Identity And Access Management (IAM)

With AWS Identity and Access Management (IAM), you can specify who or what can access services and resources in AWS. Central AWS Account Management service. **(Not requires any region selection)**

We have **4 objects** that we can create under IAM Service.

- **User:** It is a user that connects to the AWS console.
- **Group:** Groups are the objects that are consist of users with some common authorization rules. (System Admins, Developers, Read-Only etc.) Thanks to groups, we do not have to assign authorities one by one.
- **Role:** It is used for assigning permissions to the entities such as **AWS Services, applications, or users from other AWS accounts.** Roles define what actions can be performed and what resources can be accessed by the entity assuming the role. They allow you to provide access without access and secret keys.
- **Policy:** It is the object that defines the **rules for users or groups** as a JSON formatted file (Policy Document)

Policy Document Example

Policy has 4 major JSON elements: **Version, Effect, Action, and Resource.**

- **Version:** It specifies the language syntax rules that are needed by AWS to process a policy.
- **Effect:** Specifies whether the policy will **allow or deny** access.
- **Action:** Describes the type of action that should be allowed or denied. Statements **must include** either an Action or NotAction element. Each AWS service has its own set of actions that describe tasks that you can perform with that service. To find which actions are available, read the documentation of the services.
    - We can specify multiple values for the action element. **Format:** "Action": ["action-1","action-2"...]
    - We can use wildcard (*) in the action. **Format:** "s3:*" or "iam:*AccessKey*"
- **Resource:** Specifies the object or objects that the policy statement covers. Statements **must include** either a Resource or a NotResource element. The permissions policy language requires you to specify the resource or resources using the Amazon Resource Name (ARN) format.
    - Format: arn:partition:service:region:account:resource
    - We can use wildcard (*) in the resource.

Optional Policy JSON elements: **Condition, Principal, Sid.**

- **Condition:** Specify the circumstances under which the policy grants permission.
    - Format: "Condition" : { "{condition-operator}" : { "{condition-key}" : "{condition-value}" }}
    - Example: "Condition": {"StringLike": {"s3:prefix": ["janedoe/*"]}}

To learn more about policy elements, click here.

Role Options

Custom Trust Policy: Creates a role that includes some policies which is assumable by the principal account in the policy. In AWS console, it can be used via switch role.

Note: Trust Policy is the only resource-based policy that the IAM service supports.

Notes

- To give **programmatic access** to the user, you should create an **access key** based on your use cases.
- The users that are first created will have only 1 permission which allow them to change their passwords.
- Identity Providers allow us to connect IAM with 3rd Party Provider like Active Directory etc to enable **Single Sign On** (SSO). It supports two types of providers: **SAML, OpenID Connect.**

Best Practices

- Create groups and assign policy to the group instead of assigning them directly to the user.
- Do not use the root account and secure the accounts using MFA.
- Only assign a user the **minimum** amount of privileges they need to do their job.
- Always setup password rotations. Create and customize password rotation policies.
- Prefer using roles instead of credentials due to security perspective.

IAM Database Authentication

IAM has database authentication capabilities that would allow an **RDS database to only be accessed** using the **profile credentials** specific to your EC2 instances, which means you dont need to use password.

Storage Services

Simple Storage Service (S3)

- S3 is an **Object Based Storage**.
- You can upload any file to the S3.
- **Cannot be used** to run an operating system or database.
- Objects can be up to **5TB**. (There is no limit for object count.)
- Store files in the buckets.
- When you store an object, it keeps the **metadata** that includes some information like **content-length, created_at, last-modified.**
- Buckets are permanent storage entities and only removable when they are empty. After deleting a bucket, the name becomes available for reuse by any account after 24 hours if not taken by another account.
- Built for 99.95% - 99.99% **availability**. Designed for 99.9999% **durability.**
- **Buckets are private by default.** You have to allow public access on both the bucket and the object in order to make the bucket public.

Naming and URLs

- Bucket name must be globally unique. (Once created, you cannot change a bucket name.)
- URL format: **<bucket-name>.s3.<region>.amazonaws.com/<key-name>**

Security

- **Secure data with server-side encryption:** You can set default encryption on a bucket to encrypt objects they are stored in the bucket. Encrpytion Options can be listed as **SSE-S3, SSE-KMS, SSE-C**
    - When you use (SSE-S3), each object is encrypted with a unique key.
- **Access Control Lists (ACLs):** Define which **AWS accounts or groups are granted access**. You can attach S3 ACLs to **individual objects** within a bucket.
- **Bucket Policies:** S3 Bucket policies specify what actions are allowed or denied. (exp: allow user X to PUT but not DELETE objects in bucket)
- You can optionally add another layer of security by configuring a bucket to enable **MFA delete** on any request for deletion.

Types of Access Control	AWS Account Level Control	User Level Control
IAM Policies	No	Yes
ACLs	Yes	No
Bucket Policies	Yes	Yes

Versioning

- **Disabled by default.** You should enable it from the properties. **Once enabled, can not be disabled** – only suspended.
- Writing and deleting effects the version. All versions of an object are stored in S3.
- Deleting an object only adds **delete marker** to the object It does not delete the object.
- It costs money, as you'll be paying for every additional copy of your objects that you upload.

Lifecycle Management

It has use cases like lifecycle rules, replication rules. Basically you are defining some rules to be applied for the bucket or the objects. **Moving data between different S3 storage classess automatically**, replication and versioning rules etc. **Lifecycle policies can't work backwards.**

Note: Objects **must be stored at least 30 days** in the current storage class **before you can transition** them.

Object Locks

- Use S3 Object Locks to store objects using write once, read many (WORM) model.
- Object locks can be applied to the individual objects or across the bucket.
- Comes with two modes: **Governance Mode** (protected object version can not be touched without permissions), **Compliance Mode** (protected object version can not be touched by any user including root user)

Storage Classes (Tiers)

S3 Standard

- Designed for frequent access.

- Highly available AZ >=3

S3 Standard-Infrequent Access (S3 Standard-IA)

- Less frequently, **rapid access.**
- Highly available AZ >=3

S3 One Zone-Infrequent Access

- %20 less cost than S3 Standard-IA
- **Single AZ**
- Great for non-critical data

S3 Intelligent-Tiering

- **Automatically moves** your data to the most cost-effective tier based on access rate of the objects.

S3 Glacier

- Use only for **archiving** data.
- It is cheap.
- Optimized for data that is very infrequently accessed.

Glacier Options

- **Glacier Instant Retrieval:** Providers data archiving with instant retrieval time.
- **Glacier Flexible Retrieval:** Retrieving data can take minutes to 12 hours.
- **Glacier Deep Archive:** Retrieving data can take 12 hours to 48 hours.

Performance across the S3 Storage Classes

	S3 Standard	S3 Intelligent-Tiering*	S3 Standard-IA	S3 One Zone-IA+	S3 Glacier Instant Retrieval	S3 Glacier Flexible Retrieval	S3 Glacier Deep Archive
Designed for Durability	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)
Designed for Availability	99.99%	99.9%	99.9%	99.5%	99.9%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99.9%	99.9%	99.9%
Availability Zone(s)	≥3	≥3	≥3	1	≥3	≥3	≥3
Min Capacity Charge per Object	N/A	N/A	128KB	128KB	128KB	40KB	40KB
Minimum Storage Duration Charge	N/A	30 days	30 days	30 days	90 days	90 days	180 days
Retrieval Fee	N/A	N/A	Per GB retrieved	Per GB retrieved	Per GB retrieved	Per GB retrieved	Per GB retrieved
Storage Type	Object	Object	Object	Object	Object	Object	Object

Storage Class	Availability and Durability	AZ(s)	Use Case
S3 Standard	99.99% Availability 11 9's Durability	>=3	Suitable for most workloads (e.g., websites, content distribution, mobile and gaming applications, and big data analytics)
S3 Standard-Infrequent Access	99.9% Availability 11 9's Durability	>=3	Long-term, infrequently accessed critical data (e.g., backups, data store for disaster recovery files, etc.)
S3 One Zone-Infrequent Access	99.5% Availability 11 9's Durability	1	Long-term, infrequently accessed, non-critical data
S3 Intelligent-Tiering	99.9% Availability 11 9's Durability	>=3	Unknown or unpredictable access patterns
S3 Glacier Instant Retrieval	99.99% Availability 11 9's Durability	>=3	Provides **long-term data archiving** with instant retrieval time for your data.
S3 Glacier Flexible Retrieval	99.99% Availability 11 9's Durability	>=3	Ideal storage class for archive data that does not require immediate access but needs the flexibility to retrieve large sets of data at no cost, such as backup or disaster recovery use cases. Can be minutes or up to 12 hours.
S3 Glacier Deep Archive	99.99% Availability 11 9's Durability	>=3	Cheapest storage class and designed for customers that retain data sets for 7-10 years or longer to meet customer needs and regulatory compliance requirements. The standard retrieval time is 12 hours, and the bulk retrieval time is 48 hours.

S3 Lifecycle Transactions Model:

Amazon S3 supports a waterfall model for transitioning between storage classes, as shown in the following diagram.

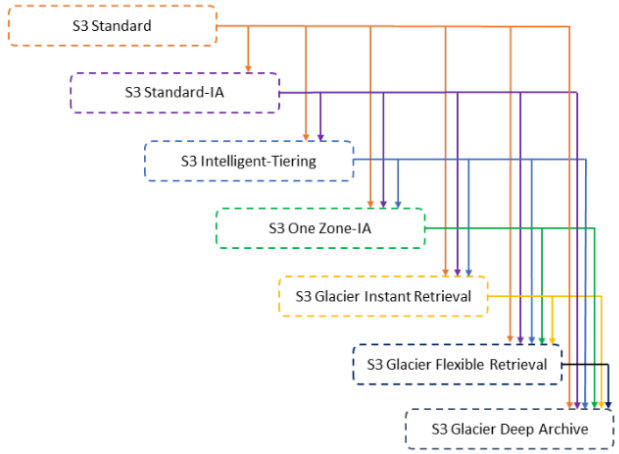


Figure 1 S3 Lifecycle Transactions Model

S3 Event Notifications:

The Amazon S3 event notification feature enables you to receive notifications **when certain events happen in your bucket**. To enable notifications, you must first add a notification configuration that identifies the events you want Amazon S3 to publish and the destinations where you want Amazon S3 to send the notifications.

Amazon S3 can send event notification messages to the following destinations.

- Amazon Simple Notification Service (Amazon SNS) topics
- Amazon Simple Queue Service (Amazon SQS) queues

- AWS Lambda
- Amazon EventBridge

S3 Object Lambda:

New capability that allows you to add your own code to **process data retrieved from S3 before returning it** to an application. Uses AWS Lambda functions to automatically process and transform your data as it is being retrieved from S3.

Best Practices

- Dividing objects into subfolders (prefixes) increases the performance.
- Parellelize **uploads** to increase efficiency with **Multipart uploads.** (Should be used for files over 100MB, must be used files over 5GB)
- Parellelize **downloads** by specifying byte ranges with **S3 byte-range fetches or S3 Select ScanRange.** (Should be used for files over 100MB, must be used files over 5GB)
- Best for **static website hosting**.

Compute Services

Elastic Compute Cloud (EC2)

Pricing Options

On-Demand: Pay by the hour or the second, depending on instance type.
Reserved: Reserved capacity for **1 or 3 years**. Up to %72 discount on the hourly charge.
Spot: Purchase unused capacity at a discount of up to %90. Prices change based on market demands and when the **price changes**, your spot instances **goes down.**
Dedicated: A pyshical EC2 Server dedicated for your use only. **The most expensive.**
Capacity Reservations: Reserve capacity for your EC2 instances in a specific Availability Zone.
Spot Fleet: Set of **Spot Instances** and optionally **On-Demand** Instances.

Note: Reserved instances **can be sold** on the Reserved Instance Marketplace.

Note: You can only change the tenancy of an instance from dedicated to host, or from host to dedicated after you've launched it.

Dedicated Hosts vs Dedicated Instances

- Dedicated instances **cannot be used** for **existing** server-bound **software licenses**.
- Dedicated instances **may share hardware with other instances** from the same AWS account **that are not dedicated instances.**

Security Groups

- All outbound traffic is allowed by default.
- All inbound traffic is blocked by default.

Note: If either Launch Template Tenancy or VPC Tenancy is set to dedicated, then the instance tenancy is also dedicated.

Network Adaptors available for EC2

ENI: Basic networking. Low cost.
Enhance Networking: 10Gbs and 100Gbs. Anywhere you need **high throughput**.
EFA: When you need **high performance computing**, machine learning scenarios.

Types of Placement Groups

Cluster Placement Groups: Low network **latency**, **high** network **thoughput**. Usecase for High Performance Computing (HPC) applications.

Spread Placement Groups: Individual critical EC2 instances. Strictly **places a small group of instances** across distinct underlying hardware to reduce correlated failures. **(maximum 7 instances per AZ).**

Partition Placement Groups: Many EC2 instances, **large distributed and replicated workloads**; Usecase for HDFS, HBase, Cassandra, Hadoop, Kafka.

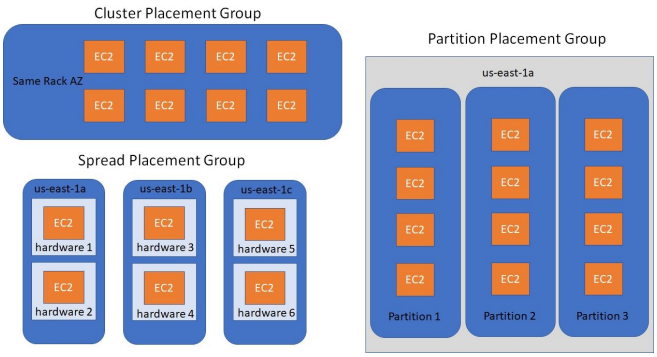


Figure 2 Placement Group Types

Outposts: Scenario about extending AWS to your data center. Outposts solutions allow you to **extend and run native AWS services on premises**.

EC2 Hibernation: It preserves the in-memory RAM on persistent storage (EBS) to **boot the instance up faster**.

Instances Types

General Purpose: Provide a balance of compute, memory and networking resources.
Memory Optimized: Designed to deliver fast performance for workloads that process large data sets in memory.

Storage Optimized: Designed for workloads that require high, sequential read and write access to very large data sets on local storage.

Note: There is a vCPU-based On-Demand Instance limit per Region.
https://docs.aws.amazon.com/general/latest/gr/ec2-service.html

Note: Retrieving data from ec2: http://169.254.169.254/latest/...

- **For meta-data:** http://169.254.169.254/latest/meta-data/. Metadata includes ip-address, instance ID, public keys etc.
- **For user-data:** http://169.254.169.254/latest/user-data/. Userdata is the scripts or parameters that are provided by the user before launch of the instance.

Elastic Block Storage (EBS) and Elastic File Storage (EFS)

Note: EC2 instances must be in the same AZ as the EBS volumes.

IOPS vs Throughput

IOPS	Throughput
Measures the number of read an write operations per second.	Measures the number of bits processed per second.

Provisioned Throughput: Specify **a level of throughput** that the file system can drive **independent of the file system's size** or burst credit balance.

*Bursting Throughput:*Throughput **scales with the amount of storage** in your file system.

Note: Any time that you change throughput mode, you **must wait at least 24 hours** before you can **change the throughput mode again**.

EFS Performance Modes

Max I/O performance mode is used to scale to **higher levels of aggregate throughput** and operations per second. This scaling is done with a tradeoff of slightly higher latencies for file metadata operations. Highly parallelized applications and workloads, such as **big data analysis, media processing**, and genomic analysis, can benefit from this mode.

SSD Volumes (EBS)

SSD Volumes are highly available and scalable storage volumes that you can attach to an EC2 instance.

gp2	gp3	io1	io2	io2 Block Express

Up to 16,000 IOPS per volume	Predictable 3,000 IOPS baseline performance and 125/MiB/s regardless of the volume size	Up to 64,000 IOPS per volume	Up to 64,000 IOPS per volume	Up to 256,000 IOPS per volume
Suitable for boot disks and general applications.	Suitable for high performance applications.	Suitable for OLTP and latency-sensitive applications	Suitable for OLTP and latency-sensitive applications	
		50 IOPS/GiB	500 IOPS/GiB	

Tips

- Volumes exists on EBS, whereas snapshots exist on S3.
- Snapshots are point-in-time photographs of volumes.
- The first snapshot will take some time to create.
- For consistent snapshots, stop the instance and detach the volume.
- You can share snapshots between AWS account as well as between regions.
- You can resize EBS volumes on the fly as well as changing the volume types.

Encryption

- Encryption operations occur on the servers that host EC2 instances, ensuring the security of both data-at-rest and data-in-transit between an instance and its attached EBS storage.
- You can attach both encrypted and unencrypted volumes to an instance simultaneously.
- You can't encrypt existing unencrypted Amazon EBS volumes.

Steps to Encrypt EBS Volumes

- Create a snapshot from the existing ebs which is unencrypted.
- Create a copy of the snapshot and select the encryption option.
- Create an AMI from the encrypted snapshot.
- Use that AMI to launch new encrypted instances.

EFS

- Supports NFSv4 protocol.
- Only pay what you use.
- Can scale up to petabytes. **Automatically** scales down and up itself based on the requirements.
- Can support thousands of **concurrent NFS connections** which means it **can be reached by multiple EC2** instances simultaniously.
- Data is stored across multiple AZs.
- Also have storage classes: EFS Standard, EFS Infrequent Access(IA), EFS Archive

EFS, FSx for Windows, FSx for Lustre

When to choose what?

- **EFS:** When you need distributed, highly resilient storage storage forbased applications and linux instances. **NFS protocol Based.**
- **FSx for Windows:** When you need centralized storage for Windows based applications such as Sharepoint, MSSQL, Workspaces, IIS Web Server etc. **SMB protocol based.**
    - It's possible to connect Linux instances to this file system using SMB/CIFS protocol.
- **FSx for Lustre:** When you need high-speed, high capacity distributed storage. This will be for the **machine learning**, AI, financial, video processing applications that do high performance computing. Provides the ability to both process the **'hot data'** in a parallel and distributed fashion as well as **easily store the 'cold data' on Amazon S3**.
    - **Lustre Persistent volumes**
    - **Scratch volumes**

EBS vs Instance Store

EBS

Amazon Elastic Block Store (Amazon EBS) is block-level storage that you can attach to an Amazon EC2 instance.

- EBS **can be stopped and detached or attached** to the another machine.
- It will be deleted on termination by default. However **you can keep it** if you wish **by changing DeleteOnTermination attribute**. However, **for running instances, you can change it via command line only**.

Note: AWS announced the Amazon **EBS multi-attach feature** that permits **Provisioned IOPS SSD (io1 or io2) volumes** to be attached to multiple EC2 instances at one time. This **feature is not available for all instance types.**

Raid

A RAID array uses **multiple EBS volumes** to improve performance or redundancy.
Raid 0: When **I/O performance** is important.
Raid 1: When **fault tolerance** is important.

Instance Store

Instance store provides **temporary block-level storage** for your instance. This storage is located on disks that are physically attached to the host computer.

- Instance store volumes are sometimes called ephemeral storage.
- **Instance stores cannot be stopped**. You loose data if;
    - The underlying disk drive fails
    - The instance stops
    - The instance hibernates
    - The instance terminates
    - Hardware disk failure
- It will be deleted on termination of the instance. There is **no choice to not delete**.

- **Only attachable on launch.** Can not be attached on restart.
- **Can not** detach and attach to another instance.

Instance store is ideal for the temporary storage of information that changes frequently such as buffers, caches, scratch data, and other temporary content, or for data that is replicated across a fleet of instances.

As Instance Store based volumes provide **high random I/O performance at low cost**.

AWS BACKUP

- Provides a backup console, public APIs, and a command line interface to **centrally manage backups** across the AWS storage, compute, database, and hybrid services your applications run on such as EC2, EBS, EFS, Amazon FSx for Lustre, Amazon FSx for windows, storage gateway.
- You can use AWS Organizations and AWS Backup to back up your aws services acrross multiple AWS accounts.

DATABASES

Note: **With Multi-AZ**, RDS creates an **exact copy** of your production database in another availability zone automatically.
Note: Aurora is always Multi-AZ. You can't have Aurora as Single-AZ.

Tip: You cannot alter the encryption state of an RDS database after you have deployed it. You also cannot create encrypted replicas from unencrypted instances.

Online Transaction Processing (OLTP) vs Online Analytical Processing (OLAP)

OLTP: Processing data from transactions in real time. OLTP is all about data processing and completing large number of small transactions in real time.
OLAP: Processing complex queries to analyze historical data. OLAP is all about data analysis using large amounts of data, as well as complex queries that take long time to complete.

Note: **Amazon RDS Custom** is a managed database service for applications that require **customization of the underlying operating system** and database environment.

Increasing Performance of the RDS

Read Replica

A read replica is read only copy of your RDS database. It is only used for performance.
Automatic backups must be enabled in order to deploy read replica.
You can create multiple replicas.

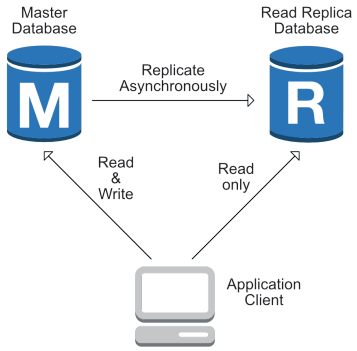


Figure 3 RDS Read Replica

Multi-AZ vs Read Replica

Multi-AZ

- An exact copy of the production database in another availability zone.
- Used for disaster recovery.
- Automatic failover to the standby instance.
- **When failing over,** Amazon RDS simply **flips the CNAME** for your DB instance to point at the standby. typically it **takes one to two minutes** to complete**.**
- Synchronous replication.

Read Replica

- A read-only copy of the primary database in the same AZ, cross-AZ, or even cross-Region.
- Used for better performance.
- Great for read-heavy workloads and takes the load from your primary database for read-only workloads.
- Asynchronous replication.
- Only charges on data replication across regions.
- If the master database is encrypted, the read replicas are also encrypted. Can not be unecrypted.

Note: For Amazon Aurora, each Read Replica is associated with a priority tier (0-15). In the event of a failover, Amazon Aurora will promote the Read Replica that has the highest priority (the lowest numbered tier). If two or more Aurora Replicas share the same priority, then Amazon RDS promotes the replica that is largest in size. If two or more Aurora Replicas share the same priority and size, then Amazon Aurora promotes an arbitrary replica in the same promotion tier.

AURORA

- 6 copies of your data within minimum 3 AZ's and 2 copies for each AZ. 3x2 = 6
- Aurora snapshots can be shared with other AWS Accounts.
- 3 types of replicas available: Aurora, MySQL, and PostgreSQL. **Automated failover is only avaliable with Aurora replicas.**
- Automated backups are turned on by default.
- Use **Auorora Serverless** if you want simple, cost-effective option for infrequent, intermittent or upredictable workloads. It can **start, shut and scale capacity automatically**, according to individual application's requirements.

Note: **Amazon RDS Proxy** effectively manages and **optimizes database connections**, particularly beneficial in scenarios with a substantial number of concurrent connections from serverless components.

Note: Amazon **Aurora Global Database** is designed for **globally distributed applications**, allowing a single Amazon Aurora database to **span multiple AWS regions**. Provides **disaster recovery from Region-wide** outages.

DynamoDB

NoSQL database for AWS.

Note: **Global tables** provide us **multi-region replication** of the table. To enable global tables, you **must enable dynamodb 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.

Tip: ACID (Atomic-Consistent-Isolated-Durable)
Any question with ACID requirements; think **DynamoDB Transactions.**

OTHER DATABASES

Amazon DocumentDB: Allows you to run **MongoDB on the AWS** cloud.
Amazon KeySpaces: Allows you to run **Apache Cassandra on the AWS** cloud.
Amazon TimeStream: Think when **time-series database** mentioned.
Amazon Neptune: A **graph database** and would be suitable to handle graph queries.

Consistency
https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/HowItWorks.ReadConsistency.html

Amazon Quantum Ledger Database (Amazon QLDB)

Amazon Quantum Ledger Database (Amazon QLDB) is a fully managed ledger database that provides a transparent, **immutable**, and **cryptographically** verifiable transaction log.

NETWORKING (VPC)

Note: Everytime you create a VPC, it creates Route table and ACL(Access Control List) by default.
Note: VPC can only have 1 Internet Gateway.

SUBNET

- Each subnet associated to Availability zone (AZ).
- The **first four IP** and the **last IP** in each subnet is reserved by AWS. (**5 IP reserved**)
- To create public subnet, you should create an internet gateway and attach it to VPC. Then you should create a route table which has route to the Internet Gateway and associate it with the subnet that you want to be public.

NAT GATEWAY

- To enable instances in a private subnet to connect to the internet or other AWS services while preventing the internet from initiating a connection with those instances.
- We should provision NAT gateway on the public subnet.
- Automatically assigned a public IP address.

NAT GATEWAY VS NAT INSTANCE

NAT GATEWAY	NAT INSTANCE
Port forwarding **not supported**	**Supports** port forwarding
Can not be used as a bastion server	**Can be used** as a bastion server
You **cannot associate** security groups with NAT gateways.	You **can associate** security groups with your NAT instance
Uniform offering; you don't need to decide on the type or size.	Choose a suitable instance type and size, according to your predicted workload.

Network ACL (Access Control List)

- **Subnet level** security.
- Each subnet has to be attached to 1 Network ACL.
- You can associate multiple subnets to Network ACL.
- You can **block IP addresses**. And you can also deny traffic on any **malicious ports**.
- **Network ACLs are stateless.** Basically, you have to go and allow for both inbound and outbound traffic.
- You **can not edit or remove * All Traffic Deny rule** which is default.

Security Groups

- **Resource level** security.
- As default, it blocks all inbound traffic and allows all outbound traffic (when you create). Allows inbound traffic from itself (when aws creates initially)
- **Security Groups are stateful**, responses to allowed inbound traffic are allowed to flow out, regardless of outbound rules.
- You can change the attached security groups of the EC2 instances even the instance is in running state. And add multiple security groups **up to 5** to single instance.

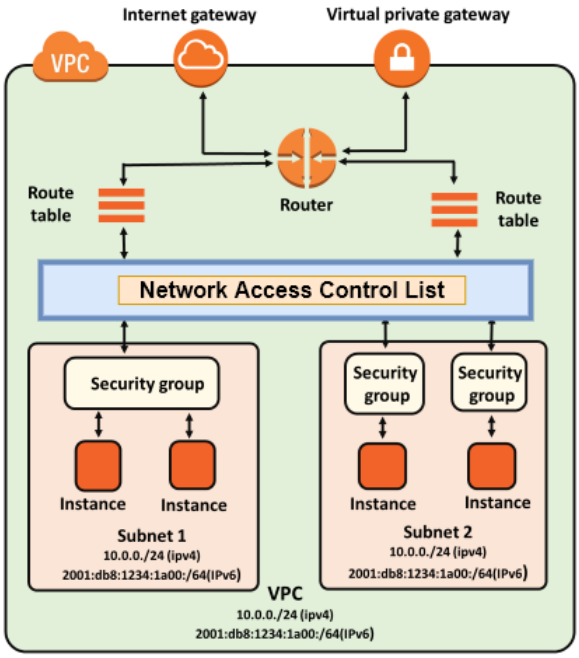


Figure 4 Overall Network Diagram

VPC Endpoints

Provides **quick, private connection** from the internal network **instead of going to the Internet** for the determined endpoints which means the traffic between your VPC and the other service **does not leave the Amazon Internal Network**.

Has two options; **Interface endpoints**, **Gateway endpoints**.

Gateway endpoint **supports S3 and Dynamodb** right now.

Tip: Avoid regional data transfer charge - VPC endpoint.

VPC Peering Connections

- Allows you to connect 1 VPC with another.
- **CIDR blocks** of the VPC's **can not overlap** eachother.
- You can peer regions, and VPC's which are exitst in another account.

Note: You need to **set routing** to the peering connection **manually** when the connection is created.

VPC Sharing

VPC sharing allows multiple AWS accounts to create their application resources, such as Amazon EC2 instances, Amazon Relational Database Service (RDS) databases, Amazon Redshift clusters, and AWS Lambda functions, into shared, centrally-managed virtual private clouds (VPCs). In this model, the account that owns the **VPC (owner) shares one or more subnets with other accounts (participants) that belong to the same organization** from AWS Organizations.

AWS PrivateLink

Best way to **expose a service VPC to** tens, hundreds, thousands of **customer VPCs**.
Requires a NLB on the service VPC and an **ENI** on the customer VPC.

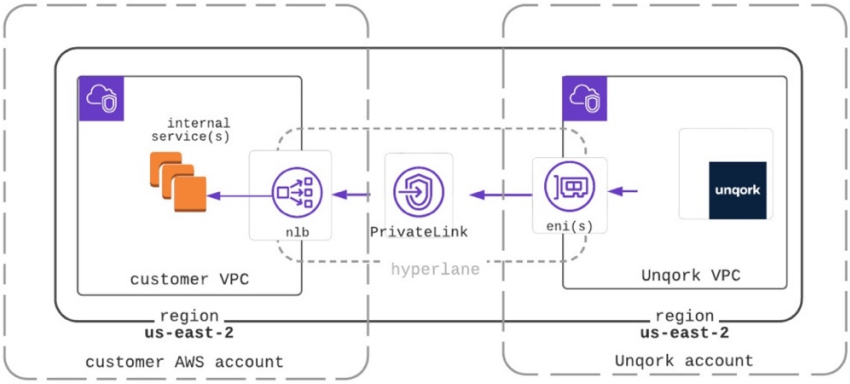


Figure 5 Private Link

AWS VPN CloudHub

If you have multiple sites, each **with their own VPN Connection**, you can use CloudHub to connect those sites together.

AWS VPN

It uses internet protocol security **(IPSec)** communications to create **encrypted VPN tunnels** between two locations.

Direct Connect

AWS Direct connect is a cloud service solution that makes it easy to **establish a dedicated network connection between AWS and on-premises.**

- **Secure than VPN**. Because the **connection is not goes over the internet**.
- AWS **Direct Connect bandwidth** starts at 50 Mbps and **goes up to 100 Gbps**.
- Lead times are often **longer than 1 month**.
- Low latency and high throughput connection.
- **Cannot provide an encrypted connection** between a data center and AWS Cloud **by itself**.

Note: **VPN and Direct Connect can be combined** together for dedicated, encrypted, low latency, and high throughput connection.

Note: **VPN tunnels** can only have a **maximum bandwidth of 1.25 Gbps.**

Transit Gateway

Connects VPCs and on-premises networks **through a central hub**.

- Simplified network topology.

- Supports IP multicast.
- Works with Direct Connect and VPN Connections.

Note: AWS Transit Gateway also enables you to scale the IPsec VPN throughput with equal cost multi-path (ECMP) routing support over multiple VPN tunnels.

Tip: Any question about 5G Mobile Edge Computing; think AWS Wavelength**.** Private cellular network; think Private 5G.

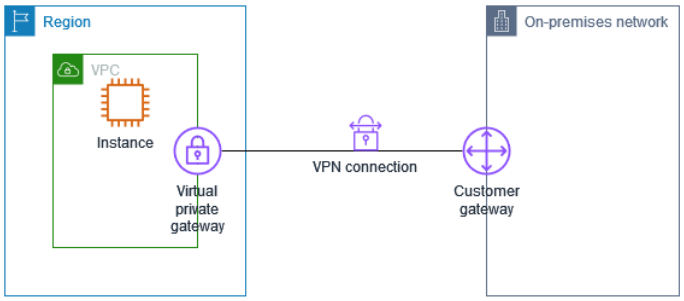
Gateway Load Balancer

Gateway Load Balancer helps you easily deploy, scale, and manage your third-party virtual appliances. It gives you one gateway for distributing traffic across multiple virtual appliances while scaling them up or down, based on demand. This decreases potential points of failure in your network and increases availability.

Virtual Private Gateway

A virtual private gateway is the VPN concentrator on the Amazon side of the Site-to-Site VPN connection. You create a virtual private gateway and attach it to a virtual private cloud (VPC) with resources that must access the Site-to-Site VPN connection.

The following diagram shows a VPN connection between a VPC and your on-premises network using a virtual private gateway.



ROUTE 53

Simple routing policy — Use for a single resource that performs a certain role for your domain, for example, a web server that provides content to the example.com site.

Failover routing policy — Use when you want to configure active-passive failover.

Geolocation routing policy — Use when you want to route traffic based on the location of users.

Geo-proximity routing policy — Use when you want to route traffic based on the location of your resources and optionally switch resource traffic at one location to resources elsewhere.

Latency Routing Policy — Use when you have resources across multiple AWS regions and want to route traffic to the region that provides the best latency.

Multi-Value Response Routing Policy — Use when you want Route 53 to respond to DNS queries with up to eight randomly selected healthy records.

Weighted routing policy — Use to route traffic to multiple resources in the proportions you specify.

Note: For each VPC that you want to associate with the Route 53 hosted zone, change the following VPC settings to true: **enableDnsHostnames**, **enableDnsSupport.**

ELASTIC LOAD BALANCING (ELB)

Note: X-Forwarded-For header in the request gives us the IP address of the end user.

Sticky Session: It enables users to stick to the same EC2 Instance. Can be usefull if you are storing information locally to that instance.

Deregistration Delay/Connection Draining: Keep existing connections for some time open if the EC2 instance becomes unhealthy.

SNI Support: With SNI support AWS makes it easy to use **more than one certificate** with the same ALB.

MONITORING

CloudWatch

- There is no default alarm. Everything that you want to hear about, you have to set an alarm.
- Default and Custom metrics: Default metrics (CPU Utilization, network throughput). Custom metrics such as memory utilization, disk space, Network latency etc. You have to create **custom metric using the agent** on the resource.
- Basic and detailed monitoring: **Basic** is **5-min intervals** where **detailed** is **1-min interval**. For detailed monitoring you actually pay more.

CloudWatch Logs:

Terms:

Log Event: This is the record of what happened. It contains the data and timestamp.
Log Stream: A collection of log events from the **same source**.
Log Group: This is a collection of log streams.

CloudWatch Logs Insights: **Running SQL like queries** on the CloudWatch log groups.

CloudWatch Agent Setup Guide:
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/QuickStartEC2Instance.html

CloudWatch Container Insights: Use CloudWatch Container Insights to collect, aggregate, and summarize metrics and logs from your containerized applications and microservices.

VPC Flow Logs: VPC Flow Logs is a feature that enables you to capture information about the **IP traffic going to and from network interfaces in your VPC**. Flow log data can be published to the following locations: Amazon CloudWatch Logs, Amazon S3, or Amazon Data Firehose.

Note: Instance store volumes are not supported for automatic recovery by Amazon CloudWatch alarms.

Tip: Any questions about **real-time logging**; think **Kinesis**

High Availability and Scaling

Launch Templates

Launch templates specify all the needed settings that go into builiding out an EC2.

- Can be versioned. (Launch configuration can not be versioned.)
- Network configurations included.
- Userdata included.

Auto Scaling Groups

An auto scaling group contains a **collection of EC2 instances** that are treated as a collective group for purposes of **scaling** and **management**.

Auto Scaling Lifecycle: https://docs.aws.amazon.com/autoscaling/ec2/userguide/ec2-auto-scaling-lifecycle.html

Auto Scaling Rebalancing

EC2 Auto Scaling **launches new instances before terminating the old ones**, so that rebalancing does not compromise the performance or availability of your application.

Auto Scaling Group Configuration Opportunities

- We can choose different types of instances in the group. (We can determine it as a percentage)
- When purchasing, we can choose between on demand and spot in percentage terms.
- We can choose a load balancer.
- It is possible to choose which subnets the instances can be created in. For high availability, we must choose at least 2 different subnets. It balances itself according to the number. For example, we created 4 different instances and 2 different subnets were selected (AZ). In this case, it automatically balances 2 to 2.

Note: Amazon EC2 Auto Scaling uses the **default termination policy**. It selects the Availability Zone with two instances, and **terminates** the instance that was launched from **the oldest launch template or launch configuration (first, oldest launch configuration)**. If the instances were launched from the same launch template or launch configuration, Amazon EC2 Auto Scaling **selects the instance that is closest to the next billing hour** and terminates it.

Note: Auto Scaling Group does not have a dynamic Elastic IPs attachment feature.

Auto Scaling Policies:

Scaling Policy	What it is	When to use
Target Tracking Policy	Adds or removes capacity as required to keep the metric at or close to the specific target value.	You want to keep the CPU usage of your ASG at 70%
Simple Scaling Policy	Waits for the health check and cool down periods to expire before re-evaluating.	Useful when load is erratic. AWS recommends step scaling instead of simple in most cases.
Step Scaling Policy	Increases or decreases the configured capacity of the Auto Scaling group based on a set of scaling adjustments.	You want to vary adjustments based on the size of the alarm breach

Scaling types:

- Maintain – keep a specific or minimum number of instances running.
- Manual – use maximum, minimum, or a specific number of instances.

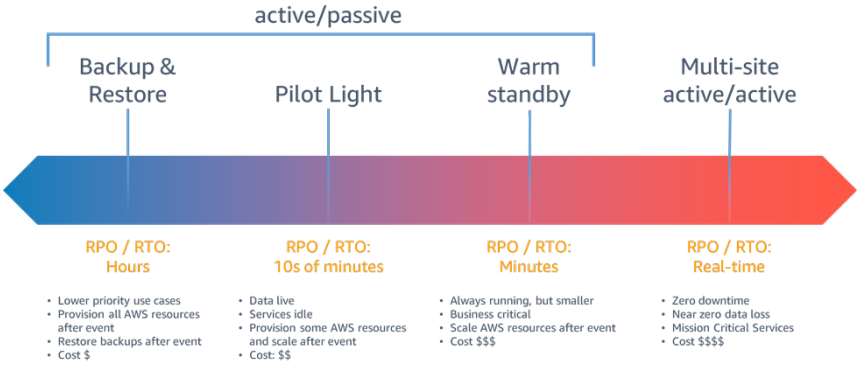
- Scheduled – increase or decrease the number of instances based on a schedule.
- Dynamic – scale based on real-time system metrics (e.g. CloudWatch metrics).
- Predictive – machine learning to schedule the right number of EC2 instances in anticipation of approaching traffic changes.

Disaster Recovery

Recovery Time Objective(RTO): In the event of failure. **How quickly you need to recover?**
Recovery Point Objective (RPO): In the event of failure. **How much data you can afford to lose?**

Disaster Recovery Strategies

- Backup and Restore
- Pilot Light
- Warm Standby (Active/Passive)
- Active/Active (Multi-Site)



Decoupling Workflows

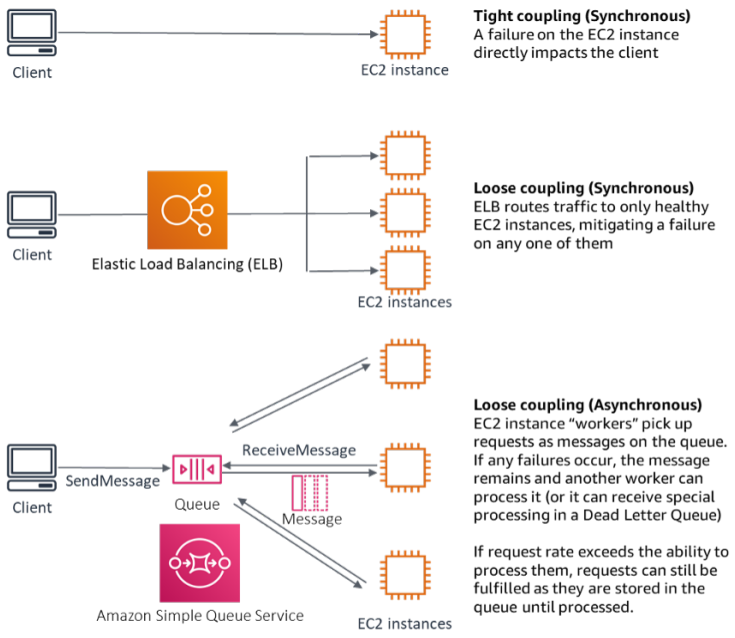


Figure 6 Coupling diagram

Simple Queue Service (SQS)

SQS is a managed service for sending and receiving messages between different parts of your application. Easy to configure.

Settings

Delivery Delay: Default is 0; can be set up to 15 mins.
Message Size: Up to 256KB of text in any format.
Message Retention: Default is 4 days; can be set between 1 minute up to 14 days.
Long vs Short Polling: Default is short. But the long should be the default. (Reconnections to ask messages)
Queue Depth: Item count. Can be listanable with CloudWatch to set an alarm.
Visibility Timeout: To prevent other consumers from processing the message again, Amazon SQS sets a visibility timeout, a period of time during which Amazon SQS prevents all consumers from receiving and processing the message. Default is 30 seconds; can be set between 0 up to 12 hours.

Note: Using **long polling can reduce the cost** of using SQS because you can **reduce the number of empty receives.**

Dead-Letter Queue

In message queueing a dead letter queue (DLQ) is a service implementation to store messages that the messaging system **cannot or should not deliver**.

FIFO Queues

- The name of a FIFO queue **must end with the .fifo** suffix.
- SQS FIFO **guarantees the order of processing** and ensure that **each message processed exactly once.**
- SQS FIFO **defaultly has 300 messages per second**. Can be increased with batching (maximumum 10 batch 300*10 = 3000) **up to 3,000 messages** per second.

- If you are creating a Dead-Letter Queue for a FIFO Queue, the type of dead-letter queue also must be FIFO.

Note: Standard SQS Queues has nearly unlimited messages per second.

Note: AWS recommend **using separate queues** when you need to provide **prioritization** of work.

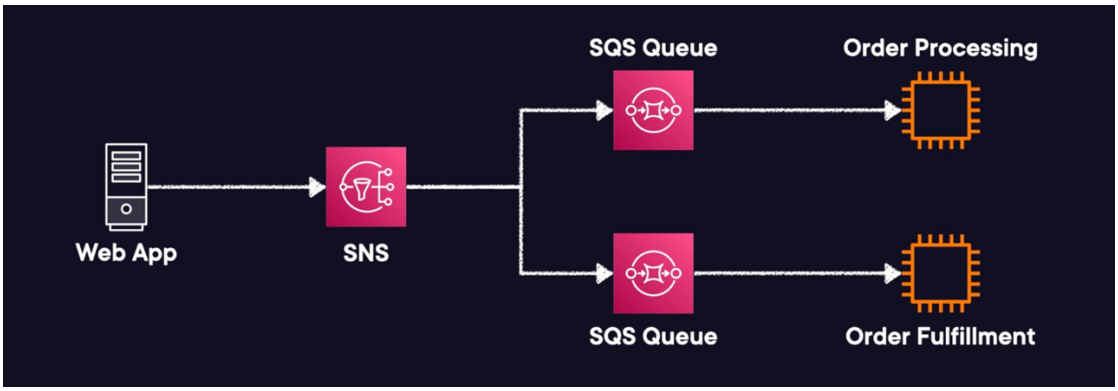
Temporary Queues

Temporary queues help you save development time and deployment costs when using **common message patterns** such as **request-response**.

Simple Notification Service (SNS)

Push-based message application.

SNS Fanout architecture



Simple Email Service (SES)

SES allows sending emails in an HTML and text format.

API Gateway

Amazon API Gateway allows you to create RESTful APIs and WebSocket APIs to provide access to your backend services, such as AWS Lambda functions, AWS Step Functions, or HTTP endpoints. It acts as a front door for your applications, allowing you to control access, apply security measures, and monitor usage.

Tip: You can give custom domain names to API Gateway.

Note: To prevent your API from being overwhelmed by too many requests, Amazon API Gateway throttles requests to your API using the token bucket algorithm, where a token counts for a request. Specifically, API Gateway sets a limit on a steady-state rate and a burst of request submissions against all APIs in your account. In the token bucket algorithm, the burst is the maximum bucket size.

Note: You can enable Amazon API caching in Amazon API Gateway to cache your endpoint's responses.

AWS Batch

Anything related to batch workflows that is **long running > 15mins**. Will likely involve AWS Batch. It is an on demand **alternative to AWS Lambda**. Questions regarding an alternative solution to AWS Lambda due to runtime requirements.

Amazon MQ

- Fully managed **service for open-source message brokers**
- Message broker service allowing easier migration of existing applications to the AWS Cloud.
- Currently supports **Apache ActiveMQ and RabbitMQ**.
- Protocols: AMQP, JMS, MQTT, OpenWire, STOMP.

When to choose Amazon MQ or SNS with SQS

If **migrating existing applications** with messaging system in place, you might **consider Amazon MQ**. If **creating new applications**, look at **SNS and SQS**. These are simpler to use and are highly scalable. Amazon MQ **requires private networking**. Whereas, SNS and SQS are **publicly accessable** by default.

Amazon AppFlow

Fully managed integration service for **transferring data to and from SaaS vendors and applications**. Flows can be bi-directional between AWS services and SaaS applications.

Usecases:

- Transferring Salesforce records to Amazon Redshift
- Ingesting and analyzing Slack conversations in S3
- Migrating Zendesk and other help desk support tickets to Snowflake

Big Data

Redshift

Fully managed **data warehouse** service in the cloud. Its a very large **relational database** traditionally used in big data applications.

Supports MULTI and Single AZ deployments.

EMR (Elastic Map Reduce)

- EMR is made up of EC2 instances. This means you can employ your standard EC2 instance cost savings measures.

- Managed cluster platform that simplifies running big data frameworks, such as **Apache Hadoop and Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi, and Presto** on AWS to process and analyze vast amounts of data.

Kinesis

Only service with **real time response.** It is a bit more complicated to configure than SQS.

Kinesis Data Firehose:

Amazon Kinesis Data Firehose is a fully managed service designed **to simplify the process of loading streaming data into storage and analytic services**. It provides an efficient way to capture and deliver streaming data directly to other AWS services or external destinations. (**Near real-time**)

Kinesis Data Streams:

Amazon Kinesis Data Stream is a fully managed real-time data streaming service that enables developers **to collect, process, and analyze large volumes of data** in **real-time**.

Kinesis Data Analytics (Apache Flink):

allows you to process streaming data in real time using standard SQL.

When to choose SQS or Kinesis

SQS does not offer real-time message delivery. So, if your application **needs real-time choose Kinesis**. And kinesis mostly used in big data applications.

Athena

Serverless sql, can query on S3 objects.

Glue

Serverless ETL(Extract-Transform-Load) service

- You can use AWS Glue to export the data from DynamoDB, transform the data, and then load the data back into DynamoDB
- AWS Glue job is meant to be used for batch ETL data processing.

QuickSight

Creating dashboard for visualizing data.

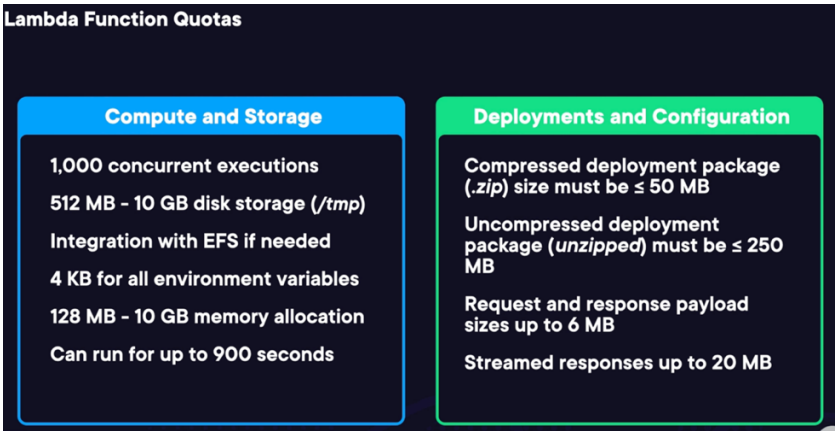
SERVERLESS

Lambda

- Whenever you are talking about credentials and Lambda, ensure you are attaching a role to the function.
- S3, Kinesis, Event Bridge **can trigger** lambda.
- Can allocate **up to 10GB of RAM** and **15 minutes** of runtime.

Limits and quotas;

- AWS Lambda currently supports **1000 concurrent executions** per AWS account per region.



Lambda Best Practices

- If you intend to reuse code in more than one AWS Lambda function, you should consider creating an AWS Lambda Layer for the reusable code.
- Since AWS Lambda functions can scale extremely quickly, it's a good idea to deploy a Amazon CloudWatch Alarm that notifies your team when function metrics such as ConcurrentExecutions or Invocations exceeds the expected threshold
- Lambda functions require you to package all needed dependencies (or attach a Layer) — the bigger your deployment package, the slower your function will cold-start. Remove all unnecessary items, such as documentation and unused libraries.

AWS Fargate

It does not work by itself. It requires ECS or EKS. Fargate allows you to run containers without using EC2 instances.

Amazon EventBridge

Think EventBridge **if you want to trigger an action based on something** that happened in AWS.
Common usecase is triggering Lambda functions when an AWS API call happens.
Successor of the CloudWatch Events.

Amazon EventBridge is recommended when you want to build an application that reacts to events from SaaS applications and/or AWS services. **Amazon EventBridge is the only event-based service that integrates directly with third-party SaaS partners.**

AWS X-RAY

Application insights: Collects data for gaining **insights to application requests and responses**.
Terms: Traces, tracing headers and segments.
Scenerios involving app request insights, viewing response times of downstream resources, and HTTP response anlysis.

AWS AppSync

AWS AppSync provides a robust, **scalable GraphQL interface for application developers** to combine data from multiple sources, including Amazon DynamoDB, AWS Lambda, and HTTP APIs.

SECURITY

CloudTrail

Remember that CloudTrail is basically just CCTV for your AWS Account. It **logs all API calls made to your account** and **stores these logs in S3**.

AWS Shield

Free DDOS protection that **protects against layer 3 and layer 4 only**. Protects all AWS customers on ELB, CloudFront, and Route 53.

AWS Shield Advanced: It is the advanced version of the Shield. It costs 3k usd per month.

AWS WAF (Web Application Firewall)

AWS WAF is a web application firewall that lets you monitor HTTP/HTTPS requests that are forwarded to CloudFront or an Application Load Balancer. **(Layer 7)**

WAF can **block Layer 7 DDoS attacks as well as things like SQL injection and cross-site scripting**. If you need to **block access to specific countries or ip address**, you can also do it.

Amazon GuardDuty

Uses AI to learn what normal behaviour looks like **in your account** and to alert you of **any abnormal or malicious** behaviour. **Monitors CloudTrail events, VPC Flow logs and DNS logs**.

AWS Firewall Manager

See a scenario about multiple AWS accounts and resources that need to be **secured centrally**.
Making sure that all your **firewall rules across these multiple accounts** and regions are consistent.
You can centrally configure AWS WAF rules, AWS Shield Advanced protection, VPC security groups, AWS Network Firewalls, and Amazon Route 53 Resolver DNS Firewall rules.

Note: It does not support Network ACL's as of today.

Amazon Macie

Uses AI to analyze data in S3 and helps **identify PII, PHI**, and finacial data.

Amazon Inspector

It is used to perform **vulnerability scans** on both **EC2 instances and VPCs**.

AWS KMS (Key Management Service) and CloudHSM

AWS KMS is a managed service that makes it easy for you to create and control the **encryption keys**.
You start using the service by requesting the creation of a CMK and you can control the lifecycle of the CMK as well as who can use or manage it.

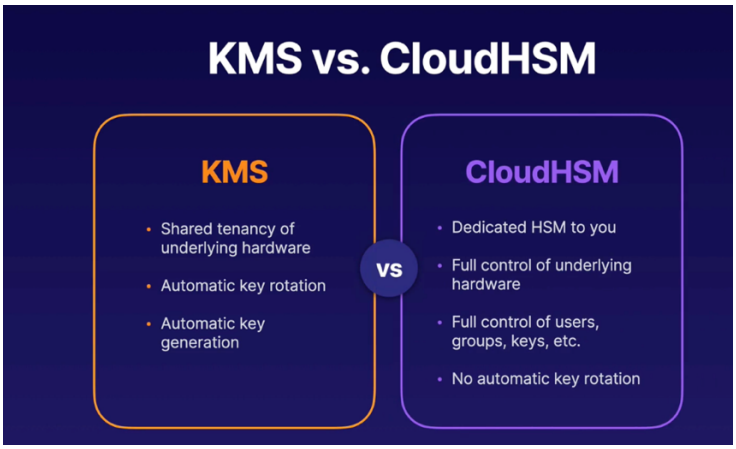
Note: Minimum length of time before you can schedule a KMS key to be deleted is **7 days.**

Ways to generate a CMK

1. AWS creates CMK for you. The key material for a CMK is generated within HSMs managed by AWS KMS.
2. Import key material from your own key management infra and associate it with a CMK.
3. Have the key material generated and used in a AWS CloudHSM cluster as part of the custom key store features in AWS KMS.

Ways to control permissions

1. Use the key policy.
2. Use combination of IAM policies and key policy.

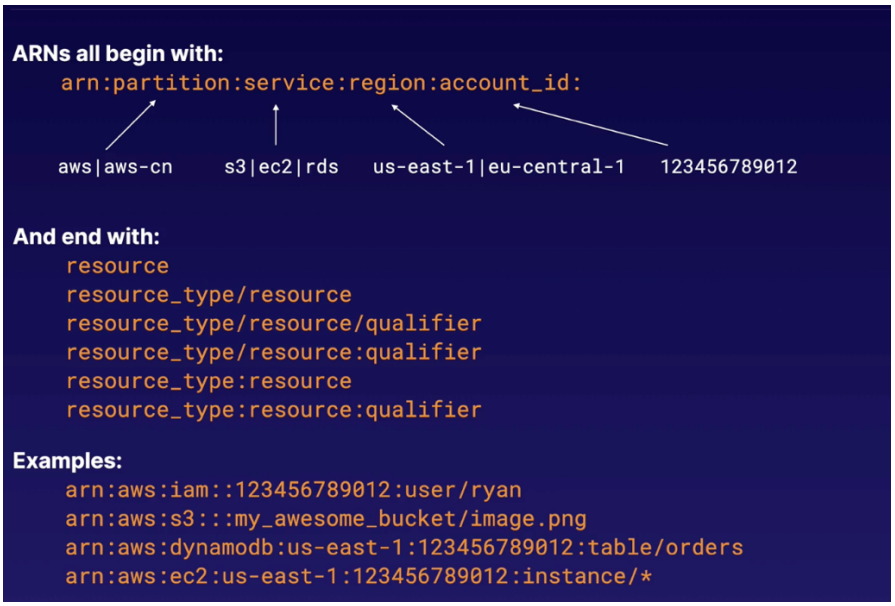


Secrets Manager

Secret manager can be used to securely store applications secrets. Applications get the secrets via API Calls.
It also provide us functionality for **rotating the credentials**. You can set up **automatic rotation** for your secrets.

Note: **Presigned URLs: Temporarly Sharing private objects** in the S3 Bucket**.**

AMAZON ARNs (Amazon Resource Names)



AWS Audit Manager

Think it when you see a scenario question about **HIPAA or GDPR compliance** that asks about continues auditing or automating.

AWS Artifact

Think it when you see a scenario question asking about **audits** and the need for **complience reports**.

AWS Network Firewall

AWS Network Firewall is a managed service that makes it easy to deploy **physical firewall protection** across your VPCs via its managed infrastructure (e.g., a physical firewall that is managed by AWS).

Amazon Cognito

User Pool: User directories that provide sign-in, sign-up for users to your application.
Identity Pool: Allows users to access other AWS services.

AWS Security Hub

Single place to view all your **security alerts** across multiple AWS security services (such as GuardDuty, Inspector, Macie, Firewall Manager) and accounts.

AUTOMATION

CloudFormation

AWS service that allows you to declare your AWS **infrastructure as code (IaC).** Support both **YAML and JSON** format. Each template definition is named as **stack** in cloudformation. Before applying stack, you can create **Change sets** to observe the changes without applying.

Parameters, mappings, resource sections, parameter store.

Template (YAML)

```yaml
AWSTemplateFormatVersion: "2010-09-09"
Description:
  this template does XXXX
Metadata:
  template metadata
Parameters:
  set of parameters
Mappings:
  set of mappings
Conditions:
```

```
  set of conditions
Transform:
  set of transforms
Resources:
  set of resources
Outputs:
  set of outputs
```

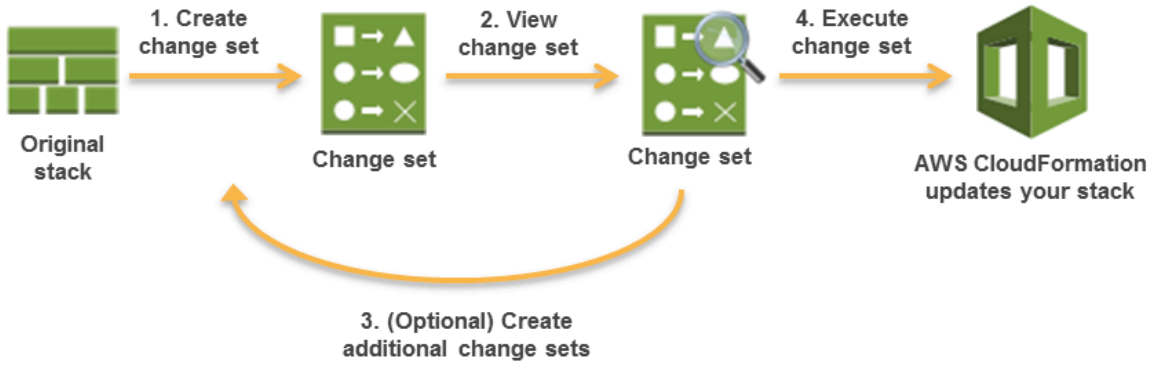


Figure 7 CloudFormation Change Set

Note: After you execute a change, CloudFormation removes all change sets that are associated with the stack because they aren't applicable to the updated stack.

Systems Manager

AWS Systems Manager is a secure end-to-end management solution for resources on AWS and in multicloud and hybrid environments.

With installing SSM Agent on your resources you can control them. You can manage their folder system, you can apply upgrades & updates, you can run commands and you can even see them in a fleet.

It is more secure cause you can connect them via System Manager directly which means you don't have to open any port on them for ssh or rdp.

OpsWorks

Configuration management of systems using **Chef or Puppet**.

CACHING

CloudFront

CloudFront is a fast **content delivery network (CDN)** that securely delivers data. It helps reduce latency and provide higher transfer speeds using **AWS edge locations.**

Important Settings

- Can be used to **front AWS endpoints as well as non AWS endpoints**.
- Defaults to https connections with the ability to add custom ssl certificate.
- You can't pick speficic countries, but you can pick continents which effects the price.
- You can force an expiration of content from the cache if you can't wait for the TTL.
- You can **restrict viewer access**. (geographical restrictions etc.)
- Amazon CloudFront **can route to multiple origins based on the content type.**
- You can set up Amazon CloudFront with **origin failover** for scenarios that require high availability with **origin groups**.
- **Field level encryption** to protect sensitive data for specific content.

CloudFront Signed URLS

Signed URLs are useful when you want to access **individual private files**.

CloudFront Signed Cookies

Provide access to **hundreds of private files** (multiple files) served by your CloudFront distribution.

CloudFront Functions

Functions can manipulate the requests and responses that flow through CloudFront, perform basic authentication and authorization, generate HTTP responses at the edge, and more.

CloudFront Lambda@Edge

Lambda@Edge is a feature of Amazon CloudFront that lets you run code closer to users of your application, which improves performance and reduces latency.

When the requests directly goes to the origin instead of edge caches?

- Proxy HTTP methods (PUT, POST, PATCH, OPTIONS, and DELETE) go directly to the origin from the POPs and do not proxy through the regional edge caches.
- Dynamic requests, as determined at request time, do not flow through regional edge caches, but go directly to the origin.
- When the origin is an Amazon S3 bucket and the request's optimal regional edge cache is in the same AWS Region as the S3 bucket, the POP skips the regional edge cache and goes directly to the S3 bucket.

ElastiCache

ElastiCache is a managed version of 2 open source projects: **Redis and Memcached.**

MEMCACHED	REDIS
Simple database caching solution.	Supported as a caching solution.
No backup support	Support backups

No failover or Multi-AZ Support	Failover and Multi-AZ Support
Not a database itself	Can be used as a standalone database
Stores key-value pairs as a String and has a 1MB size limit per value	Supports data structures like list, set, and hash, and can store values of up to 512MB in size
Multi-threaded	Not Multi-threaded

DynamoDB Accelerator (DAX)

DAX is specific to DynamoDB and can only be used with it.

- **In-memory cache.** It reduces dynamodb response times from milliseconds to microseconds.
- This cache is highly available and lives inside the VPC you specify.

Global Accelerator

AWS Global Accelerator is a networking service that sends your user's traffic through AWS's global network infrastructure via accelerators. It can increase performance and help deal with **IP caching.**

Note: Meant for **TCP or UDP traffic. Major difference from CloudFront.**

GOVERNANCE

AWS Organizations

Its a service that helps you to **manage multiple AWS accounts from a centralized place**.

Management Account (Payer Account): Primary account that hosts and manages the organization.
Member Account: All the other accounts in the organization.

Note: Accounts can be migrated between organizations. **Steps do to it;** Remove the member account from the old organization. Send an invite to the member account from the new Organization. Accept the invite to the new organization from the member account.

Features

Consolidated billing: Collects all the bills up to payer account to pay with single payment method.
Share: Reserved Instances and Shared Plans can be shared across accounts in the organization.
Tag enforcement: capability to require specific tags for all aws resources. Force users to use tags.
Organization Unit (OU): Logical grouping of the accounts.
Service Control Policies(SCPs): Policies that get applied to the organization units or acounts to restrict actions.

Note:

- A service control policy (SCP) will take precedent over any other permissions.
- Service control policy (SCP) does not affect service-linked role.
- Service control policy (SCP) **affects all users and roles** in the member accounts, **including root user** of the member accounts.

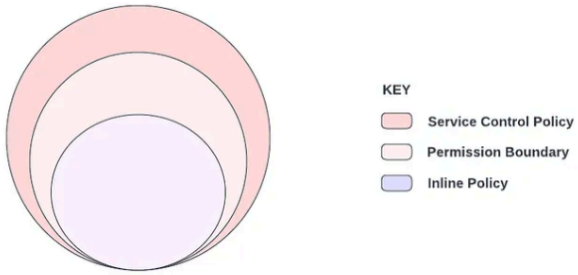
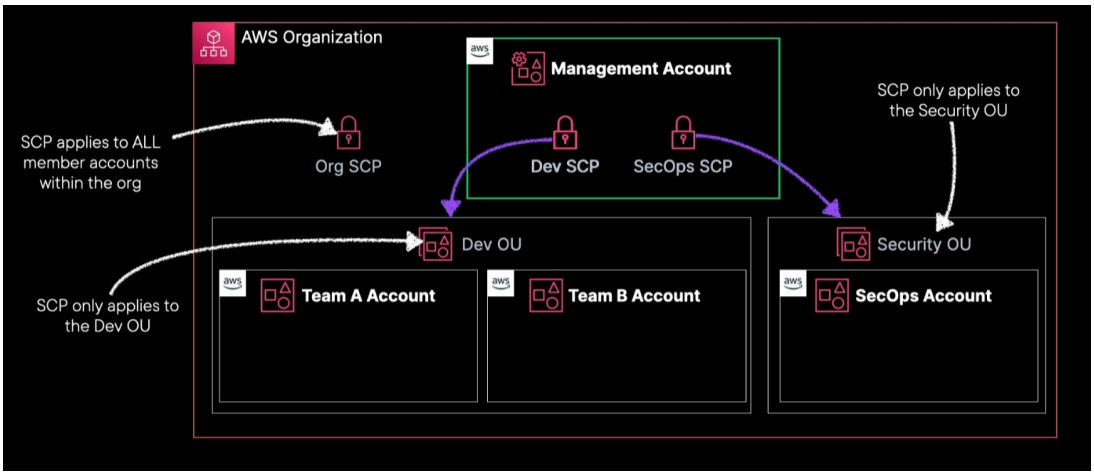


Figure 8 SCP, Permission Boundary, Inline Policy



Note: Its important to know you can assign a single account to hold your CloudTrail logs.

AWS Resource Access Manager (RAM)

Thanks to this service, you can share resources with other accounts.
Common resources to share; Transit Gateways, VPC subnets, License Manager, Route53, Dedicated Hosts etc.

Note: You pay for the resources that are being shared.

AWS Config

Provide visibility (monitoring) and alerting (via SNS) for resource config changes. It does not prevent the changes.

Should be enabled for per region that you want to keep the records for the resources inside the region.

Note: AWS Config is the best way to check what standards are applied to your architecture.

AWS Directory Service

Managed Microsoft AD: This is entire AD suite. You can easily build out AD in AWS.
AD Connector: Creates a tunnel between AWS and your on-premises AD.

AWS Cost Explorer

Its a tool that allows you to **visualize and analyze your cloud costs**.

You can **generate custom reports** based on variety of factors, including resource tags.
Built-in forecasting up to 12 months.

Not only offers historical records but also **creates forecasts and savings recommendations.**

Cost And Usage Reports (CUR)

The AWS Cost and Usage Reports (AWS CUR) contains the most comprehensive set of cost and usage data available. You can use Cost and Usage Reports to publish your AWS billing reports to an Amazon Simple Storage Service (Amazon S3) bucket that you own.

Usecases:

- What if you want to know your AWS cost from 2 years ago?
- What if your boss want to view your AWS cost from PowerBI/Tableau/QuickSight? (Of course, usually your boss does not want to log in to an AWS account which is not familiar with him/her)
- Thanks automatic refresh, AWS updates the report in the budget at least once a day.
- Integration of the report data with Athena, Redshift and QuickSight.

AWS Budgets

Budgets are the best way to **let users know** when they are getting close to **overspending**.
You can use Cost explorer to create fine-grained budgets.

AWS Compute Optimizer

Provides **rightsizing recommendations** based on collected utilization and configuration metrics for EC2, Auto Scaling Groups, Lambda and EBS. Rightsize workloads according to your workload preferences through artificial intelligence and machine learning-based analytics to reduce costs by up to 25%.

Savings Plans

Similar to reserved instances, offers flexible pricing on compute usage. However, **Savings Plans are applicable to EC2, Lambda and Fargate instances**, RIs are only applicable to EC2 instances.

One-year and three-year agreements. And you can pay All Upfront, Partial Upfront or No Upfront.

AWS Trusted Advisor

It provides **recommendition based on best practices** for **cost, security, performance, fault tolerance and service limits.**

Free to use, however you will need a business or enterpise support plan.

AWS Control Tower

Automated multi-account governance, guardrails, account orchestration, governed user account provisioning.

AWS Health

Questions about service alerts or notifications of EC2 **hardware maintanence** reboots.

MIGRATION

AWS Snow Family

Physically transferring data to AWS.

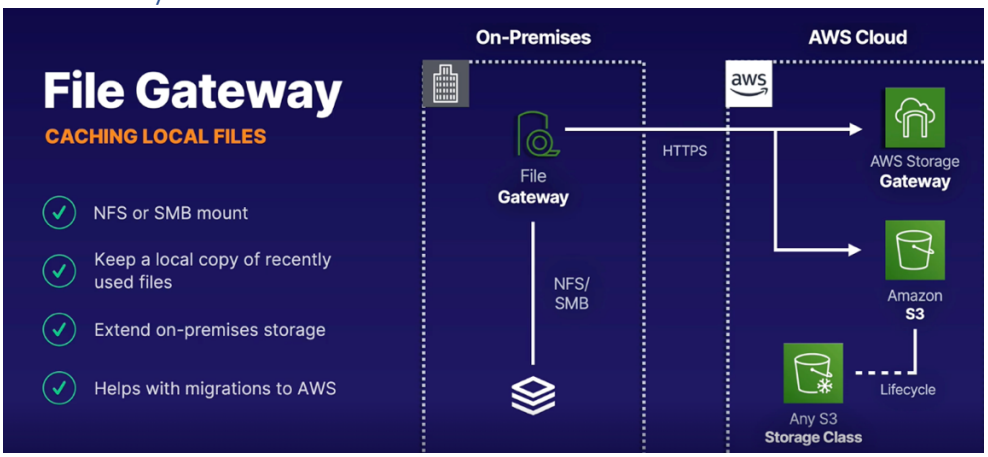
Snowcone: Up to 8TB of data.
Snowball Edge: Up to 80TB of data. (Has two option; Compute Optimized, Storage Optimized)
SnowMobile: Up to 100PB of data. (Choose if the data is greater than 10PB+)

Note: You can't move data directly from AWS Snowball into a Amazon S3 Glacier Vault or a Glacier Deep Archive Vault. You need to go through Amazon S3 first and then use a lifecycle policy.

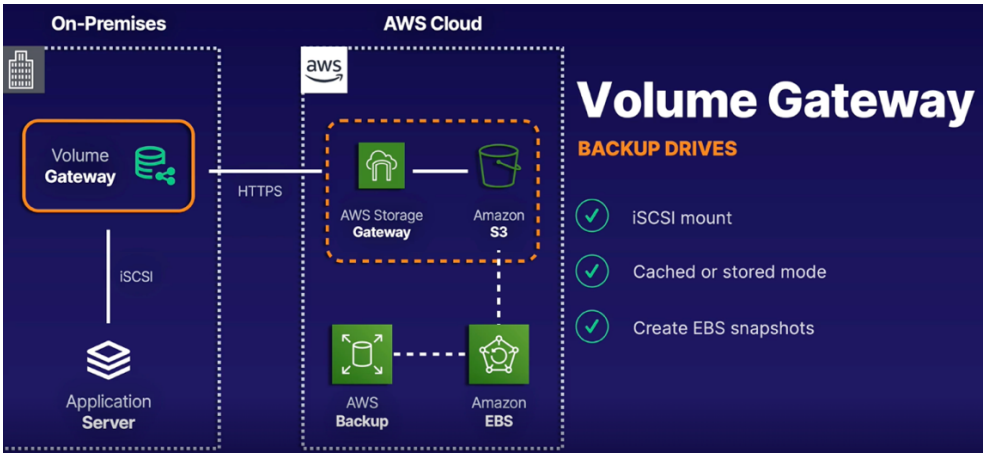
Storage Gateway

Its a hybrid storage for data migration.

File Gateway:



Volume Gateway:



Tape Gateway:

It's suitable for backup and archiving use cases.

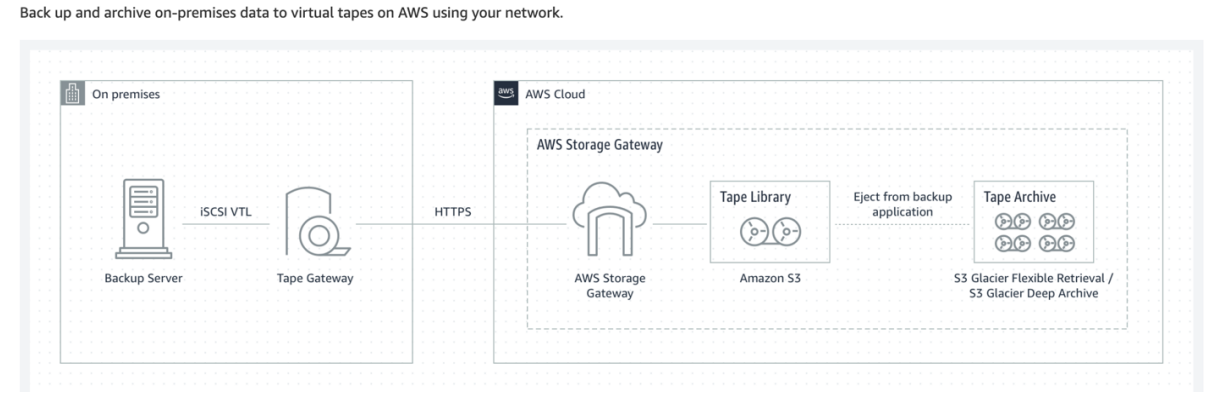


Figure 9 Tape Gateway

AWS DataSync

AWS DataSync can be used to **move large amounts of data online between on-premises storage and Amazon S3, Amazon Elastic File System (Amazon EFS), Amazon FSx for Windows**. DataSync eliminates or automatically handles many of these tasks, including scripting copy jobs, scheduling and monitoring transfers, validating data, and optimizing network utilization. The source datastore can be Server Message Block (SMB) file servers.

Note: AWS recommends that you **should use AWS DataSync to migrate existing data** to Amazon S3, and subsequently **use the AWS Storage Gateway** to retain access to the migrated data and **for ongoing updates** from your on-premises applications.

AWS Transfer Family

AWS Transfer Family securely scales your recurring business-to-business file transfers to AWS Storage services using SFTP, FTPS, FTP, and AS2 protocols.

AWS Migration Hub

AWS Migration Hub is a singular place to discover existing servers, plan your migration efforts, and track migration statuses.

Database Migration Service (DMS)

Migrating on-premise database to AWS Cloud. It also allow us to migrate databases with different engines (exp: oracle to postgresql). However, for migrating different engine types, you will need **Schema conversion tool (SCT).**

Device Farms: The service enables you to run your tests concurrently on **multiple desktop browsers or real mobile devices** to speed up the execution of your test suite, and generates videos and logs to help you quickly identify issues with your app.

Pinpoint:

Amazon Pinpoint allows users to easily engage millions of customers via different communication channels. **Group customers based on specific criteria** for targeted messaging. The service also offers the ability to leverage **machine learning models to better understand customer interactions** for future engagements.

MACHINE LEARNING

Rekognition: Image and video analysis.

Comprehend: Natural language processing service.

SageMaker: Build, train and deploy ML models.

Amazon Transcribe:

Amazon Transcribe **converts speech to text** automatically. You can use this service to generate subtitles.

AWS Translate:

Amazon Translate is a neural machine translation service that delivers fast, high-quality, affordable, and customizable **language translation**.

Amazon Detective:

Amazon Detective helps you analyze, investigate, and quickly identify the **root cause of security findings or suspicious activities**. Detective automatically collects log data from your AWS resources.

Amazon Lex: Amazon Lex is an AWS service for building conversational interfaces for applications using voice and text. With Amazon Lex, the same conversational engine that powers Amazon Alexa is now available to any developer, enabling you to build sophisticated, **natural language chatbots** into your new and existing applications.

Amazon Polly:

Amazon Polly uses deep learning technologies to synthesize natural-sounding human speech, so you can **convert articles to speech.**

Amazon Textract:

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

Amazon Forecast:

Amazon Forecast is a time-series forecasting service based on machine learning (ML) and built for business metrics analysis.

AWS STS:

AWS provides AWS Security Token Service (AWS STS) as a web service that enables you to request temporary, limited-privilege credentials for users. This guide describes the AWS STS API. For more information, see Temporary Security Credentials in the IAM User Guide.