, buckets are **private** (recommended)
- You can change storage class and encryption on the fly by using the More menu

Setting public access

- Trying to open a file through a URL won't work by default because public read access is not enabled. Need to enable when uploading.
- Click box next to file > More > Make public
- Another way: Click into file > Permissions tab > Everyone (under public access) > Read Object

Transfer acceleration

1. Click Properties in bucket
2. Advanced Settings > Transfer acceleration > Enable
- S3 has a feature to allow you to test your transfer speeds to different regions around the world

Cross Region Replication

- Management > Replication
- Allows you to replicate bucket in one region to bucket in another region in the world
- Useful for disaster recovery
- Any object upload to first bucket is automatically replicated to second bucket

S3 for Web Pages

- S3 can host **static** web pages (not dynamic like WordPress or PHP)
- It will scale automatically, will scale with demand. Useful for large number of requests.
- You can use bucket policies to make an entire bucket public (used for static websites)

CloudFront

- Amazon's CDN network

- Used to deliver entire website, including dynamic, static, streaming, and interactive content using edge locations
- Requests for content automatically routed to nearest ege location so content is delivered with best performance possible

CDN

- Content-delivery network
- System of services around the world that deliver web pages or web page content to a user based on user geographic location, origin of the webpage, and content delivery server
- Works with origin types listed below
- Also works with non-AWS origins

Edge Location

- Location where content will be cached
- Similar to AWS Region/AZ
- As close to user as possible

Origin

- Origin of files that CDN will distribute
- S3 bucket, EC2 instance, Elastic Load Balancer, or Route53

Distribution

- Name given to the CDN which consists of a collection of edge locations
- First time a user goes to a website, it'll check a local edge location to see if website asset is there
- If not, it will download the asset from the origin and cache it to the edge location
- Next time someone tries to access, they will get the cached version from a local edge location
- Reduces stress on web servers and increases speed to download large files

Distribution Types

1. Web distribution – Used for websites
2. RTMP – Used for media streaming

Setting up CloudFront

1. Choose Web distrbution
2. Origin Domain Name: Choose an S3 bucket
3. Origin Path: You can choose subdirectories for your origin

- Once it's deployed, you will see a domain name. Use that and the name of a file in your bucket to access.

Exam Tips

- Content comes from Origin
- Cached at a local Edge Location
- Takes awhile for the first person to access, much quicker every time after that because it's cached geographically close to you

EC2 (Elastic Cloud Compute)

Setup

- **VPC:** Virtual data center in the cloud
  - Deploy all EC2 instances into a VPC
- **AMI:** Using Amazon Linux AMI because it includes stuff to connect to AWS
- **Instance:** Choosing t2.micro because it's usually used to test in dev
- **Instance Details:**
  - **Network:** Keep default VPC
  - **Subnet:** Choose which availability zone you want to be put into
  - **Auto-assign Public IP:** Allows you to assign a public IP so you can SSH into instance
  - **Shutdown behavior:** Choose what happens if your EC2 instance turns out (stop or have Amazon terminate for you)
  - **Enable termination protection:** Prevents people from accidentally shutting down your instance
- **Storage:**
  - 8GB is default

    - **Volume Type:** General purpose is most common, Provisioned IOPS lets you choose a very fast disk (database server), Magnetic is a very slow disk (file server)
- **Tags:** Allow us to add tags like Department and Employee ID to help with cost tracking later on
- **Security Group:** Virtual firewall in the cloud
    - Open ports like 22 for SSH or 3389 for RDP (Windows) or 80 for HTTP

Connect to the EC2 server

- Open Terminal
- chmod 400 MyVirginiaKP.pem – Protects file from accidental overwriting
- ssh ec2-user@54.242.147.206 -i MyVirginiaKP.pem to connect
- sudo su for root
- yum update to update security patches

Exam Tips

- EC2 is compute-based, it's not serverless. It is a server!
- Use private key to connect to EC2
- Security groups are virtual firewalls in the cloud. Need to open ports in order to use them (22 for SSH, 80 for HTTP, 443 for HTTPS, 3389 for RDP)
- Always design for failure, have one EC2 instance in each availability zone

AWS Command Line

- Use aws configure on command line to set up login details
    - Enter Access Key and Secret Access Key
    - Region name: us-east-1
    - No output format
- aws s3 ls to view s3 buckets
- aws s3 mb s3://myacloudgurubucket2018 to make a bucket
- aws s3 cp hello.txt s3://myacloudgurubucket2018sheil to upload EC2 file to S3 bucket

Tips

- Interact with AWS in 3 ways:
    1. Using the console
    2. Using the command line interface (CLI)
    3. Using the software development kits

Using Roles

- Prevent account from getting hacked
- cd /~.aws and rm -rf credentials to remove credentials file
- Roles are a secure way to grant permission to entities that you trust
- AWS Console > IAM > Roles
- Create new EC2 role, choose S3 Admin access, and name the role
- In EC2, find the instance, choose Actions > Instance Settings > Attach/Replace IAM Role
- No Role to My Admin S3 Access (the role I created in previous step)
- This process allows me to access S3 via CLI without having to store credentials on the EC2 instance itself

Tips

- Roles are much more secure than using access key IDs and secret access keys
- Much easier to manage
- Can apply roles to EC2 instances at any time (not just when it boots up)
- Changes take place immediately
- Roles are universal, no need to specify region. Similar to users

Build a Website

- Connect to EC2 with CLI
- Web servers need either Apache (Linux) or IIS (Windows)
- yum install httpd -yes to install

- service httpd start to start server
- Anything you put into /var/www/html will be on the website
- aws s3 cp s3://myacloudgurubucket2018sheil /var/www/html --recursive to copy files from S3 to web server on EC2

Databases

Relational Database Service (RDS)
Types:

1. SQL Server
2. Oracle
3. MySQL Server
4. PostgreSQL
5. Aurora (Amazon's own database)
6. MariaDB

Two key features:

1. Multi availability zones for disaster recovery
2. Read replicas for performance improvement

- **Multi AZ:** Exact copy of your database in case the primary goes down
  - Disaster recovery
- **Read replica:** Spread read access across five databases, only one is for writing
  - Scaling out / performance

Nonrelational Databases

1. Collection (table)
2. Documents (row)
3. Key-value pairs (fields)

- Allows you to add in extra fields all the time
- **Amazon DynamoDB** is Amazon's nonrelational/NoSQL database
  - Fast, flexible
  - Scales with your application

Aurora

- Relational, Amazon's own
- 6 copies of itself
- 5 times better performance than MySQL, 1/10 price point
- Choose Aurora if you have an RDS
- Choose DynamoDB if you have nonrelational

Data Warehousing

- Used for business intelligence
- Used to pull in large and complex datasets
- Used by management to do queries (current performance targets, etc)
- **Redshift** is Amazon's data warehouse in the cloud for business intelligence
  - Start with a few hundred GB of data, scale to petabyte or more

Autoscaling

- Review: EC2 connects to one database that is duplicated to a second database (redundancy).
- No redundancy on the EC2 itself. Autoscaling group will fix this.
- You can set up how many instances you want with Autoscaling. When one fails, it will automatically create a new one
- You can set a startup script to run when each new instance starts

Route 53

- Amazon's DNS service
- Domain registration

Elastic Beanstalk

- Allows you to deploy everything (provisions everything like EC2 and RDS and everything else) all at one button
- Creates load balancers, auto-scaling groups, security groups, etc.
- Provisioning EC2 instances, installs PHP

CloudFormation

- Way of scripting out infrastructure
- Turning infrastructure into code
- Codify creating EC2 instances, security groups, etc
- JSON that describes your cloud environment – this is a template
- Elastic Beanstack and CloudFormation are free, but you pay for the resources that are provisioned as a result of using EB and CF

Architecting for the Cloud – Best Practices

Why Cloud Computing?

- IT assets becoming programmable resources
- Global availability and unlimited capacity
- High-level managed services, incl call center functionality, text to voice, machine learning, etc
- Security built in (AWS manages security)

Design Principles – Scalability

1. Scale Up – Start with a small virtual machine and increase size
2. Scale Out – Start with an elastic load balancer, add more virtual machines as your project gets bigger
    1. Stateless Applications – Lambda (no state is stored)
    2. Stateless Components – Instead of storing state on server, it stores state on cookies on user's browser
    3. Stateful Components – Can store some stuff with databases that can scale with you (add replicas or increase size)
    4. Distributed Processing – Break your data into pieces and have EC2 instances work on them separately in parallel (Elastic MapReduce)

Design Priciples – Disposable Resources

- Treat your services like cattle, not pets
- If a server dies, just replace with another one
    1. Bootstrapping – Scripts allow you to set up an instance automatically, setup Apache
    2. Golden images – Take an Amazon Machine Image (AMI) and use it for autoscaling
    3. Hybrid of the two

Design Principles – Infrastructure as Code

- Cloudformation
- Allows you to deploy infrastructure to many clients very easily without manually setting anything up

Design Principles – Automation

- Use alarms, events to automate creation/maintenance of infrastructure
- Loose coupling: Make sure failure in one component doesn't affect other pieces of infrastructure
    - Well defined interfaces: Use RESTful API
    - Service discovery: Don't use fixed IP addresses. Instead use DNS names/endpoints.
    - Asyncronous integration: Messages (actions) remain in queue so if one EC2 goes down, the actions are stored in queues for the next EC2 to pick up
    - Graceful failure: If something breaks, nicely tell the user and report to developers

Design Principles – Serverless not services

- Managed Services (other compnanies like Paypal)
- Serverless Architectures (Lambda, DynamoDB, etc)

Design Principles – Databases

- Relational: Aurora
  - High scalability
  - High availability (6 copies of data at any given time)
  - Data needs joins or complex transactions
- Nonrelational: DynamoDB
  - High scalability
  - High availability
  - Data does not need joins or complex transactions
- Data Warehouse: Red Shift
  - Meant for data for business analysis
  - Red Shift is highly scalability and available
  - Red Shift not meant for online transaction processing (not production database)

Design Principles – Search

- Cloud Search or Elastic Search
- Cloud search: less control, easier
- Elastic search: more control
- Both are very scalability

Design Principles – Misc

- Remove single points of failure, everything should have redundancy
- Detect failure with monitoring (Health checks)
- Durable data storage
  - Don't store all on an EC2 instances
  - Store instead in S3 or Dynamo
- Automate multi-center resilience (multiple Availability Zones)
- Introduce fault isolation and horizontal scaling

Design Principles – Financial

- Optimize for cost
  - Elasticity: More servers when busy, less when not busy with auto saling
  - Purchasing options:
    - Reserviced Capacity
    - Spot Instances

Design Principles – Caching

- Application Caching
- Edge Caching

Design Principles – Security

- Offload security to AWS
- Reduce privileged access
- Treat security as code

Tips

- Understand the basic services:
  - Databases – RDS, DynamoDB, Red Shift
  - Compute – EC2 vs Lambda
  - Storage – S3 (great for static hosting)

Summary of Cloud Concepts and Tech Summary

General

1. 6 Advantages of Cloud
2. 3 Types of Cloud Computing
   1. Infrastructure as a Service (IAAS) – Lightsail
   2. Platform as a Service (PAAS)

    3. Software as a Service (SAAS)
3. 3 Types of Cloud Computing Deployment
    1. Public Cloud (AWS, Azure, Google Cloud)
    2. Hybrid (mix)
    3. Private Cloud (managed locally)
4. Difference between:
    1. Regions – London, Frankfurt, N. Virginia
    2. Availability Zones – Collections of data centers, geographically distributed
    3. Edge Locations – Caching
5. Access AWS Console by:
    1. Via AWS console
    2. Programatically using command line
    3. SDKs
6. Root account has full admin, never give out. Create user for each individuals and secure with multi-factor auth.
7. Groups are places to store users
8. Set permissions in group with policies with JSON

S3

1. S3 bucket is a place to store objects
2. S3 unique namespace
3. Object based only, 200 status code when complete
4. Storage places:
    1. S3 – Current data
    2. Glacier – Archival (2-5 hour retrieval)
5. Restrict access with bucket policy
6. Restrict access to indiv objects with access control lists
7. S3 transfer acceleration – Upload to edge locations. Edge locations then send to central place.
8. Cross-region replication – Replicate to other buckets
9. S3 hosts static websites
10. Scales automatically to meet demand (movie preview)

Cloudfront

1. Edge Location: Location where content is cached
2. Origin: Origin of files that CDN will distribute (S3, EC2, Elastic Load Balancer, Route53)
3. Distribution: Name given to CDN, consists of edge locations
    1. Web – Websites
    2. RTP – Media streaming
4. Can write to edge locations (S3 transfer acceleration)

EC2

- NOT SERVERLESS, compute-based
- Private key to connect
- Security Groups: Virtual firewalls in the cloud, open ports to use
- Design for failure, have one EC2 instance in each Avail Zone
- Pricing models
- Types of EC2 depending on the purpose of EC2
- EBS: Elastic block storage where you install operating system and file
- 4 kinds of EBS:
    - General Purpose SSD
    - Provisioned IOPS SSD
    - Throughput Optimized HDD
    - Cold HDD
- Roles much more security and easier to managethan using access and secret access keys
- Roles are universal, no need to specify users

RDS

1. Multi Avail Zone: Disaster Recovery
2. Read replicas: Scaling out or performance
3. DynamoDB for nonrelational, Aurora for relational, Red Shift for data warehousing

Billing

- Philosophy on pricing: Pay for what you use, start or stop using product at any time. No long-term contracts required.
- Free Tier to help new AWS users get started

Pricing policies

- **Pay as you go: **EC2 used to be pay by hour, pay by second as it's used
- **Pay less when you reserve:** If you reserve time ahead of time, you get a discount
- **Pay even less by unit when using more: **If you use more, you pay less per GB
- **Pay even less as AWS grows**
- Custom pricing for enterprise

What's free?

1. Amazon VPC
2. Elastic Beanstalk (services it provisions are not free)
3. CloudFormation (services it provisions not free)
4. Identity Access Management (IAM)
5. Auto Scaling (EC2 instances it uses are not free)
6. Opsworks
7. Consolidated Billing (add all AWS accounts into one bill)

3 Fundamental Charges

1. Compute
2. Storage
3. Data Out to Internet (Data In is free)

What determines price?

1. Clock hours of server time (time server is running)
2. Machine configuration (more resources consumed = more paid)
3. Machine purchase type (some instance types cost more)
4. Number of instances
5. Load balancing
6. Detailed monitoring (monitor EC2 by minute instead of 5-min intervals)
7. Auto scaling (EC2 instances cost money)
8. Elastic IP Addresses
9. Operating systems (Windows) and software packages

- Elastic Compute Cloud can reserve instances ahead of time, even cheaper if you pay upfront

S3 – What determines price?

1. Storage class (Standard or IA)
2. Storage amount
3. Number of requests
4. Data transfer (data transfer out)

RDS – What determines price?

1. Number of hours RDS is running
2. Database characteristics (licensed?)
3. Database purchase type (huge, nano?)
4. Number of instances
5. Provisioned storage (how big?)
6. Requests made to database
7. Deployment type (multi A-Z, read replicas)
8. Data transfer out

Cloudfront – What determines price?

1. Traffic distribution
2. Requests
3. Data transfers out

Billing: Support Plans

1. Basic
    1. Free, no tech acct mgt, no open cases
2. Developer
    1. $29/mo, business hr access via email, no TAM, 1 person can open unlim cases
    2. General guidance: < 24 business hours
    3. System impaired: < 12 business hours
3. Business
    1. $100/mo, 24×7 email, chat, and phone support, no TAM, unlimited cases for support
    2. General guidance: < 24 business hours
    3. System impaired: < 12 hours
    4. Prod system impaired: < 4 business hours
    5. Prod system down: < 1 hour
4. Enterprise
    1. $15,000/mo, 24×7 email chat and phone, TAM, unlimited cases for support
    2. General guidance: < 24 business hours
    3. System impaired: < 12 hours
    4. Prod system impaired: < 4 business hours
    5. Prod system down: < 1 hour
    6. Business critical down: < 15 mins

- Pricing can be higher if you use AWS a lot

Billing: Resource Groups

- Tags are key-value pairs attached to resources
- Tags can be inherited (created by one service, moves to another service)
- **Resource groups:** Make it easy to group resources based on tags assigned to them
- Resource groups contain info like:
    - Region
    - Name
    - Healthchecks
    - EC2 – Public/Private IP Addresses
    - ELB – Port Configs
    - RDS – Database Engine
- You can search for resources by a specific tag (used by a particular department, user ID, etc)
- Tag Editor allows you to find resources not tagged and add tags

Billing: Consolidated Billing

- **AWS Organization:** Enables you to consolidate multiple AWS accounts into an organization that you create and centrally manage
- **Consolidated billing:** One monthly bill (paying account) for all linked accounts in organization
- 20 linked accounts for consolidating billing
- Easy to track charges and allocate costs
- Volume pricing
- You can also reserve EC2 instances and if one group isn't using them, you can carry them over to another group to save money
- Best practices:
    - Always enable multi factor auth
    - Strong and complex factor
    - Restrict root access
- Billing alerts

Exam Tips

- Consolidated billing allows you to get volume discounts for all your accounts
- Unsused reserved instances for EC2 are applied across group
- CloudTrail is on per-account and per-region basis , can be aggregated into single bucket in paying account

AWS Quick Starts

- Allow you to enable a particular type of technology very quickly
- Templates to get you started with a server that runs a particular technology
- Uses CloudFormation based on a template URL

AWS Cost Calculators

Simple Monthly Calculator

- Allows you to quickly add the resources you're going to use and the types of resources and it'll tell you the cost of each and total monthly bill
- Not comparing what you have on premise and in cloud

Total Cost of Ownership Calculator

- Compares against your current costs for total cost of ownership
- Takes into account:
  - Server costs (hardware & software)
  - Storage costs (hardware & storage admin)
  - Networking costs (network hardware & network admin)
  - IT labor costs

Billing & Pricing Summary

- Remember the free services!
- AWS Support Plans and features of each
- What are tags?
- What are resource groups? Group resources based on tags
- What is the benefit of consolidated billing?
- What's the benefit of AWS Quick Starts?
- Two different AWS calculators

AWS Compliance

Certifications / Attestations

AWS certified with:

1. ISO 27001
2. PCI DSS Level 1
3. SOC 1
4. SOC 2
5. SOC 3

Laws, Regulations, Privacy

1. HIPAA compliant – Meets standards to store health information

Alignments / Frameworks

1. G-Cloud (UK) – Frameworks for government customers to meet these requirements in UK

Shared Responsibility Model

- AWS manages security of cloud, security in cloud itself is responsibilty of customer. Customers are responsibility for security of how AWS is set up. AWS is responsible for the infrastructure
- Do you have the ability to stop something from happening? If you dont have the ability to stop it, it's Amazon's responsibility
- You have control over encryption, customer data

AWS Web Application Firewall and AWS Shield

AWS WAF

- Application firewall that helps protect your web apps from common web exploits that could affect availability, compromise security, or consume excessive resources
- AWF can read data hacker is sending and can intervene on your behalf
- Prevents common attacks
- Goes down to Layer 7

AWS Shield

- Managed DDOS service
- Provides safeguards for web apps running on AWS Two tiers:

1. Standard – Free, avail automatically
2. Advanced – Advanced protection for $3000/mo

AWS Inspector vs AWS Trusted Advisor

AWS Inspector

- Automated security assessment service
- Automatically asses apps for vulnerabilities or deviations from best practices
- Assessment done, provides detailed list of security findings prioritized by leve of severity

AWS Trusted Advisor

- Optimizes AWS environment to reduce cost, increase performance and improve security

1. Cost Optimization (do you have an EC2 with nothing happening on it or an empty DB?)
2. Performance
3. Security
4. Fault Tolerance (are you using multiple avail zones?)

- Two options:

1. Core checks and recommendations 2 Full trusted advisor – business/enterprise only

Security Summary

- Name some of the compliance that AWS meets (above)
- Define what shared responsibility means
- AWS WAF reads data and blocks traffic if it will cause problems
- AWS shield blocks DDOS attacks. Two tiers: Standard and Advanced
- Inspector looks for vulnerabilies on your EC2 instances.
- Advisor gives suggestions for improvement, advanced one requires business subscription

Support Plans

At AWS, we want you to be successful. Our Support plans are designed to give you the right mix of tools and access to expertise so that you can be successful with AWS while optimizing performance, managing risk, and keeping costs under control.

Basic Support is included for all AWS customers and includes:

- Customer Service and Communities - 24x7 access to customer service, documentation, whitepapers, and support forums.
- AWS Trusted Advisor - Access to the 7 core Trusted Advisor checks and guidance to provision your resources following best practices to increase performance and improve security.

- AWS Personal Health Dashboard - A personalized view of the health of AWS services, and alerts when your resources are impacted.

Developer

	Developer	Business	Enterprise
	Recommended if you are experimenting or testing in AWS.	*Recommended if you have production workloads in AWS.*	*Recommended if you have business and/or mission critical workloads in AWS.*
AWS Trusted Advisor Best Practice Checks	7 Core checks	Full set of checks	Full set of checks
Enhanced Technical Support	Business hours** email access to Cloud Support Associates Unlimited cases / 1 primary contact	24x7 phone, email, and chat access to Cloud Support Engineers Unlimited cases / unlimited contacts (IAM supported)	24x7 phone, email, and chat access to Cloud Support Engineers Unlimited cases / unlimited contacts (IAM supported)
Case Severity / Response Times*	General guidance: < 24 hours** System impaired: < 12 hours**	General guidance: < 24 hours System impaired: < 12 hours Production system impaired: < 4 hours Production system down: < 1 hour	General guidance: < 24 hours System impaired: < 12 hours Production system impaired: < 4 hours Production system down: < 1 hour Business-critical system down: < 15 minutes
Architectural Guidance	General	Contextual to your use-cases	Consultative review and guidance based on your applications
Programmatic Case Management		AWS Support API	AWS Support API
Third-Party Software Support		Interoperability and configuration guidance and troubleshooting	Interoperability and configuration guidance and troubleshooting
Proactive Programs and Services		Access to Infrastructure Event Management for additional fee	Infrastructure Event Management Well-Architected Reviews Access to proactive reviews, workshops, and deep dives
Technical Account Management			Designated Technical Account Manager (TAM) to proactively monitor your environment and

			assist with optimization and coordinate access to programs and AWS experts
Training			Access to online self-paced labs
Account Assistance			Concierge Support Team
Pricing	Greater of $29 / month*** - or - 3% of monthly AWS usage See pricing detail and example.	Greater of $100 / month*** - or - 10% of monthly AWS usage for the first $0–$10K 7% of monthly AWS usage from $10K–$80K 5% of monthly AWS usage from $80K–$250K 3% of monthly AWS usage over $250K See pricing detail and example.	Greater of $15,000 - or - 10% of monthly AWS usage for the first $0–$150K 7% of monthly AWS usage from $150K–$500K 5% of monthly AWS usage from $500K–$1M 3% of monthly AWS usage over $1M See pricing detail and example.

Developer

Greater of $29.00

- or -

3% of monthly AWS charges

Business

Greater of $100.00

- or -

10% of monthly AWS charges for the first $0--$10K

7% of monthly AWS charges from $10K--$80K

5% of monthly AWS charges from $80K--$250K

3% of monthly AWS charges over $250K

Enterprise

Greater of $15,000.00

- or -

10% of monthly AWS charges for the first $0--$150K

7% of monthly AWS charges from $150K--$500K

5% of monthly AWS charges from $500K--$1M

3% of monthly AWS charges over $1M