AWS - Key Concepts

Table of Contents

- Audit / Advise
  - Artifact
  - Config
  - CloudTrail
  - Compute Optimizer
- Account Management
  - IAM Identity Center
  - Secrets Manager
  - Control Tower
  - Resource Access Manager
  - Trusted Advisor
  - Organizations
- AI / ML
  - Rekognition
  - Textract
  - Comprehend
  - Macie
  - Lex
  - EKS
- Automation
  - Batch
  - Lambda
    - Lambda function URLs
    - Possible issues
- Backup
  - AWS Backup
  - Disaster Recovery
- Caching
  - ElastiCache
- Centralization
  - Systems Manager
- Containerized
  - Prometheus
  - Proton
  - Fargate
  - ECS
- Computing
  - EC2
    - ASG
      - Types of scaling
        - Step Scaling
        - Target tracking scaling
      - Cooldown Period
      - Which instance does ASG terminate first?
      - What is the difference between Horizontal scaling and Vertical scaling?
    - Type of EC2
      - Bastion Host
      - On-Demand
      - Reserved instance
      - Spot Instances
      - Spot Fleet
      - Reserved Instances (RIs)
    - Networking
      - ENA
  - Fanout pattern
- Communication
  - Simple Workflow Service (Amazon SWF)
  - Pinpoint

- SQS
  - Long vs Short polling - ReceiveMessageWaitTimeSeconds
  - Message Deletion
  - Visibility Timeout
- MQ
- SNS

Data
- Big Data
  - EMR
- Data backup
  - Data Lifecycle Manager
- Database
  - DynamoDB
    - DynomoDB Streams
  - Neptune
  - Redshift
  - Aurora
    - Aurora Global Database
    - Failover behavior
  - RDS
    - Amazon RDS Multi-AZ deployments
    - IAM DB Authentication
    - Enhanced Monitoring
    - What would happen to RDS if the primary database instance fails?
- Data Transform
  - AppSync
  - Glue
- Data Analytics
  - Athena
- Data Stream
  - Amazon Kinesis
    - Kinesis Data Streams (KDS)
- Data Transfer
  - AWS Transfer Family
  - Amazon FSx File System Types
    - Amazon FSx for NetApp ONTAP
    - Amazon FSx for Windows File Server
- DataSync
    - On-premises storage transfers
    - AWS storage transfers
  - Storage Gateway
  - S3 File Gateway
  - Snowball Edge

Distribution
- Local Zones
- CloudFront
  - What could be the possible cause if the requests are hitting the origin server instead of the AWS Edge location?

Deployment
- CloudFormation
- Elastic Beanstalk
  - Where does it store the application files and server log files?

High Performance Computing (HPC)
- Lustre

Security
- Security Token Service
- Guarduty
- Encryption
  - Key Management Service (KMS)
- AWS Shield
  - AWS Shield Advanced
- CloudHSM
- Inspector
- WAF
  - Preventing SQL injection
  - Mitigate DDoS Attack

- Storage
  - S3
    - Types of Tiers
      - Intelligent Tiering
    - Cross-origin resource sharing (CORS)
    - S3 Object Lock
      - Legal Hold vs. Retention Period
  - EFS
  - EBS
  - Encryption
  - Lake Formation
- Key Management
  - KMS
- General
  - OLTP (online transaction processing)
- Migration
  - AWS Directory Service
  - Application Migration Service
  - Database Migration Service (DMS)
  - Application Discovery Service
- Monitoring
  - CloudWatch
    - Custom metric
    - CloudWatch Logs
  - EventBridge
- Networking
  - Direct Connect
  - Network Load Balancer
  - VPC
    - VPC endpoint
    - VPC Peering Connection
    - Inter-Region VPC Peering
    - Transit Gateway
    - managed prefix
    - gateway endpoint
    - Customer Gateway
  - NAT
  - Route 53
    - Latency Routing
    - Geoproximity Routing
    - Geolocation Routing
    - Weighted Routing
    - Geolocation vs Geoproximity Routing
  - Global Accelerator
- Miscellaneous
  - AppSync pipeline resolvers
  - Systems Manager Run Command
  - Auto Scaling

Audit / Advise

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).**

Config

AWS Config provides a detailed view of the configuration of AWS resources in your AWS account. This includes how the resources are related to one another and how they were configured in the past so that you can see how the configurations and relationships change over time.

- **AWS Config is a service that enables you to assess, audit, and evaluate the configurations of your AWS resources.** Config continuously monitors and records your AWS resource configurations and allows you to automate the evaluation of recorded configurations against desired configurations. With Config, you can review changes in configurations and relationships between AWS resources, dive into detailed resource configuration histories, and determine your overall compliance against the configurations specified in your internal guidelines. This enables you to simplify compliance auditing, security analysis, change management, and operational troubleshooting.

An AWS resource is an entity you can work with in AWS, such as an Amazon Elastic Compute Cloud (EC2) instance, an Amazon Elastic Block Store (EBS) volume, a security group, or an Amazon Virtual Private Cloud (VPC). For a complete list of AWS resources supported by AWS Config, see Supported Resource Types for AWS Config.

Whitepaper Link: What Is AWS Config? - AWS Config

CloudTrail

WS **CloudTrail** is an AWS service that helps you enable operational and risk auditing, governance, and compliance of your AWS account. Actions taken by a user, role, or an AWS service are recorded as events in CloudTrail. Events include actions taken in the AWS Management Console, AWS Command Line Interface, and AWS SDKs and APIs.

- **CloudTrail is active in your AWS account when you create it.** When activity occurs in your AWS account, that activity is recorded in a CloudTrail event.

1. Event history
2. CloudTrail Lake
3. Trails

- **AWS CloudTrail Lake** lets you run SQL-based queries on your event logs in AWS CloudTrail. It provides a robust and efficient way to directly analyze CloudTrail logs.

Whitepaper Link: What Is AWS CloudTrail? - AWS CloudTrail

Compute Optimizer

AWS **Compute Optimizer** recommends optimal AWS resources for your workloads to reduce costs and improve performance by using machine learning to analyze historical utilization metrics. Overprovisioning resources can lead to unnecessary infrastructure costs, and underprovisioning resources can lead to poor application performance. Compute Optimizer generates recommendations for the following resources:

- Amazon Elastic Compute Cloud (Amazon EC2) instances
- Amazon EC2 Auto Scaling groups
- Amazon Elastic Block Store (Amazon EBS) volumes
- AWS Lambda functions

Account Management

IAM Identity Center

AWS **IAM Identity Center** (successor to AWS Single Sign-On) **provides single sign-on access for all of your AWS accounts and cloud applications.** It connects with Microsoft Active Directory through AWS Directory Service to allow users in that directory to sign in to a personalized AWS access portal using their existing Active Directory user names and passwords. From the AWS access portal, users have access to all the AWS accounts and cloud applications that they have permission for.

Secrets Manager

AWS Secrets Manager is an AWS service that makes it easier for you to manage secrets. Secrets can be database credentials, passwords, third-party API keys, and even arbitrary text. You can store and control access to these secrets centrally by using the Secrets Manager console, the Secrets Manager command line interface (CLI), or the Secrets Manager API and SDKs.

Control Tower

AWS Control Tower provides a single location to easily set up your new well-architected multi-account environment and govern your AWS workloads with rules for security, operations, and internal compliance. You can automate the setup of your AWS environment with best-practices blueprints for multi-account structure, identity, access management, and account provisioning workflow. For ongoing governance, you can select and apply pre-packaged policies enterprise-wide or to specific groups of accounts.

Resource Access Manager

AWS **Resource Access Manager (RAM)** is a service that enables you to easily and securely share AWS resources with any AWS account or within your AWS Organization. You can share AWS Transit Gateways, Subnets, AWS License Manager configurations, and Amazon Route 53 Resolver rules resources with RAM.

- Many organizations use multiple accounts to create administrative or billing isolation, and limit the impact of errors. RAM eliminates the need to create duplicate resources in multiple accounts, reducing the operational overhead of managing those resources in every single account you own.
- You can create resources centrally in a multi-account environment, and use RAM to share those resources across accounts in three simple steps:

1. Create a Resource Share
2. Specify resources
3. Specify accounts.

RAM is available to you at no additional charge.

Trusted Advisor

Trusted Advisor draws upon best practices learned from serving hundreds of thousands of AWS customers. Trusted Advisor inspects your AWS environment, and then makes recommendations when opportunities exist to save money, improve system availability and performance, or help close security gaps.

- Trusted Advisor inspects your AWS environment and then makes recommendations when opportunities exist to save money, improve system availability and performance, or help close security gaps.
- If you have a **Basic or Developer Support plan**, you can use the Trusted Advisor console to access all checks in the Service Limits category and six checks in the Security category.

Organizations

AWS Organizations helps you centrally manage and govern your environment as you grow and scale your AWS resources. Using Organizations, you can create accounts and allocate resources, group accounts to organize your workflows, apply policies for governance, and simplify billing by using a single payment method for all of your accounts.

The following diagram shows a **high-level explanation** of how you can use AWS Organizations:

- Add accounts
- Group accounts
- Apply policies
- Enable AWS services.

AI / ML

Rekognition

Amazon **Rekognition** can help you streamline or automate image and video moderation workflows using machine learning. **Using fully managed image and video moderation APIs, you can proactively detect inappropriate, unwanted, or offensive content containing nudity, suggestiveness, violence, and other such categories.**

Textract

Amazon Textract is a machine learning (ML) service that automatically extracts text, handwriting, layout elements, and data from scanned documents.

Comprehend

Amazon **Comprehend** uses natural language processing (NLP) to extract insights about the content of documents without the need of any special preprocessing. Amazon Comprehend processes any text files in UTF-8 format. **It develops insights by recognizing the entities, key phrases, language, sentiments, and other common elements in a document.** Use Amazon Comprehend to create new products based on understanding the structure of documents. With Amazon Comprehend you can search social networking feeds for mentions of products, scan an entire document repository for key phrases, or determine the topics contained in a set of documents.

- 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.

Macie

Amazon Macie is a data security service that discovers sensitive data by using machine learning and pattern matching, provides visibility into data security risks, and enables automated protection against those risks.

- To help you manage the security posture of your organization's Amazon Simple Storage Service (Amazon S3) data estate, Macie provides you with an inventory of your S3 general purpose buckets, and automatically evaluates and monitors the buckets for security and access control. **If Macie detects a potential issue with the security or privacy of your data, such as a bucket that becomes publicly accessible, Macie generates a finding for you to review and remediate as necessary.**
- With Amazon Macie, **users can quickly identify and categorize sensitive data, such as personally identifiable information (PII),** financial information, and intellectual property.

Lex

Amazon Lex is a service that allows developers to build conversational interfaces for applications using voice and text. It uses artificial intelligence (AI) and natural language processing (NLP) to create chatbots and voice assistants.

- Amazon Lex enables you to build applications using a speech or text interface powered by the same technology that powers Amazon Alexa. Amazon Lex provides deep functionality and flexibility in natural language understanding (NLU) and automatic speech recognition (ASR), so you can build highly engaging user experiences with lifelike conversational interactions and create new categories of products.
- Amazon Lex enables any developer to build conversational chatbots quickly. With Amazon Lex, no deep learning expertise is necessary—to create a bot, you just specify the basic conversation flow in the Amazon Lex console. The console provides a graphical user interface that you can use to build a production-ready bot for your application.

EKS

Amazon **Elastic Kubernetes Service (Amazon EKS)** is a managed Kubernetes service that eliminates the need to operate and maintain the availability and scalability of Kubernetes clusters in Amazon Web Services (AWS) and in your own data centers. Kubernetes is an open source system that automates the management, scaling, and deployment of containerized applications. To get started, see the Quickstart: Deploy a web app and store data page in the Amazon EKS User Guide.

- **Autoscaling** is a function that automatically scales your resources up or down to meet changing demands. This is a major Kubernetes function that would otherwise require extensive human resources to perform manually.
- Amazon EKS supports two autoscaling products:
  - Karpenter
  - Cluster Autoscaler

Automation

Batch

AWS Batch is a powerful tool for developers, scientists, and engineers who need to run a large number of batch and ML computing jobs. By optimizing compute resources, AWS Batch enables you to focus on analyzing outcomes and resolving issues, rather than worrying about the technical details of running jobs.

- With AWS Batch, you can define and submit multiple simulation jobs to be executed concurrently. AWS Batch will take care of distributing the workload across multiple EC2 instances, scaling up or down based on the demand, and managing the execution environment. **It provides an easy-to-use interface and automation for managing the simulations**, allowing you to focus on the software itself rather than the underlying infrastructure.

Lambda

You can use AWS Lambda to run code without provisioning or managing servers.

- Lambda runs your code on a high-availability compute infrastructure and performs all of the administration of the compute resources, including server and operating system maintenance, capacity provisioning and automatic scaling, and logging. With Lambda, all you need to do is supply your code in one of the language runtimes that Lambda supports.
- **AWS Lambda lets you run code without provisioning or managing servers.** You pay only for the compute time you consume. With Lambda, you can run code for virtually any type of application or backend service – all with zero administration. **Just upload your code, and Lambda takes care of everything required to run and scale your code with high availability.** You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

Lambda function URLs

Lambda function URLs are HTTP(S) endpoints dedicated to your Lambda function. You can easily create and set up a function URL using the Lambda console or API. Once created, Lambda generates a unique URL endpoint for your use.

Possible issues

- **Timeout:** To prevent your Lambda function from running indefinitely, you specify a timeout. The default timeout is 3 seconds, and **the maximum execution duration per request in AWS Lambda is 900 seconds, which is equivalent to 15 minutes.**

Backup

AWS Backup

AWS Backup is a fully-managed service that makes it easy to centralize and automate data protection across AWS services, in the cloud, and on premises. Using this service, you can configure backup policies and monitor activity for your AWS resources in one place. It allows you to automate and consolidate backup tasks that were previously performed service-by-service, and removes the need to create custom scripts and manual processes. With a few clicks in the AWS Backup console, you can automate your data protection policies and schedules.

- AWS Backup does not govern backups you take in your AWS environment outside of AWS Backup. Therefore, if you want a centralized, end-to-end solution for business and regulatory compliance requirements, start using AWS Backup today.

Whitepaper Link: What is AWS Backup? - AWS Backup

Disaster Recovery

AWS Elastic Disaster Recovery (AWS DRS) provides continuous block-level replication, recovery orchestration, and automated server conversion capabilities. These allow customers to achieve a crash-consistent recovery point objective (RPO) of seconds, and a recovery time objective (RTO) typically ranging between 5–20 minutes.

Caching

ElastiCache

The Amazon **ElastiCache** architecture is based on the concept of deploying one or more cache clusters for your application. After your cache cluster is up and running, the service automates common administrative tasks, such as resource provisioning, failure detection and recovery, and software patching. Amazon ElastiCache provides detailed monitoring metrics associated with your cache nodes, enabling you to diagnose and react to issues very quickly. For example, you can set up thresholds and receive alarms if one of your cache nodes is overloaded with requests.

Centralization

Systems Manager

AWS Systems Manager helps you centrally view, manage, and operate nodes at scale in AWS, on-premises, and multicloud environments. With the launch of an unified console experience, Systems Manager consolidates various tools to help you complete common node tasks across AWS accounts and Regions.

- To use Systems Manager, nodes must be managed, which means SSM Agent is installed on the machine and the agent can communicate with the Systems Manager service.
-

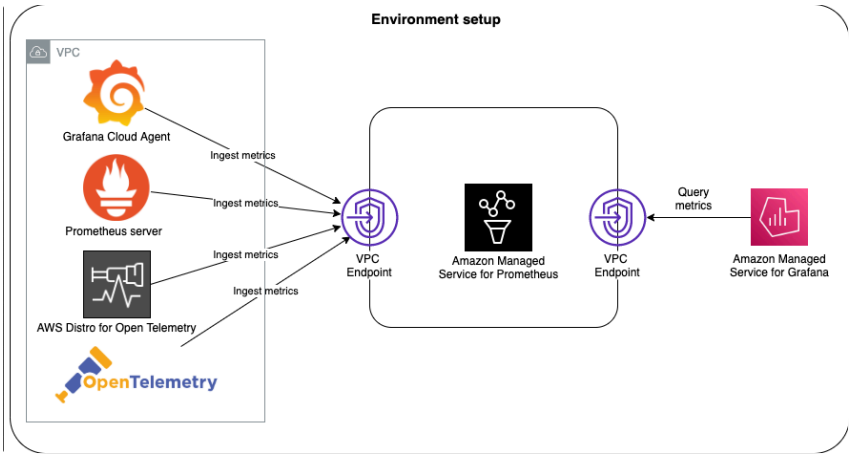
Containerized

Prometheus

Amazon Managed Service for **Prometheus** is a serverless, Prometheus-compatible monitoring service for container metrics that makes it easier to securely monitor container environments at scale. With Amazon Managed Service for Prometheus, you can use the same open-source Prometheus

data model and query language that you use today to monitor the performance of your containerized workloads.

- Monitor containers running on Amazon EC2, Amazon ECS, and Amazon EKS (on Amazon EC2 and on AWS Fargate) in the cloud as well as in hybrid environments. Use it together with **Amazon Managed Grafana** for monitoring, alerts, and dashboard views across all your Kubernetes environments, including both host- and application-level monitoring.



Proton

AWS Proton is:

- **Automated infrastructure as code provisioning and deployment of serverless and container-based applications**

  The AWS Proton service is a two-pronged automation framework.
- **Standardized infrastructure**

  Platform teams can use AWS Proton and versioned infrastructure as code templates.
- **Deployments integrated with CI/CD**

  When developers use the AWS Proton self-service interface to select a *service template*, they're selecting a standardized application stack definition for their code deployments.

Whitepaper Link: What is AWS Proton? - AWS Proton

Fargate

AWS Fargate is a serverless compute engine for containers that works with both Amazon Elastic Container Service (ECS) and Amazon Elastic Kubernetes Service (EKS). Fargate makes it easy for you to focus on building your applications.

ECS

Amazon Elastic Container Service (Amazon ECS) is a fully managed container orchestration service that helps you easily deploy, manage, and scale containerized applications. As a fully managed service, Amazon ECS comes with AWS configuration and operational best practices built-in. It's integrated with both AWS tools, such as Amazon Elastic Container Registry, and third-party tools, such as Docker.

- This integration makes it easier for teams to focus on building the applications, not the environment. You can run and scale your container workloads across AWS Regions in the cloud, and on-premises, without the complexity of managing a control plane.

Whitepaper Link: https://docs.aws.amazon.com/AmazonECS/latest/developerguide/Welcome.html

Computing

EC2

ASG

An **Auto Scaling group** *contains a collection of EC2 instances that are treated as a logical grouping for the purposes of automatic scaling and management.* An Auto Scaling group also lets you use Amazon EC2 Auto Scaling features such as health check replacements and scaling policies. Both maintaining the number of instances in an Auto Scaling group and automatic scaling are the core functionality of the Amazon EC2 Auto Scaling service.

Types of scaling

Step Scaling

- With **step scaling**, you choose scaling metrics and threshold values for the CloudWatch alarms that trigger the scaling process as well as define how your scalable target should be scaled when a threshold is in breach for a specified number of evaluation periods. **Step scaling policies increase or decrease the current capacity of a scalable target based on a set of scaling adjustments, known as step adjustments.** The adjustments vary based on the size of the alarm breach. After a scaling activity is started, the policy continues to respond to additional alarms, even while a scaling activity is in progress. Therefore, all alarms that are breached are evaluated by Application Auto Scaling as it receives the alarm messages.

Target tracking scaling

With a **target tracking scaling** policy, **you can increase or decrease the current capacity of the group based on a target value for a specific metric.** This policy will help resolve the over-provisioning of your resources. The scaling policy adds or removes capacity as required to keep the metric at, or close to, the specified target value. In addition to keeping the metric close to the target value, a target tracking scaling policy also adjusts to changes in the metric due to a changing load pattern.

Cooldown Period

Facts about the cooldown period.

1. It ensures that the Auto Scaling group does not launch or terminate additional EC2 instances before the previous scaling activity takes effect.
2. Its default value is 300 seconds.
3. It is a configurable setting for your Auto Scaling group.

Which instance does ASG terminate first?

ASG terminates instances based on below criteria.

1. Choose the Availability Zone with the most number of instances, which is the us-west-1a Availability Zone in this scenario.
2. Select the instance that is closest to the next billing hour.
3. Select the instances with the oldest launch template.

What is the difference between Horizontal scaling and Vertical scaling?

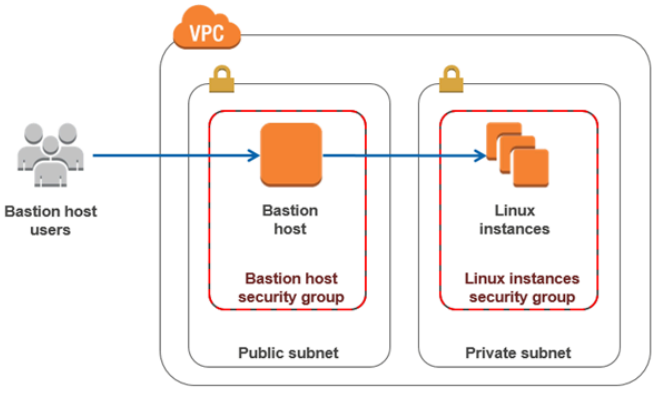
1. Vertical scaling means running the same software on bigger machines which is limited by the capacity of the individual server.
2. Horizontal scaling is adding more servers to the existing pool and doesn't run into limitations of individual servers.

Documentation Link: Auto Scaling groups - Amazon EC2 Auto Scaling

Type of EC2

Bastion Host

A **bastion host** is a **special purpose computer on a network specifically designed and configured to withstand attacks**. If you have a bastion host in AWS, it is basically just an EC2 instance. **It should be in a public subnet with either a public or Elastic IP address with sufficient RDP or SSH access defined in the security group**. Users log on to the bastion host via SSH or RDP and then use that session to manage other hosts in the private subnets.



On-Demand

On-Demand Capacity Reservations enable you to reserve compute capacity for your Amazon EC2 instances in a specific Availability Zone for any duration.

Reserved instance

When you purchase **Reserved Instances**, you make *a one-year or three-year commitment and receive a billing discount of up to 72 percent in return.* When used for the appropriate workloads, Reserved Instances can save you a lot of money.

- **The Reserved Instance Marketplace is a platform that supports the sale of third-party and AWS customers' unused Standard Reserved Instances**, which vary in terms of length and pricing options. For example, you may want to sell Reserved Instances after moving instances to a new AWS region, changing to a new instance type, ending projects before the term expiration, when your business needs change, or if you have unneeded capacity.

Whitepaper Link: Amazon EC2 Reserved Instances - Amazon EC2 Reserved Instances and Other AWS Reservation Models

Spot Instances

- Can get a discount up to 90% compared to On-demand
- Used for batch jobs, data analysis, or workloads that are resilient to failures
- Not great for critical jobs or databases.

Spot Fleet

- Strategies to allocate Spot Instances
  - **lowestPrice**: from the pool with the lowest price (cost optimization, short workload)
  - **Diversified**: distributed across all pools (great for availability, long workloads)
  - **capacityOptimized**: pool with the optimal capacity for the number of instances
  - **priceCapacityOptimized** (recommended): pools with highest capacity available then select the pool with the lowest price (best choice for most workloads)

Reserved Instances (RIs)

Reserved Instances (RIs) provide you with a significant discount (up to 75%) compared to On-Demand instance pricing. You have the flexibility to change families, OS types, and tenancies while benefiting from RI pricing when you use Convertible RIs. One important thing to remember here is that Reserved Instances are not physical instances, but rather a billing discount applied to the use of On-Demand Instances in your account.

- You can modify Standard and Convertible Reserved Instances. Take note that in Convertible Reserved Instances, you are allowed to exchange another Convertible Reserved instance with a different instance type and tenancy.

Networking

ENA

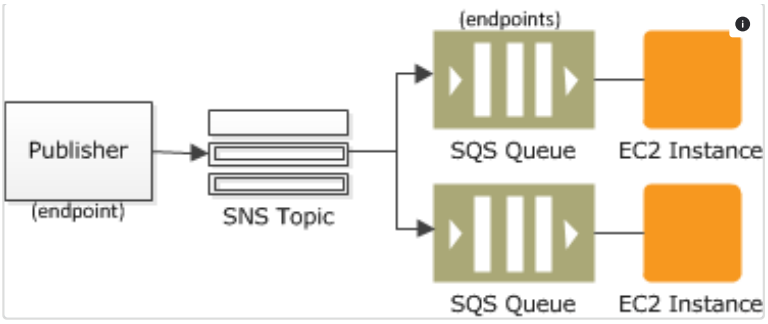
An **Elastic Network Adapter (ENA)** is a virtual network interface provided by Amazon Web Services (AWS) that delivers high-performance networking capabilities to EC2 instances, enabling increased network throughput, low latency, and improved packet per second (PPS) performance compared to standard network adapters; essentially, it's a custom-designed network interface optimized for high-bandwidth cloud computing needs on AWS.

What are differences between EFA and ENA?

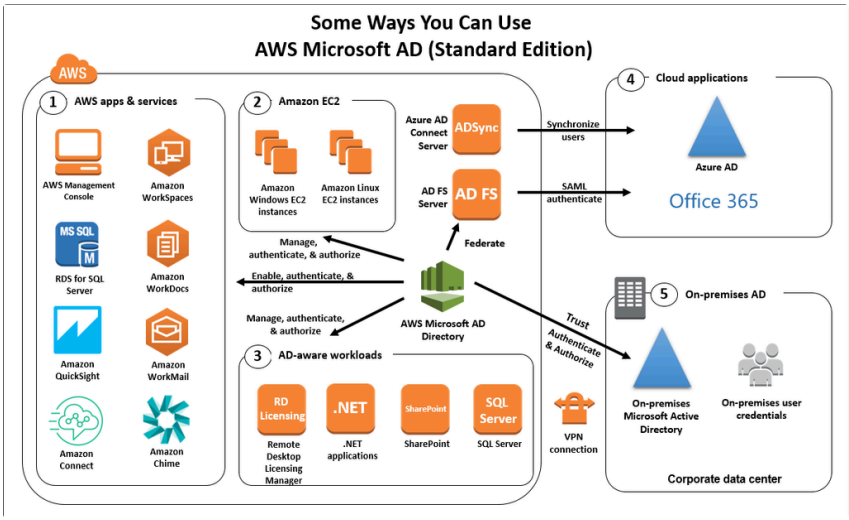
An **Elastic Fabric Adapter (EFA)** is simply an **Elastic Network Adapter (ENA)** with added capabilities. It provides all of the functionality of an ENA, with additional OS-bypass functionality. OS-bypass is an access model that allows HPC and machine learning applications to communicate directly with the network interface hardware to provide low-latency, reliable transport functionality.

Fanout pattern

A "**fanout**" pattern is when an Amazon SNS message is sent to a topic and then replicated and pushed to multiple Amazon SQS queues, HTTP endpoints, or email addresses. This allows for parallel asynchronous processing. For example, you could develop an application that sends an Amazon SNS message to a topic whenever an order is placed for a product. Then, the Amazon SQS queues that are subscribed to that topic would receive identical notifications for the new order. The Amazon EC2 server instance attached to one of the queues could handle the processing or fulfillment of the order, while the other server instance could be attached to a data warehouse for analysis of all orders received.



Microsoft AD



Communication

Simple Workflow Service (Amazon SWF)

The Amazon **Simple Workflow Service (Amazon SWF)** provides a way to build, run, and scale background jobs that have parallel or sequential steps. With Amazon SWF, you can coordinate work across distributed components, tracking the state of tasks.

- In Amazon SWF, a task represents a logical unit of work that is performed by a component of your application. Coordinating tasks across the application involves managing intertask dependencies, scheduling, and concurrency in the logical flow of your application. Amazon SWF gives you control over implementing tasks and coordinating them without worrying about underlying complexities such as tracking their progress and maintaining their state.

Pinpoint

In Amazon Pinpoint, an event is an action that occurs when a user interacts with one of your applications, when you send a message from a campaign or journey, or when you send a transactional SMS or email message. For example, if you send an email message, several events occur:

- When you send the message, a send event occurs.
- When the message reaches the recipient's inbox, a delivered event occurs.
- When the recipient opens the message, an open event occurs.

SQS

Amazon Simple Queue Service (Amazon SQS) is a fully managed message queuing service that makes it easy to decouple and scale microservices, distributed systems, and serverless applications. Amazon SQS moves data between distributed application components and helps you decouple these components.

*What are differences between a **long polling** and a **short polling**?*

In Amazon SQS, short polling returns a response immediately, while long polling waits for a message to arrive. Long polling can reduce costs and improve message processing efficiency.

Long vs Short polling - ReceiveMessageWaitTimeSeconds

The **ReceiveMessageWaitTimeSeconds** is the queue attribute that determines whether you are using Short or Long polling. By default, its value is zero which means it is using Short polling. If it is set to a value greater than zero, then it is Long polling.

Message Deletion

Amazon SQS automatically deletes messages that have been in a queue for more than the maximum message retention period. The default message retention period is 4 days. Since the queue is configured to the default settings and the batch job application only processes the messages once a week, the messages that are in the queue for more than 4 days are deleted. This is the root cause of the issue.

Visibility Timeout

The **visibility timeout** is a period of time during which Amazon SQS prevents other consuming components from receiving and processing a message.

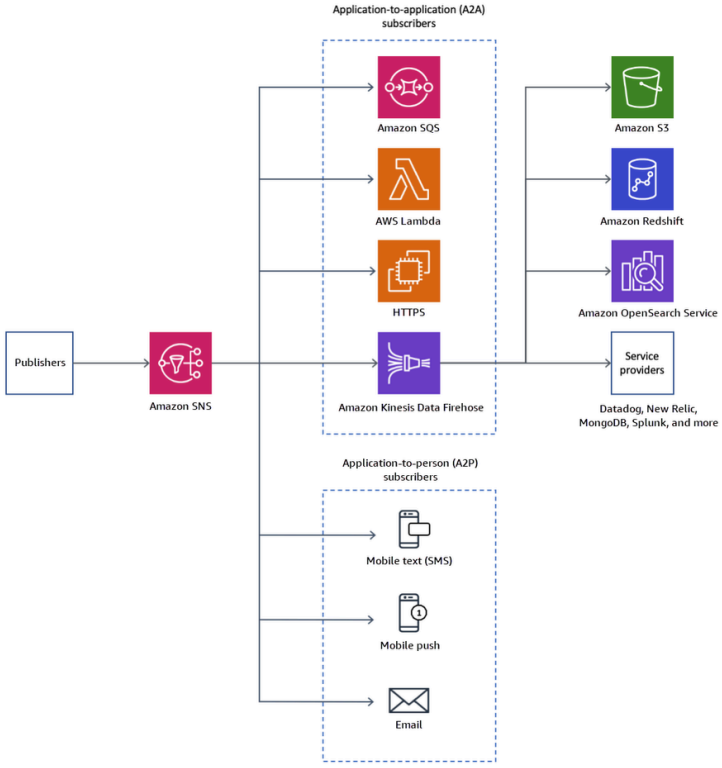
MQ

Amazon MQ is a managed message broker service for Apache ActiveMQ Classic and RabbitMQ that manages the setup, operation, and maintenance of message brokers. You can create a new Amazon MQ broker using industry standard messaging protocols, or migrate existing message brokers to Amazon MQ without rewriting messaging code.

SNS

Amazon Simple Notification Service (Amazon SNS) is a managed service that provides message delivery from publishers to subscribers (also known as producers and consumers). Publishers communicate asynchronously with subscribers by sending messages to a topic, which is a logical access point and communication channel.

- Clients can subscribe to the Amazon SNS topic and receive published messages using a supported endpoint type, such as Amazon Data Firehose, Amazon SQS, AWS Lambda, HTTP, email, mobile push notifications, and mobile text messages (SMS).



Data

Big Data

EMR

Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. By using these frameworks and related open-source projects, such as Apache Hive and Apache Pig, you can process data for analytics purposes and business intelligence workloads. Additionally, you can use Amazon EMR to transform and move large amounts of data into and out of other AWS data stores and databases.

Data backup

Data Lifecycle Manager

You can use Amazon Data Lifecycle Manager to automate the creation, retention, and deletion of EBS snapshots and EBS-backed AMIs. When you automate snapshot and AMI management, it helps you to:

- Protect valuable data by enforcing a regular backup schedule.
- Create standardized AMIs that can be refreshed at regular intervals.
- Retain backups as required by auditors or internal compliance.
- Reduce storage costs by deleting outdated backups.
- Create disaster recovery backup policies that back up data to isolated Regions or accounts.

Database

DynamoDB

Amazon DynamoDB is a serverless, **NoSQL, fully managed database with single-digit millisecond performance at any scale.**

DynomoDB Streams

- **DynamoDB Streams** captures a time-ordered sequence of item-level modifications in any DynamoDB table and stores this information in a log for up to 24 hours. Applications can access this log and view the data items as they appeared before and after they were modified, in near-real-time.

- **DynamoDB addresses your needs to overcome scaling and operational complexities of relational databases.** DynamoDB is purpose-built and optimized for operational workloads that require consistent performance at any scale.
- For example, **DynamoDB delivers consistent single-digit millisecond performance for a shopping cart use case, whether you've 10 or 100 million users.** Launched in 2012, DynamoDB continues to help you move away from relational databases while reducing cost and improving performance at scale.

Neptune

Amazon Neptune is a fully managed graph database service. Neptune makes it easy to build and run applications that work with highly connected datasets, including for ID, graph/C360, security, fraud, and knowledge graph applications. Some key features of Amazon Neptune include:

- **High performance** — Provides low-latency and high-throughput performance for both read and write operations, making it suitable for real-time applications.
- **Scalability** — Neptune can handle billions of vertices and edges, and is designed to automatically scale to meet the demands of your application.
- **Compatibility** — Supports the popular graph query languages, including Apache TinkerPop Gremlin and SPARQL, making it easy to use with existing applications and tools.
- **Durability** — Automatically replicates data across multiple Availability Zones (AZs) for high availability (HA) and data durability.
- **Integration with other AWS services** — Integration with other AWS services such as Amazon S3, Amazon OpenSearch Service, and Amazon SageMaker AI, making it easy to build data-driven applications.
- **Management and monitoring** — Provides an easy-to-use, web-based console for monitoring and managing your database, as well as integration with Amazon CloudWatch for metrics and alerts.

Whitepaper Link: Amazon Neptune - Choosing an AWS NoSQL Database

Redshift

Welcome to the Amazon **Redshift** Management Guide. Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. Amazon Redshift Serverless lets you access and analyze data without all of the configurations of a provisioned data warehouse. Resources are automatically provisioned and data warehouse capacity is intelligently scaled to deliver fast performance for even the most demanding and

unpredictable workloads. You don't incur charges when the data warehouse is idle, so you only pay for what you use. You can load data and start querying right away in the Amazon Redshift query editor v2 or in your favorite business intelligence (BI) tool. Enjoy the best price performance and familiar SQL features in an easy-to-use, zero administration environment.

- Regardless of the size of the dataset, Amazon Redshift offers fast query performance using the same SQL-based tools and business intelligence applications that you use today.
- **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 a high-performance disk.

Whitepaper Link: What is Amazon Redshift? - Amazon Redshift

Aurora

Amazon Aurora (Aurora) is a fully managed relational database engine that's compatible with MySQL and PostgreSQL. You already know how MySQL and PostgreSQL combine the speed and reliability of high-end commercial databases with the simplicity and cost-effectiveness of open-source databases. The code, tools, and applications you use today with your existing MySQL and PostgreSQL databases can be used with Aurora. With some workloads, Aurora can deliver up to five times the throughput of MySQL and up to three times the throughput of PostgreSQL without requiring changes to most of your existing applications.

Aurora Global Database

- Amazon **Aurora Global Database** is designed for globally distributed applications, allowing a single Amazon Aurora database to span multiple AWS regions. It replicates your data with no impact on database performance, enables fast local reads with low latency in each region, and provides disaster recovery from region-wide outages.
- Aurora includes a high-performance storage subsystem. Its MySQL- and PostgreSQL-compatible database engines are customized to take advantage of that fast distributed storage. **The underlying storage grows automatically as needed, up to 64 tebibytes (TiB)**. Aurora also automates and standardizes database clustering and replication, which are typically among the most challenging aspects of database configuration and administration.
- **Aurora includes a high-performance storage subsystem.** Its MySQL- and PostgreSQL-compatible database engines are customized to take advantage of that fast distributed storage.

Failover behavior

Failover is automatically handled by Amazon Aurora so that your applications can resume database operations as quickly as possible without manual administrative intervention.

In the event of system failure on the primary database instance, **Aurora will attempt to create a new DB Instance in the same Availability Zone** as the original instance and is done on a best-effort basis.

- **The underlying storage grows automatically as needed.** An Aurora cluster volume can grow to a maximum size of 128 tebibytes (TiB). Aurora also automates and standardizes database clustering and replication, which are typically among the most challenging aspects of database configuration and administration.
- By using **Aurora cloning**, you can create a new cluster that uses the same Aurora cluster volume and has the same data as the original.
- An Aurora Serverless DB cluster is a DB cluster that automatically starts up, shuts down, and scales up or down its compute capacity based on your application's needs. Aurora Serverless provides a relatively simple, cost-effective option for infrequent, intermittent, sporadic, or unpredictable workloads.
- **Amazon Aurora Serverless v2** suits the most demanding, highly variable workloads. In contrast, Aurora provisioned clusters are suitable for steady workloads.

RDS

Amazon Relational Database Service (Amazon RDS) is an easy-to-manage relational database service optimized for total cost of ownership. It is simple to set up, operate, and scale with demand. Amazon RDS automates undifferentiated database management tasks, such as provisioning, configuring, backing up, and patching.

Amazon RDS Multi-AZ deployments

- **Amazon RDS Multi-AZ deployments** provide enhanced availability and durability for Database (DB) Instances, making them a natural fit for production database workloads. When you provision a Multi-AZ DB Instance, Amazon RDS automatically creates a primary DB Instance and synchronously replicates the data to a standby instance in a different Availability Zone (AZ). Each AZ runs on its own physically distinct, independent infrastructure and is engineered to be highly reliable.

IAM DB Authentication

You can authenticate to your DB instance using AWS Identity and Access Management (IAM) database authentication. IAM database authentication works with MySQL and PostgreSQL. With this authentication method, **you don't need to use a password when you connect to a DB instance.**

Instead, you use an authentication token.

- Amazon RDS allows customers to create a new database in minutes and offers flexibility to customize databases to meet their needs across eight engines and two deployment options. Customers can optimize performance with features like Multi-AZ with two readable standbys, optimized writes and reads, and AWS Graviton3-based instances, and they can choose from multiple pricing options to effectively manage costs.

Enhanced Monitoring

Amazon RDS provides metrics in real time for the operating system (OS) that your DB instance runs on. You can view the metrics for your DB instance using the console, or consume the **Enhanced Monitoring** JSON output from CloudWatch Logs in a monitoring system of your choice.

- By default, Enhanced Monitoring metrics are stored in the CloudWatch Logs for 30 days.
- To modify the amount of time the metrics are stored in the CloudWatch Logs, change the retention for the RDSOSMetrics log group in the CloudWatch console.

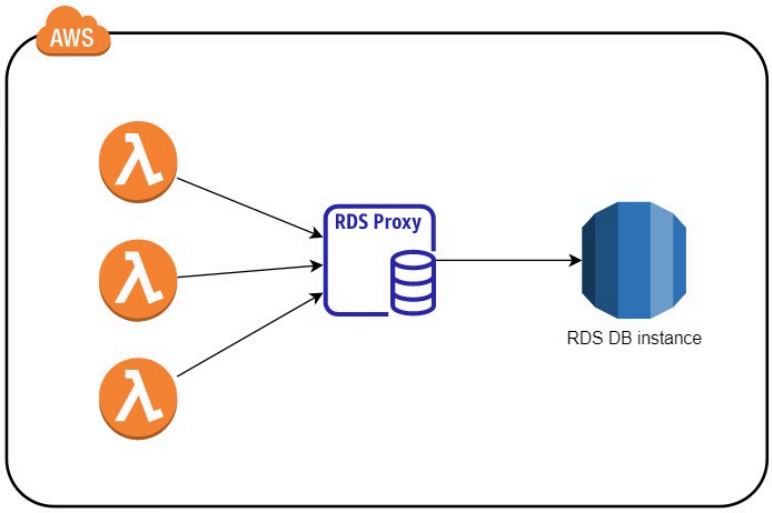
What would happen to RDS if the primary database instance fails?

- In Amazon RDS, **failover is automatically handled** so that you can resume database operations as quickly as possible without administrative intervention in the event that your primary database instance goes down.
  - When failing over, Amazon RDS simply flips the canonical name record (CNAME) for your DB instance to point at the standby, which is in turn promoted to become the new primary.

How do you troubleshoot "too many connections" error?

Answer: Provision an RDS Proxy between the Lambda function and RDS database instance

- If a "Too Many Connections" error happens to a client connecting to a MySQL database, it means all available connections are in use by other clients.
- RDS Proxy helps you manage a large number of connections from Lambda to an RDS database by establishing a warm connection pool to the database.



Data Transform

AppSync

AWS AppSync is a serverless GraphQL and Pub/Sub API service that simplifies building modern web and mobile applications. It provides a robust, scalable GraphQL interface for application developers to combine data from multiple sources, including Amazon DynamoDB, AWS Lambda, and HTTP APIs.

What is Graph API?

GraphQL is a data language to enable client apps to fetch, change and subscribe to data from servers. In a GraphQL query, the client specifies how the data is to be structured when it is returned by the server. This makes it possible for the client to query only for the data it needs, in the format that it needs it in.

Glue

AWS Glue is a serverless data integration service that makes it easy for analytics users to discover, prepare, move, and integrate data from multiple sources. You can use it for analytics, machine learning, and application development. It also includes additional productivity and data ops tooling for authoring, running jobs, and implementing business workflows.

- You can use **AWS Glue crawlers** to automatically infer database and table schema from your data in **Amazon S3** and store the associated metadata in the AWS Glue Data Catalog.
  - With AWS Glue, you can discover and connect to more than 70 diverse data sources and manage your data in a centralized data catalog. You can visually create, run, and monitor extract, transform, and load (ETL) pipelines to load data into your data lakes. Also, you can immediately search and query cataloged data using Amazon Athena, Amazon EMR, and Amazon Redshift Spectrum.
- AWS Glue consolidates major data integration capabilities into a single service. These include data discovery, modern ETL, cleansing, transforming, and centralized cataloging. It's also serverless, which means there's no infrastructure to manage. With flexible support for all workloads like ETL, ELT, and streaming in one service, AWS Glue supports users across various workloads and types of users.
- Also, AWS Glue makes it easy to integrate data across your architecture. It integrates with AWS analytics services and Amazon S3 data lakes. AWS Glue has integration interfaces and job-authoring tools that are easy to use for all users, from developers to business users, with tailored solutions for varied technical skill sets.
-

Data Analytics

Athena

Amazon Athena supports a wide variety of data formats like CSV, TSV, JSON, or Textfiles and also supports open-source columnar formats such as Apache ORC and Apache Parquet. Athena also supports compressed data in Snappy, Zlib, LZO, and GZIP formats. By compressing, partitioning, and using columnar formats you can improve performance and reduce your costs.

- Parquet and ORC file formats both support predicate pushdown (also called predicate filtering). Parquet and ORC both have blocks of data that represent column values. Each block holds statistics for the block, such as max/min values. When a query is being executed, these statistics determine whether the block should be read or skipped.

Data Stream

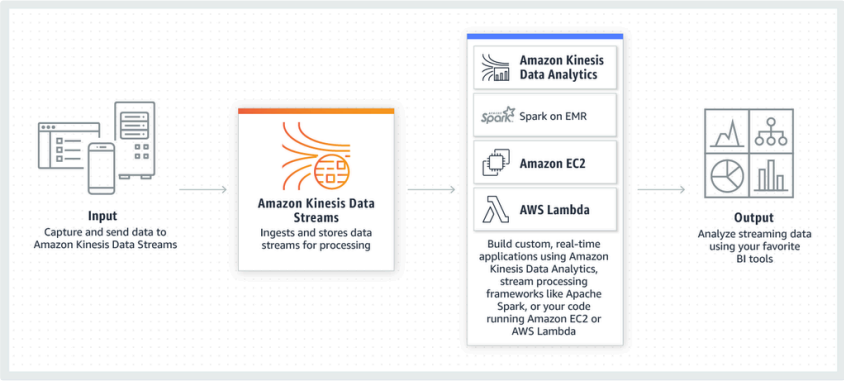
Amazon Kinesis

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. It 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 responds instantly instead of having to wait until all your data are collected before the processing can begin.

Kinesis Data Streams (KDS)

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.



Data Transfer

AWS Transfer Family

AWS Transfer Family is a service that allows for secure file transfer over protocols such as SFTP, FTPS, and FTP directly into and out of Amazon S3 or Amazon EFS. Amazon EFS is well-known for its high IOPS performance, making it good for applications that require quick and simultaneous read/write operations. This is beneficial for scenarios where high-performance shared storage is needed, such as with SFTP. The combination of AWS EFS and AWS Transfer Family provides a serverless solution that meets the needs for high performance, scalability, and security. It allows for the customization of the security framework through network access controls and encryption, ensuring that data is protected both at rest and in transit. Additionally, it enables companies to maintain oversight over user permissions, allowing for secure management of access to the file storage system without the need for traditional server management.

Amazon FSx File System Types

Amazon FSx provides fully managed third-party file systems. Amazon FSx provides you with the native compatibility of third-party file systems with feature sets for workloads such as **Windows-based storage, high-performance computing (HPC), machine learning, and electronic design automation (EDA).** You don't have to worry about managing file servers and storage, as Amazon FSx automates time-consuming administration tasks such as hardware provisioning, software configuration, patching, and backups. Amazon FSx integrates the file systems with cloud-native AWS services, making them even more useful for a broader set of workloads.

Amazon FSx for NetApp ONTAP

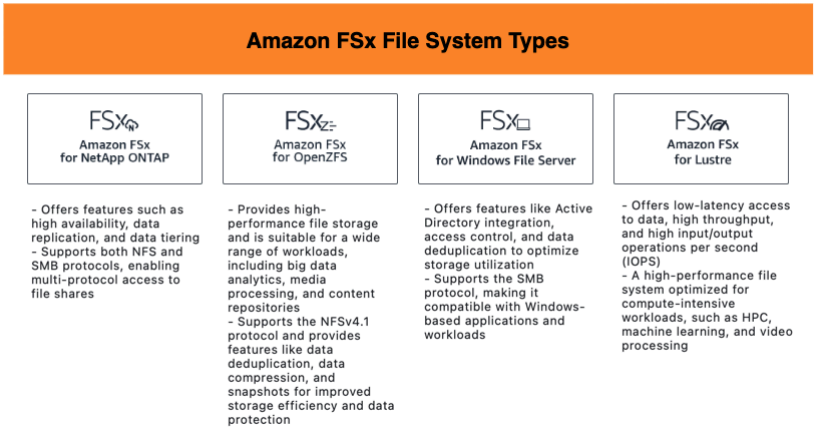
Amazon FSx for NetApp ONTAP is a fully managed service that provides highly reliable, scalable, high-performing, and feature-rich file storage built on NetApp's popular ONTAP file system. FSx for ONTAP combines the familiar features, performance, capabilities, and API operations of NetApp file systems with the agility, scalability, and simplicity of a fully managed AWS service.

- With Amazon FSx for ONTAP, there is multi-protocol access to data using the Network File System (NFS), Server Message Block (SMB), and Internet Small Computer Systems Interface (iSCSI) protocols. It supports highly available and durable Multi-AZ and Single-AZ deployment options.

Amazon FSx for Windows File Server

Finally, the files can be integrated into an **Amazon FSx for Windows File Server** file system, which provides high performance, high availability, automatic backups, and easy restoration capabilities for Windows-based applications.

- The fully managed file storage service is optimized for Windows workloads and supports the Server Message Block (SMB) protocol, ensuring a fully compatible file system.
- With this combination of AWS services, organizations can easily migrate and integrate their files into a secure, reliable, and scalable file storage solution without compromising current file permissions.



DataSync

AWS DataSync is an online data transfer and discovery service that simplifies data migration and helps you quickly, easily, and securely transfer your file or object data to, from, and between AWS storage services.

- AWS DataSync is an online data transfer service that simplifies, automates, and accelerates moving data between on-premises storage systems and AWS Storage services, as well as between AWS Storage services.
- **You can use DataSync to**
  - Migrate active datasets to AWS
  - Archive data to free up on-premises storage capacity

- Replicate data to AWS for business continuity
- Transfer data to the cloud for analysis and processing.

On-premises storage transfers

DataSync works with the following on-premises storage systems:

- Network File System (NFS)
- Server Message Block (SMB)
- Hadoop Distributed File Systems (HDFS)
- Object storage

AWS storage transfers

DataSync works with the following AWS storage services:

- Amazon S3
- Amazon EFS
- Amazon FSx for Windows File Server
- Amazon FSx for Lustre
- Amazon FSx for OpenZFS
- Amazon FSx for NetApp ONTAP
- AWS DataSync agents can be installed on the source and target storage systems. These agents facilitate data transfer between them, ensuring a highly reliable and secure transfer with support for various data transfer protocols, including NFS, SMB, and S3. DataSync tasks can be scheduled to run automatically or initiated manually, providing flexibility and control over the transfer process.

Storage Gateway

AWS Storage Gateway is a hybrid cloud storage service that connects on-premises environments with AWS cloud storage.

- It allows you to seamlessly integrate your existing on-premises infrastructure with AWS, enabling you to store and retrieve data from the cloud and run applications in a hybrid environment.
- For Windows workloads, you can use Storage Gateway to store and access data using native Windows **protocols such as SMB and NFS.** You can use Storage Gateway to reduce costs associated with running Windows workloads on AWS by using on-premises hardware and software as a bridge to the cloud. This enables you to take advantage of the scalability and cost-efficiency of AWS without having to make significant changes to your existing infrastructure.

S3 File Gateway

Amazon S3 File Gateway – Amazon S3 File Gateway supports a file interface into Amazon Simple Storage Service (Amazon S3) and combines a service and a virtual software appliance.

- **By using this combination, you can store and retrieve objects in Amazon S3 using industry-standard file protocols such as Network File System (NFS) and Server Message Block (SMB).** You deploy the gateway into your on-premises environment as a virtual machine (VM) running on VMware ESXi, Microsoft Hyper-V, or Linux Kernel-based Virtual Machine (KVM), or as a hardware appliance that you order from your preferred reseller.
- Y**ou can also deploy the Storage Gateway VM in VMware Cloud on AWS,** or as an AMI in Amazon EC2. The gateway provides access to objects in S3 as files or file share mount points. With a S3 File Gateway, you can do the following:

Snowball Edge

AWS Snowball Edge is a type of Snowball device with on-board storage and compute power for select AWS capabilities. Snowball Edge can process data locally, run edge-computing workloads, and transfer data to or from the AWS Cloud.

- Each Snowball Edge device can transport data at speeds faster than the internet. This transport is done by shipping the data in the devices through a regional carrier. The appliances are rugged, complete with E Ink shipping labels.
- Snowball Edge devices have two options for device configurations—Storage Optimized 210 TB and Compute Optimized. When this guide refers to Snowball Edge devices, it's referring to all options of the device. When specific information applies only to one or more optional configurations of devices, it is called out specifically. For more information, see Snowball Edge device configurations.

Whitepaper Link: What is AWS Snowball Edge? - AWS Snowball Edge Developer Guide

Distribution

Local Zones

AWS Local Zones are a type of AWS infrastructure deployment that places compute, storage, database, and other select services closer to large population, industry, and IT centers, enabling you to deliver applications that require single-digit millisecond latency to end-users.

- Local Zones bring AWS closer or within a customer's geographic boundary in a fully AWS-owned and operated mode and can, therefore, help meet data residency requirements. If an AWS Region is not within the desired geo-political boundary but a Local Zone is available and is able to meet your residency requirements in a geographical area, then choose the Local Zone for your regulatory and compliance needs.

CloudFront

Amazon **CloudFront** is a content delivery network (CDN) service that enables the efficient distribution of web content to users across the globe. It reduces latency by caching static and dynamic content in multiple edge locations worldwide and improves the overall user experience.

- Typically, CloudFront serves an object from an edge location until the cache duration that you specified passes — that is, until the object expires. After it expires, the next time the edge location gets a user request for the object, **CloudFront forwards the request to the origin server to verify that the cache contains the latest version of the object.**
  - The `Cache-Control` and `Expires` headers control how long objects stay in the cache. The `Cache-Control max-age` directive lets you specify how long (in seconds) you want an object to remain in the cache before CloudFront gets the object again from the origin server. **The minimum expiration time CloudFront supports is 0 seconds for web distributions and 3600 seconds for RTMP distributions.**

What could be the possible cause if the requests are hitting the origin server instead of the AWS Edge location?

In this scenario, the main culprit is that the **Cache-Control max-age directive is set to a low value**, which is why the request is always directed to your origin server.

Deployment

CloudFormation

AWS CloudFormation is a service that enables developers to create AWS resources in an orderly and predictable fashion. Resources are written in text files using JSON or YAML format. The templates require a specific syntax and structure that depends on the types of resources being created and managed. You author your resources in JSON or YAML with any code editor such as AWS Cloud9, check it into a version control system, and then CloudFormation builds the specified services in safe, repeatable manner.

- In CloudFormation, a template is a JSON or a YAML-formatted text file that describes your AWS infrastructure. Templates include several major sections.
  - Format Version
  - Description
  - Metadata
  - Parameters
  - Mappings
  - Conditions
  - Transform
  - Resources (required)
  - Outputs

Whitepaper Link: https://docs.aws.amazon.com/whitepapers/latest/introduction-devops-aws/aws-cloudformation.html

Elastic Beanstalk

With **Elastic Beanstalk** you can quickly deploy and manage applications in the AWS Cloud without having to learn about the infrastructure that runs those applications. Amazon Web Services (AWS) comprises over one hundred services, each of which exposes an area of functionality. While the variety of services offers flexibility for how you want to manage your AWS infrastructure, it can be challenging to figure out which services to use and how to provision them. Elastic Beanstalk reduces management complexity without restricting choice or control. You simply upload your application, and Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring.

- Summary: AWS Elastic Beanstalk reduces management complexity without restricting choice or control. You simply upload your application, and Elastic Beanstalk automatically handles the details of capacity provisioning, load balancing, scaling, and application health monitoring.
- Elastic Beanstalk supports applications developed in **Go, Java, .NET, Node.js, PHP, Python, and Ruby.** When you deploy your application, Elastic Beanstalk builds the selected supported platform version and provisions one or more AWS resources, such as Amazon EC2 instances, to run your application.
- AWS Elastic Beanstalk stores your application files and optionally, **server log files in Amazon S3.**

Where does it store the application files and server log files?

Application files are stored in S3. The server log files can also optionally be stored in S3 or in CloudWatch Logs.

Whitepaper Link: https://docs.aws.amazon.com/elasticbeanstalk/latest/dg/Welcome.html

High Performance Computing (HPC)

Lustre

Amazon FSx for Lustre provides a high-performance file system optimized for fast processing of workloads such as machine learning, **high performance computing (HPC)**, video processing, financial modeling, and electronic design automation (EDA). These workloads commonly require data to be presented via a fast and scalable file system interface and typically have data sets stored on long-term data stores like Amazon S3.

Whitepaper Link: https://docs.aws.amazon.com/fsx/latest/LustreGuide/what-is.html

Security

Security Token Service

AWS **Security Token Service (STS)** is the service that you can use to create and provide trusted users with temporary security credentials that can control access to your AWS resources. Temporary security credentials work almost identically to the long-term access key credentials that your IAM users can use.

Guarduty

Amazon **GuardDuty can generate findings based on suspicious activities such as requests coming from known malicious IP addresses**, changing of bucket policies/ACLs to expose an S3 bucket publicly, or suspicious API call patterns that attempt to discover misconfigured bucket permissions.

- To detect possibly malicious behavior, GuardDuty uses a combination of anomaly detection, machine learning, and continuously updated threat intelligence.

Encryption

Key Management Service (KMS)

AWS Key Management Service (AWS KMS) is an encryption and key management service scaled for the cloud. AWS KMS keys and functionality are used by other AWS services, and you can use them to protect data in your own applications that use AWS.

AWS Shield

AWS provides two levels of protection against DDoS attacks: AWS Shield Standard and AWS Shield Advanced. AWS Shield Standard is automatically included at no extra cost beyond what you already pay for AWS WAF and your other AWS services. For added protection against DDoS attacks, AWS offers AWS Shield Advanced. AWS Shield Advanced provides expanded DDoS attack protection for your Amazon EC2 instances, Elastic Load Balancing load balancers, Amazon CloudFront distributions, and Amazon Route 53 hosted zones.

AWS Shield Advanced

AWS Shield Advanced also gives you 24×7 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.

CloudHSM

AWS **CloudHSM** combines the benefits of the AWS cloud with the security of hardware security modules (HSMs).

- A **hardware security module (HSM) is a computing device that processes cryptographic operations** and provides secure storage for cryptographic keys.
- With AWS CloudHSM, you have complete control over high availability HSMs that are in the AWS Cloud, have low-latency access, and **a secure root of trust that automates HSM management (including backups, provisioning, configuration, and maintenance).**

Inspector

Amazon Inspector is a vulnerability management service that automatically discovers workloads and continually scans them for software vulnerabilities and unintended network exposure. Amazon Inspector discovers and scans Amazon EC2 instances, container images in Amazon ECR, and Lambda functions. When Amazon Inspector detects a software vulnerability or unintended network exposure, it creates a finding, which is a detailed report about the issue. You can manage findings in the Amazon Inspector console or API.
Whitepaper Link: What is Amazon Inspector? - Amazon Inspector

WAF

AWS WAF is a web application firewall that lets you monitor the HTTP(S) requests that are forwarded to your protected web application resources. You can protect the following resource types:

- Amazon CloudFront distribution
- Amazon API Gateway REST API
- Application Load Balancer
- AWS AppSync GraphQL API
- Amazon Cognito user pool
- AWS App Runner service
- AWS Verified Access instance

AWS WAF lets you control access to your content. **Based on criteria that you specify, such as the IP addresses that requests originate from** or the values of query strings, the service associated with your protected resource responds to requests either with the requested content, with an HTTP 403 status code (Forbidden), or with a custom response.

Preventing SQL injection

- **AWSManagedRulesSQLiRuleSet** – The SQL database rule group contains rules to block request patterns associated with the exploitation of SQL databases, like SQL injection attacks. This can help prevent remote injection of unauthorized queries. Evaluate this rule group for use if your application interfaces with an SQL database.

Mitigate DDoS Attack

To detect and mitigate DDoS attacks, you can use AWS WAF in addition to AWS Shield. AWS WAF is a web application firewall that helps detect and mitigate web application layer DDoS attacks by inspecting traffic inline. Application layer DDoS attacks use well-formed but malicious requests to evade mitigation and consume application resources. You can define custom security rules that contain a set of conditions, rules, and actions to block attacking traffic. After you define web ACLs, you can apply them to CloudFront distributions, and web ACLs are evaluated in the priority order you specified when you configured them.

Storage

S3

- you can use Multipart upload in S3 to improve the throughput.

Types of Tiers

Intelligent Tiering

Amazon S3 Standard

Among the options given, only **Amazon S3 Standard** has the feature of no minimum storage duration. It is also the most cost-effective storage service because you will only be charged for the last 12 hours, unlike in other storage classes where you will still be charged based on its respective storage duration (e.g. 30 days, 90 days, 180 days).

Amazon S3 Intelligent-Tiering is the only cloud storage class that delivers automatic storage cost savings when data access patterns change, without performance impact or operational overhead. The Amazon S3 Intelligent-Tiering storage class is designed to optimize storage costs by automatically moving data to the most cost-effective access tier when access patterns change. For a small monthly object monitoring and automation charge, S3 Intelligent-Tiering monitors access patterns and automatically moves objects that have not been accessed to lower-cost access tiers.

- Since the launch of S3 Intelligent-Tiering in 2018, customers have saved more than $4 billion in storage costs by using S3 Intelligent-Tiering as compared to Amazon S3 Standard.

Cross-origin resource sharing (CORS)

Cross-origin resource sharing (CORS) defines a way for client web applications that are loaded in one domain to interact with resources in a different domain. With CORS support, you can build rich client-side web applications with Amazon S3 and selectively allow cross-origin access to your Amazon S3 resources.

- Additional explanation: **Cross-origin resource sharing (CORS)** is a mechanism for integrating applications. CORS defines a way for client web applications that are loaded in one domain to interact with resources in a different domain. This is useful because complex applications often reference third-party APIs and resources in their client-side code.
  - For example, your application may use your browser to pull videos from a video platform API, use fonts from a public font library, or display weather data from a national weather database. CORS allows the client browser to check with the third-party servers if the request is authorized before any data transfers.

S3 Object Lock

S3 Object Lock provides two retention modes:

1. **Governance mode:** In governance mode, users can't overwrite or delete an object version or alter its lock settings unless they have special permissions. With governance mode, you protect objects against being deleted by most users, but you can still grant some users permission to alter the retention settings or delete the object if necessary. You can also use governance mode to test retention-period settings before creating a compliance-mode retention period.
2. **Compliance mode:** In compliance mode, a protected object version can't be overwritten or deleted by any user, including the root user in your AWS account. When an object is locked in compliance mode, its retention mode can't be changed, and its retention period can't be shortened. Compliance mode helps ensure that an object version can't be overwritten or deleted for the duration of the retention period.

Legal Hold vs. Retention Period

- With **Object Lock**, you can also place a legal hold on an object version. Like a retention period, a legal hold prevents an object version from being overwritten or deleted. **However, a legal hold doesn't have an associated retention period and remains in effect until removed.** Legal holds can be freely placed and removed by any user who has the s3:PutObjectLegalHold permission.
- Legal holds are independent from retention periods. As long as the bucket that contains the object has Object Lock enabled, you can place and remove legal holds regardless of whether the specified object version has a retention period set. **Placing a legal hold on an object version doesn't affect the retention mode or retention period for that object version.**

EFS

Amazon Elastic File System (Amazon EFS) provides serverless, fully elastic file storage so that you can share file data without provisioning or managing storage capacity and performance. Amazon EFS is built to scale on demand to petabytes without disrupting applications, growing and shrinking automatically as you add and remove files. Because Amazon EFS has a simple web services interface, you can create and configure file systems quickly and easily. The service manages all the file storage infrastructure for you, meaning that you can avoid the complexity of deploying, patching, and maintaining complex file system configurations.

- Amazon Elastic File System (EFS) is a scalable, elastic, cloud-native file storage service offered by AWS. It is designed to provide a simple, serverless, set-and-forget experience that is fully managed, eliminating the need to provision or manage capacity. EFS is built to provide massively parallel shared access to thousands of Amazon EC2 instances, making it ideal for workloads and applications that require high throughput and low latency. It supports NFS-based storage solutions, seamlessly integrating with existing applications and services. With AWS EFS, users benefit from a scalable file system that can grow and shrink automatically as files are added and removed, meaning you only pay for the storage you use. Its integration with AWS Transfer Family enables the creation of fully managed, highly secure SFTP services that facilitate file transfers directly into and out of Amazon EFS, making it an excellent choice for serverless file storage and transfer solutions.

EBS

Encryption

What happens if you encrypt EBS volumes?

When you create an encrypted EBS volume and attach it to a supported instance type, the following types of data are encrypted:

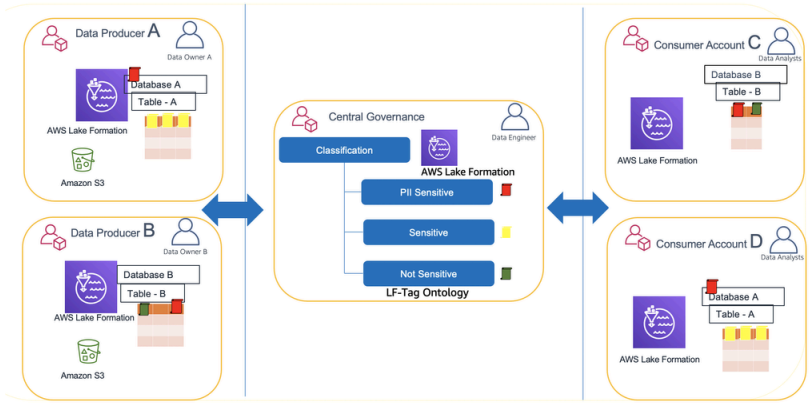
- Data at rest inside the volume
- All data moving between the volume and the instance
- All snapshots created from the volume
- All volumes created from those snapshots

Encryption by default is a Region-specific setting. If you enable it for a Region, you cannot disable it for individual volumes or snapshots in that Region.

Lake Formation

AWS Lake Formation helps you centrally govern, secure, and globally share data for analytics and machine learning. With Lake Formation, you can manage fine-grained access control for your data lake data on Amazon Simple Storage Service (Amazon S3) and its metadata in AWS Glue Data Catalog.

- **AWS Lake Formation is a robust data lake-building service that simplifies creating, securing, and managing data lakes.**
- It provides a centralized location for storing data from various sources and enables you to analyze this data using a wide range of AWS tools and services.
- One of the key features of AWS Lake Formation is **tag-based access control**, which allows administrators to grant or deny access to data based on tags attached to the data.



Key Management

KMS

AWS Key Management Service (KMS) is a service from Amazon Web Services (AWS) that helps users create, manage, and control cryptographic keys. These keys are used to encrypt and protect data across AWS services and applications.

- Lake Formation provides its own permissions model that augments the IAM permissions model. Lake Formation permissions model enables fine-grained access to data stored in data lakes as well as external data sources such as Amazon Redshift data warehouses, Amazon DynamoDB databases, and third-party data sources through a simple grant or revoke mechanism, much like a relational database management system (RDBMS). Lake Formation permissions are enforced using granular controls at the column, row, and cell-levels across AWS analytics and machine learning services, including Amazon Athena, Amazon QuickSight, Amazon Redshift Spectrum, Amazon EMR, and AWS Glue.

General

OLTP (online transaction processing)

A transactional or **OLTP** (online transaction processing) workload is a workload typically identified by a database receiving both requests for data and multiple changes to this data from a number of users over time where these modifications are called transactions.

Migration

AWS Directory Service

AWS Directory Service provides multiple ways to use Amazon Cloud Directory and Microsoft Active Directory (AD) with other AWS services. Directories store information about users, groups, and devices, and administrators use them to manage access to information and resources. **AWS Directory Service provides multiple directory choices for customers who want to use existing Microsoft AD or Lightweight Directory Access Protocol (LDAP)–aware applications in the cloud.** It also offers those same choices to developers who need a directory to manage users, groups, devices, and access.

Application Migration Service

AWS Application Migration Service (MGN) is a highly automated lift-and-shift solution that works by replicating your on-premises (physical or virtual) and/or cloud servers (referred to as “source servers”) into your AWS account. When you’re ready, AWS MGN automatically converts and launches your servers on AWS so you can quickly benefit from the cost savings, productivity, resilience, and agility of the cloud. Once your applications are running on AWS, you can leverage AWS services and capabilities to quickly and easily re-platform or refactor those applications.

Database Migration Service (DMS)

AWS Database Migration Service helps you migrate databases to AWS quickly 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.

Application Discovery Service

AWS Application Discovery Service helps you plan your migration to the AWS cloud by collecting usage and configuration data about your on-premises servers. Application Discovery Service is integrated with AWS Migration Hub, which simplifies your migration tracking as it aggregates your migration status information into a single console. You can view the discovered servers, group them into applications, and then track the migration status of each application from the Migration Hub console in your home region.

Monitoring

CloudWatch

Amazon **CloudWatch** monitors your AWS resources and the applications you run on AWS in real time. You can use CloudWatch to collect and track metrics, which are variables you can measure for your resources and applications. You can also create custom dashboards to display metrics about your custom applications, and display custom collections of metrics that you choose. You can create alarms that watch metrics and send notifications or automatically make changes to the resources you are monitoring when a threshold is breached.

Custom metric

You need to prepare a custom metric using CloudWatch Monitoring Scripts which is written in Perl. You can also install CloudWatch Agent to collect more system-level metrics from Amazon EC2 instances. Here's the list of custom metrics that you can set up:

– Memory utilization
– Disk swap utilization
– Disk space utilization
– Page file utilization
– Log collection

- For example, you can monitor the **CPU usage and disk reads and writes of your Amazon EC2 instances** and then use this data to determine whether you should launch additional instances to handle increased load. You can also use this data to stop under-used instances to save money

CloudWatch Logs

CloudWatch Logs enables you to centralize the logs from all of your systems, applications, and AWS services that you use, in a single, highly scalable service. You can then easily view them, search them for specific error codes or patterns, filter them based on specific fields, or archive them securely for future analysis.

CloudWatch Logs enables you to see all of your logs, regardless of their source, as a single and consistent flow of events ordered by time, and you can query them and sort them based on other dimensions, group them by specific fields, create custom computations with a powerful query language, and visualize log data in dashboards.

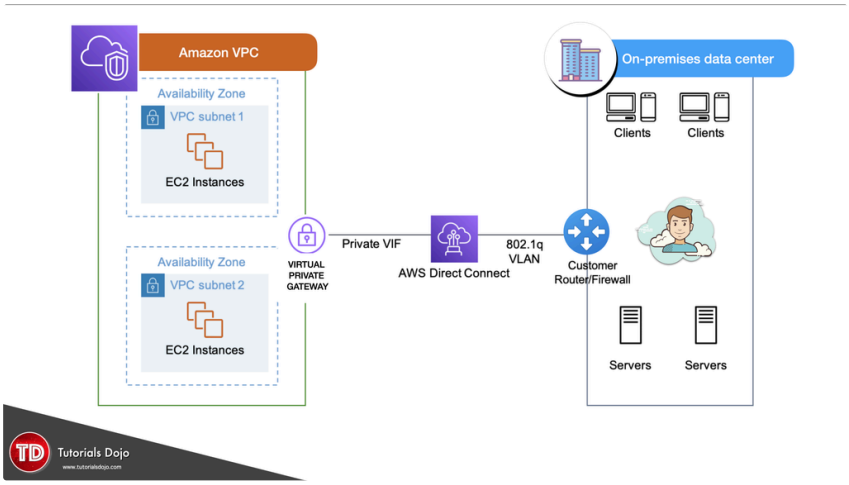
EventBridge

Amazon EventBridgeis a serverless event bus that enables integrations between AWS services, Software as a services (SaaS), and your applications. In addition to build event driven applications, EventBridge can be used to notify about the events from the services such as CodeBuild, CodeDeploy, CodePipeline, and CodeCommit.

Networking

Direct Connect

AWS Direct Connect links your internal network to an AWS Direct Connect location over a standard Ethernet fiber-optic cable. One end of the cable is connected to your router, the other to an AWS Direct Connect router.



Network Load Balancer

Network Load Balancer (NLB) is a highly scalable load-balancing service that AWS provides. It is designed to handle substantial traffic volumes and distribution of incoming connections across healthy targets within one or more target groups, including Amazon EC2 instances, containers, IP addresses, or Lambda functions. NLB operates at the transport layer (Layer 4) and handles TCP, UDP, and TLS traffic.

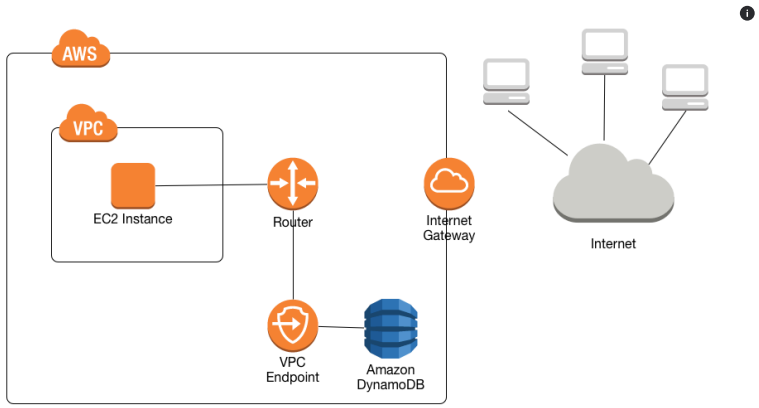
- **NLBs utilize active and passive health checks to assess a target's capability to handle requests.** These health checks are essential for the NLB component within an Amazon VPC as they regularly confirm the availability and responsiveness of registered targets. Through TCP, UDP, or TLS protocol requests, the health check guarantees that only healthy targets receive traffic. Failing targets are temporarily removed from rotation to maintain optimal performance and availability. Furthermore, NLBs are capable of managing millions of requests per second with ultra-low latencies, effectively handling sudden and volatile traffic patterns for applications with high traffic demands.

  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.

VPC

VPC endpoint

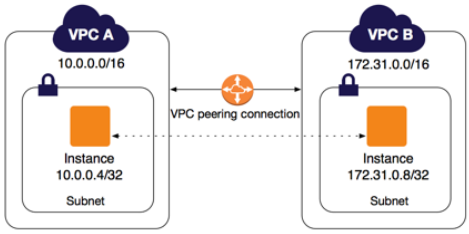
A **VPC endpoint** allows you to privately connect your VPC to supported AWS and VPC endpoint services powered by AWS PrivateLink without needing an Internet gateway, NAT computer, VPN connection, or AWS Direct Connect connection. Instances in your VPC do not require public IP addresses to communicate with resources in the service. Traffic between your VPC and the other service does not leave the Amazon network.



- There are two types of VPC endpoints: interface endpoints and gateway endpoints.
- As a rule of thumb, **most AWS services use VPC Interface Endpoint except for S3 and DynamoDB, which use VPC Gateway Endpoint.**
- An **interface endpoint** is an elastic network interface with a private IP address that serves as an entry point for traffic destined to a supported service. A gateway endpoint is a gateway that is a target for a specified route in your route table, used for traffic destined to a supported AWS service.

VPC Peering Connection

A **VPC peering connection** is a **networking connection between two VPCs that enables you to route traffic between them using private IPv4 addresses or IPv6 addresses**. Instances in either VPC can communicate with each other as if they are within the same network. You can create a VPC peering connection between your own VPCs, or with a VPC in another AWS account. The VPCs can be in different regions (also known as an inter-region VPC peering connection).



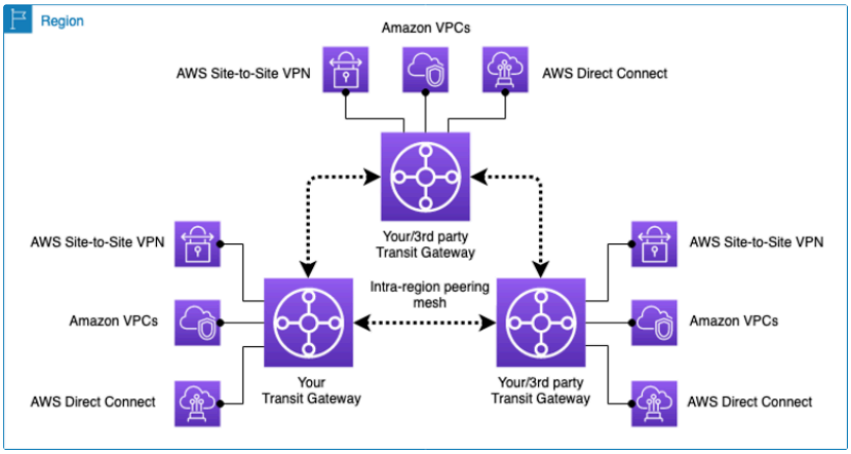
Inter-Region VPC Peering

Inter-Region VPC Peering provides a simple and cost-effective way to share resources between regions or replicate data for geographic redundancy. Built on the same horizontally scaled, redundant, and highly available technology that powers VPC today, Inter-Region VPC Peering encrypts inter-region traffic with no single point of failure or bandwidth bottleneck. Traffic using Inter-Region VPC Peering always stays on the global AWS backbone and never traverses the public internet, thereby reducing threat vectors, such as common exploits and DDoS attacks.

Transit Gateway

A transit gateway is a network transit hub that you can use to interconnect your virtual private clouds (VPCs) and on-premises networks.

- As your cloud infrastructure expands globally, inter-Region peering connects transit gateways together using the AWS Global Infrastructure. Your data is automatically encrypted and never travels over the public internet.
- A transit gateway attachment is both a source and a destination of packets. You can attach the following resources to your transit gateway:
  - One or more VPCs.
  - One or more VPN connections
  - One or more AWS Direct Connect gateways
  - One or more Transit Gateway Connect attachments
  - One or more transit gateway peering connections

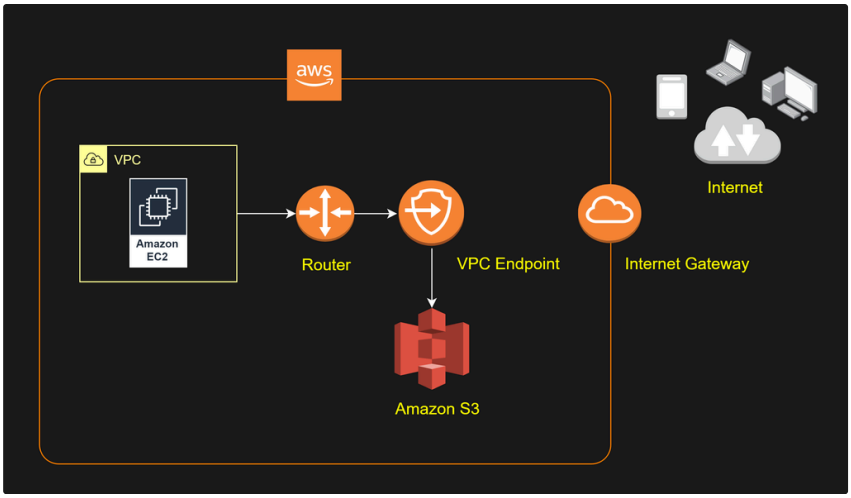


managed prefix

A managed prefix list is a set of one or more CIDR blocks. You can use prefix lists to make it easier to configure and maintain your security groups and route tables.

gateway endpoint

A **gateway endpoint** is *a gateway that you specify in your route table to access Amazon S3 from your VPC over the AWS network*. Interface endpoints extend the functionality of gateway endpoints by using private IP addresses to route requests to Amazon S3 from within your VPC, on-premises, or from a different AWS Region. Interface endpoints are compatible with gateway endpoints. If you have an existing gateway endpoint in the VPC, you can use both types of endpoints in the same VPC.

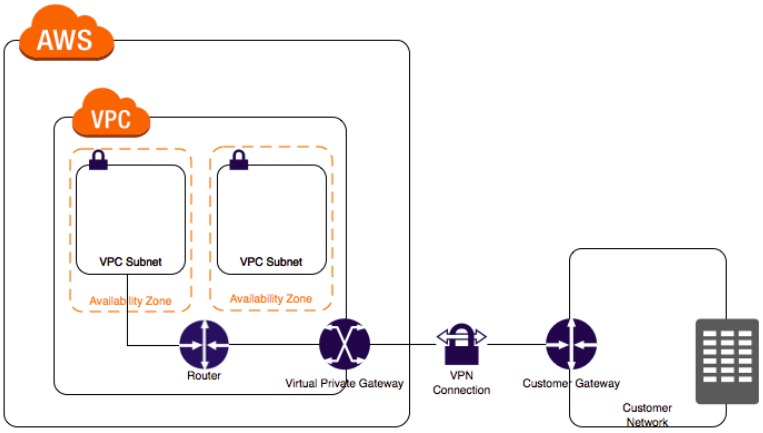


Customer Gateway

A **customer gateway** is a physical device or software application on your side of the VPN connection.

To create a VPN connection, you must create a customer gateway resource in AWS, which provides information to AWS about your customer gateway device. Next, you have to set up an Internet-routable IP address (static) of the customer gateway's external interface.

The following diagram illustrates single VPN connections. The VPC has an attached virtual private gateway, and your remote network includes a customer gateway, which you must configure to enable the VPN connection. You set up the routing so that any traffic from the VPC bound for your network is routed to the virtual private gateway.

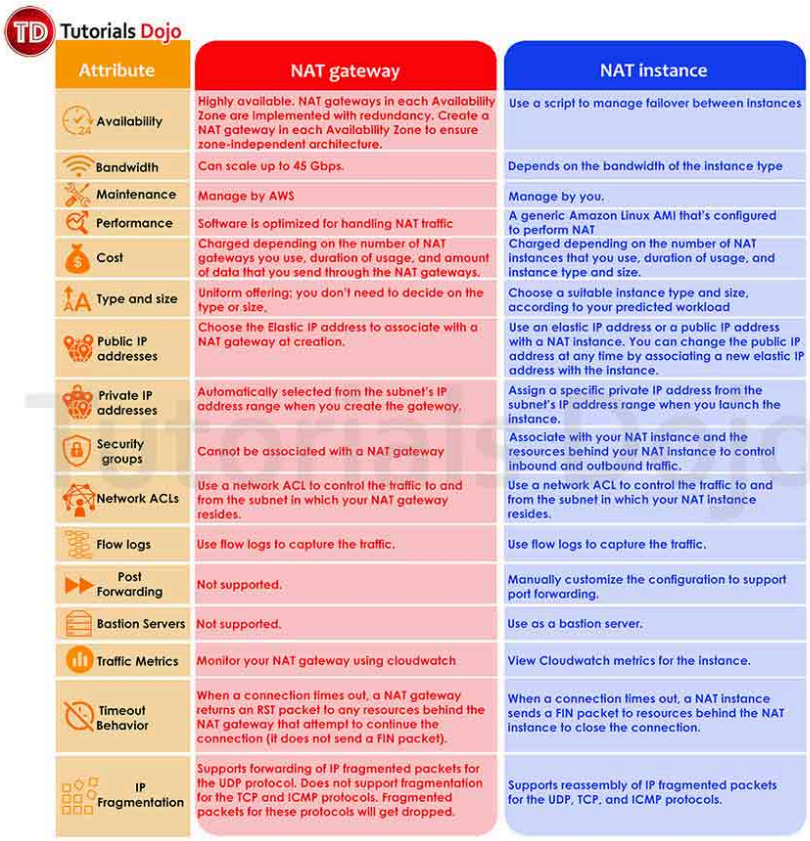


NAT

Network Address Translation (NAT) in Amazon Web Services (AWS) is a service that allows instances in a private subnet to connect to the internet and other services. NAT keeps private resources secure by preventing unsolicited traffic from reaching them.

- **AWS offers two kinds of NAT devices — a NAT gateway or a NAT instance.** *It is recommended to use NAT gateways, as they provide better availability and bandwidth over NAT instances.* The NAT Gateway service is also a managed service that does not require your administration efforts. A NAT instance is launched from a NAT AMI.

The difference between NAT gateway and NAT instance:



Route 53

Amazon **Route 53** is a highly available and scalable Domain Name System (DNS) web service. You can use Route 53 to perform three main functions in any combination: domain registration, DNS routing, and health checking.

If you choose to use Route 53 for all three functions, be sure to follow the order below:

1. Register domain names
2. Route internet traffic to the resources for your domain
3. Check the health of your resources

- You can create latency records for your resources in multiple AWS Regions by using latency-based routing.

Route 53 has different routing policies that you can choose from. Below are some of the policies:

Latency Routing

Latency Routing lets Amazon Route 53 serve user requests from the AWS Region that provides the lowest latency. It does not, however, guarantee that users in the same geographic region will be served from the same location.

Geoproximity Routing

Geoproximity Routing lets Amazon Route 53 route traffic to your resources based on the geographic location of your users and your resources. You can also optionally choose to route more traffic or less to a given resource by specifying a value, known as a bias. A bias expands or shrinks the size of the geographic region from which traffic is routed to a resource.

Geolocation Routing

Geolocation Routing lets you choose the resources that serve your traffic based on the geographic location of your users, meaning the location that DNS queries originate from.

Weighted Routing

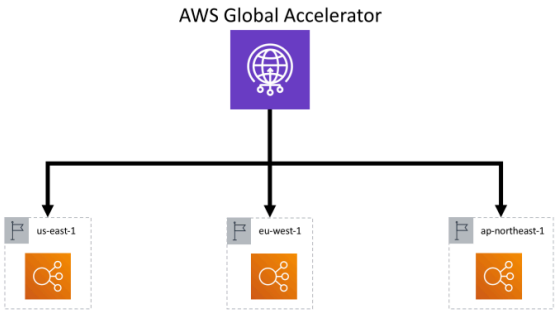
Weighted Routing lets you associate multiple resources with a single domain name (tutorialsdojo.com) or subdomain name (subdomain.tutorialsdojo.com) and choose how much traffic is routed to each resource.

Geolocation vs Geoproximity Routing

While both terms involve routing based on geographic location, "geoproximity routing" focuses on directing traffic to the geographically closest resource, taking into account both the user's location and the resource's location, while "geolocation routing" primarily routes traffic based solely on the user's location to a predefined resource, without considering the relative proximity of different resources

Global Accelerator

AWS **Global Accelerator** is a networking service that improves the performance and availability of applications for global users. It routes traffic to the nearest healthy endpoint in the AWS global network, reducing latency and improving reliability.



- AWS Global Accelerator is a service that improves the availability and performance of your applications with local or global users. It provides static IP addresses that act as a fixed entry point to your application endpoints in a single or multiple AWS Regions, such as your Application Load Balancers, Network Load Balancers, or Amazon EC2 instances.

Miscellaneous

AppSync pipeline resolvers

AppSync pipeline resolvers offer an elegant server-side solution to address the common challenge faced in web applications—aggregating data from multiple database tables. Instead of invoking multiple API calls across different data sources, which can degrade application performance and user experience, AppSync pipeline resolvers enable easy retrieval of data from multiple sources with just a single call. By leveraging Pipeline functions, these resolvers streamline the process of consolidating and presenting data to end-users.

Systems Manager Run Command

AWS Systems Manager Run Command lets you remotely and securely manage the configuration of your managed instances. A managed instance is any Amazon EC2 instance or on-premises machine in your hybrid environment that has been configured for Systems Manager. Run Command enables you to automate common administrative tasks and perform ad hoc configuration changes at scale. You can use Run Command from the AWS console, the AWS Command Line Interface, AWS Tools for Windows PowerShell, or the AWS SDKs. Run Command is offered at no additional cost.

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 set up 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.