Terminology in AWS

- **Region:** Amazon cloud computing resources are hosted in multiple locations world-wide. Each AWS Region is a separate geographical area where the cloud resources reside. They are usually in different cities.
  Ex: US East (Ohio, N. Virginia), Asia Pacific (HK, Mumbai)

- **Availability Zone:** Each ~~Area~~ AWS region has multiple, isolated locations known as Availability Zone. If one AZ goes down & if your application is highly available, it can be served from another AZ in that region.

- **Edge Location:** It is where end users access services located at AWS. They are located in most of the major cities around the world & are specifically used by CloudFront to distribute content to end users to reduce latency.

- Security & compliance is a shared responsibility between AWS & the customers. AWS operates, manages, & controls the components from the host OS & virtualization layer down to the physical security of the facilities.

The customer assumes responsibility & management of the guest OS (updates & security patches), other associated application software as well as the configuration of AWS provided security group firewall.

Customer
- Customer Data
- Platform, application, IAM
- OS, Network config

AWS
- Software
- compute, storage, DB
- hardware, region, AZ

EC2

Elastic Compute Cloud

- Amazon EC2 is a web service that provides resizable compute capacity in the cloud.
- Elastic means the capability to instantly scale to meet traffic spikes on demand.
- It takes from 2-10 minutes to get an EC2 instance ready
- EC2 provides various instance types depending on the size, no. of CPU, GPUs etc. They are general purpose, compute optimized, memory optimized, accelerated computing & many more.
- In EC2, developers have full control of the hardware that they are using.
- EC2 supports many OS including Amazon linux, Microsoft server, centos, debian etc.
- Pricing model : EC2 provides various types of instances that can be used for different use cases. The instances can be:
  - → On-demand: you pay for the compute capacity on an hourly bases. No long-term commitments or upfront payments.
  - → Spot instances: spare instances that can be used at up to 90% off on the on-demand instances. Users can bid on the instances. These instances can be terminated anytime so they cannot be used for critical

applications.

→ reserved instances: they provide discounts of upto 75% on the on-demand instances. They provide a capacity reservation and ability to launch anytime. Customers can commit to use instances over 1 to 3 year term.

→ dedicated hosts: It is a physical EC2 server dedicated to you. No one else can share the hardware with you. You can use your existing server-bound software licenses & can also help you meed compliance requirements.

- Instances can be launched with a public IP or a private IP. Public IPs can be accessible by all whereas private IPs can only be used within a private network.
- EC2 instanes are launched with a security group that controls the access of that instance. It allows roles about what ports & IPs are open to use.
- One Elastic IP can be attached to an instance for free & an account can have a max of 5 Elastic IPs w/o charge. Elastic IP is a static IP that doesn't change on stopping & restarting an instance. The public IP of an instance may change in this case.

- EC2 instances can be stored as Amazon Machine Images (AMI). AMIs store the configurations of an instance & new instances can be launched from the AMIs with the same configurations.
- Placement groups: you can use this to influence the placement of your instance to meet your workload needs. There are 3 placement groups:
  - → cluster: packs instances close together inside a single Availability zone. This is for low-latency network performance & HPC applications.
  - → partition: spreads instances across logical partitions such that instances in one partition do not share underlying hardware with other partition. For distributed workloads such as Hadoop, Cassandra etc.
  - → Spread: instances are spread across to reduce correlated hardware failures.
- Some instances come with Instance Store volumes attached as root storage. These volumes provide very high IOPS as they are physically connected to the instances. But they provide ephemeral storage, ie. the data is lost as soon as the instances

are stopped.
You can specify the volumes only at the time of the launch and can't detach or attach the instance store volumes.

- Amazon Elastic Block Storage (EBS) however, provides persistent block storage to be used with an instance. The data is not lost when the instance is stopped or terminated, & these volumes can be detached or attached any time.
- EC2 Hibernate options allow storing the RAM contents of the instance into the attached volume. This allows faster re-booting of instances.
- Hibernate doesn't have any additional cost, only the data stored in the EBS volumes is charged. Although, when the instance is in the 'stopping' state, the instance charges apply.
- To allow EC2 instance to have a DNS hostname, 'DNS resolution' & 'DNS hostname' of the VPC config should be enabled.
- A golden AMI is an AMI that you standardize through configuration, consistent security patching, and hardening.

EBS

Elastic Block Storage

- EBS provides a block level storage that can be used with Ec2 instances.
- These volumes can be attached/detached to the instances & the data can persist.
- The volumes configuration ie. capacity, type, IOPS etc. can be dynamically changed w/o requiring to provision a new volume.
- Volumes are launched in an AZ & can only be attached to instances in that AZ.
- To make the volume available outside a region, you need to create a snapshot of the volume, copy it to another region, & restore the volume in that region.
- Snapshots are created asynchronously, so while creating a snapshot, the volume can be read/written simultaneously.
- The volumes can be defined as encrypted/unencrypted at the time of their creation.
- Unencrypted volumes can be encrypted by creating a snapshot of the volume & copying it to an encrypted snapshot & then creating a volume from the encrypted snapshot.

- To enable encryption in transit for an EBS volume, detach it from the instance & reconnect selecting the option of encryption in transit.
- EBS volume is by default replicated across multiple servers inside an AZ to prevent loss of data.
- EBS volume types:
  - → General purpose (gp2): provides standard IOPS & throughput on an SSD. The IOPS change at 3 IOPS per GB with min of 100 & max of 3000 depending on the size of volume. Throughput remain constant
  - → Provisioned IOPS (io1): High performance SSD volume for low-latency applications. IOPS available are from 100 to 64000 based on the volume size. Maintain ratio of 50 IOPS per GB.
  - → Throughput optimized (st1): low cost HDD volume for frequently accessed, high throughput workloads. Can get throughput from till 500 MB/s
  - → Cold HDD (sc1): Low cost HDD for less frequently accessed workloads.
- An EBS volume can only be attached to 1 instance at a time but an instance can be attached to multiple EBS volumes.

- Amazon Data lifecycle Manager (DLM) can be used to automate the creation, retention, & deletion of snapshots to back up EBS volumes.
- EBS volumes can also be ~~use~~ used as RAID but that is done only through software & is not backed by Aws.
  - → RAID 0 : Data is divided blw 2 volumes. This provides higher IOPS but no fault tolerance
  - → RAID 1 : Data is replicated across 2 volumes. This provides no IOPS enhancements but increases fault tolerance.
- * As Aws supports creation of snapshots for fault tolerance rather than using RAID.

EFS

Elastic File Storage

- Unlike EBS volumes, EFS volumes can be shared across instances.
- It can be shared across 1000s of clients & can grow to Petabyte scale.
- Efs have 2 storage classes : standard storage & infrequent access storage class
- Lifecycle managent roles can be applied to automatically move data ~~from~~ b/w the two storage classes.
- Performance modes : general purpose & max I/O performance mode.
- Throughput mode : Bursting & provisioned throughput modes.
- They can only be used with Posix file systems (Linux).
- For windows, Aws has another service known as FSx for windows.

*FSx lustre is also available ~~for~~ for Linux. (HPC)

* EFS can only be used with 1 VPC at a time

ELB

Elastic Load Balancer

- ELB automatically distributes traffic across multiple targets like EC2 instances, IP addresses, Lambda functions, ~~&~~ ECS.
- It can handle varying load across multiple AZs.
- Types:

→ Application Load Balancer (ALB) - suited for load balancing of HTTP & HTTPS traffic. It supports path-based routing, host-based routing & can support hosting multiple SSL certificates through Server Name Indication (SNI).

- ALB can perform health check on the target before routing the traffic.
- It supports Aws Global Accelerator & for increasing the availability & performance and AWS Web Application Firewall (WAF) to prevent against common web exploits.
- ALB is highly available & uses DNS for IP resolution.

→ Network Load Balancer (NLB) - It is same as ALB but works on the TCP level.

- Used for high performance applications that require millions of connections with low latency.
- Cross-zone LB is chargeable.

fixed static

- NLB exposes 1 IP addresses unlike DNS hostname in case of ALB.

→ Classic Load Balancer (CLB) - old generation load balancer which is not backed by Aws.
- can only serve 1 SSL certificate.
- Cannot perform intelligent routing like ALB or NLB.
- Cross-zone LB is enabled by default.

• Sticky sessions: allows user to bind to a specific target (instance/IP) through web cookies. Request subsequent requests are sent to the same target.

• X-forwarded-for: used for getting the public IP of user

• ELB can be Internet-facing (public IPs) or Internal (private IPs).

ASG

Auto Scaling Group

- ASG monitors applications (instances) & automatically adjusts capacity to maintain steady, predictable performance.
- ASG works on certain parameters
  - → min capacity: the minimum no. of instances required for an application.
  - → desired capacity: the desired capacity that can run by default.
  - → max capacity: the max. no. of instances that the ASG can provision.
- The size of ASG depends on the desired capacity. You can adjust its size to meet demand & can change it manually or through automatic scaling.
- ASG monitors instances through default health checks & if it becomes unhealthy - the instance is replaced by a new one.
- Scaling policy adjusts the scaling within ASG either based on
  - → Manually: based on desired capacity
  - → Scheduled: based on date or time.
  - → Demand: based on CPU utilization, memory needs etc.

- You can create custom cloudwatch metric & alarms to trigger ASG.
- The instances are launched based on launch config configuration or launch templates.
  - Launch configurations are old way of defining the instance's config. They don't support versioning & requires creation of new config every time.
  - Launch templates are new types which support versioning. Once a new version is available, ASG automatically used that version. It also allows option to use spot instance fleet along with on demand instances to reduce cost.
- ~~Sta~~ Scaling in of instances follows this criteria.
  - If multiple AZs have got multiple instances, the instance from AZ having more no. of instances is terminated.
  - When multiple config of instances are available, the instance having the oldest configuration is terminated.
  - ~~Otherwise, the instance is terminated randomly.~~
  - Select the instance that is closest to the next billing hour.

- Dynamic scaling in ASG
  - → Target tracking - you select a scaling metric & set a target value. Ex: keep average CPU utilization of ASG at 40%. (or) keep request count per target at 1000. etc.
  - → Step & simple scaling - you choose a scaling metric & set a threshold value. Ex: CPU utilization should be b/w 20 & 80%

- warm-up: the time taken by an instance to boot. Until the warm-up time has expired, an instance is not counted towards the scaling metric.
- cool-down: helps to prevent ASG from launching/terminating additional instances before the effect of previous activity is visible. defaults to 300 sec.
- connection-draining: time to complete inflight requests while the instance is unhealthy or de-registering. No further requests are sent by ELB to this instance during this time.

- You can add lifecycle hooks to your ASG so that you can perform custom actions when instances launch or terminate.

IAM

Identity and Access Management

- IAM enables you to manage AWS services & resources securely.
- Using IAM, one can create & manage AWS users & groups, and use permissions to allow & deny their access to AWS resources.
- IAM roles are applied globally & not specific to a region.
- Terminology
  - → users: people, employees, organizations
  - → Group: group of people, employees, users.
  - → policies: certain permission for allowing the access to certain things in AWS.
  - → roles: A set of policies which can be given to a user or AWS service which allows them to access the other AWS resources.
- Features
  - → Centralized control of AWS account
  - → Shared access to AWS resources
  - → Multifactor Authentication (MFA)
  - → Temporary access to users
  - → Password rotation policy
  - → Give permissions in the form of policies.

- IAM roles for EC2 instances simplifies management & deployment of AWS access keys to EC2 instances. Instead of storing access keys in EC2, give it a role using IAM to access certain AWS resource.

- AWS Security Token service (STS) is a web service that enables you to request temporary limited-privilege credentials for AWS IAM users.
- You can call 'GetFederationToken', 'AssumeRole', 'AssumeRoleWithSAML', or 'AssumeRoleWithWebIdentity' API. for requesting temp credentials for federated users.
- IAM users can request temp credentials for their own use by calling 'GetSessionToken' API for STS.

- IAM also supports identity Federation. These external identities like Microsoft Active Directory, Amazon Cognito, Login with Amazon, Facebook, Google or any OpenID connect can be used to grant secure access to AWS resources.

S3.

Simple Storage Service

- Amazon S3 is an object storage service that offers scalability, data availability, security, & performance.
- It can be used to store & protect any amount of data for a range of use cases, such as websites, mobile applications, backup & restore, archive, enterprise application, IoT devices, & big data analytics.
- AWS can be used to backup data already in AWS cloud or use AWS Storage Gateway, a hybrid storage service, to send backups of on-premises data to AWS.
- S3 can provide object storage from 0 to 5TB, whose a maximum of object can be of size 5GB. It also supports multi-part upload, which uploads a file to S3 in small chunks of data parallely.
- The objects are stored in buckets, the whose name of which ~~see~~ should be globally unique.
- S3 allows cross-region replication to replicate objects into AWS regions for reduced latency, disaster recovery etc. CRR can be set on bucket level, shared prefix level, or an object level.
- S3 also allows same-region replication which is same as CRR but within same region. You can use SRR to change account ownership for the replicated objects to protect accidental deletion.
- For both CRR & SRR, owners can be same or different.

- Storage classes
  - → S3 standard : for general-purpose storage of frequently accessed data.
  - → S3 intelligent-tiering : to intelligently move data through different storage classes based on their usage.
  - → S3 Standard-IA : for infrequently used data
  - → S3 One zone-IA : for infrequently used data but only served from 1AZ. Therefore, is not fault tolerant.
  - → S3 Glacier : For archival storage. The retrieval time can be from few minutes to hours.
  - → S3 Glacier Deep Archive : lowest-cost storage & supports long-term retention. Retrieval time can be several hours.

- Add lifecycle management rules to your S3 bucket to automatically move objects across storage classes based on your usage. The rules need to be defined by the owner of the bucket.
- S3 allows read after write consistency for new PUTS & eventual consistency for overwrite PUTS & DELETES.
- S3 charges for storage capacity, requests (read/write) data transfer, transfer acceleration, cross-region replication, & storage management.

- S3 transfer Acceleration can speed up content transfer to & from S3 by as much as 50-500% for long distance transfer of longer objects.
- S3 has a flat structure of file storage & doesn't have folders. Each object though, has key-name Ex: photos/myphoto.jpg. The objects are retrieved by these key names.
- To support grouping of objects, S3 has prefixes. Ex: for "photos/myphoto.jpg", "photos/" is the prefix which acts as a logical folder.
- S3 supports versioning of files, when files are uploaded instead of overwriting, a new version is created.
- When a version is deleted, only a delete marker is placed over the file, but the actual file still resides.
- During replication, versioning should be enabled & the delete markers are not replicated.
- MFA deletes can be enabled.
- Max 100 buckets are allowed per account by default.
- Encryption in S3
  - → SSL/TLS is allowed during transit
  - → At rest, default AES-256 encryption or Through KMS using Aws managed or customer managed keys.

- S3 server access logging provides detailed records for the requests that are made to a bucket. They can be useful in security & access audits & also to learn more about customer base.
- The logs are itself stored in an S3 bucket.
- An access log record contains details ~~about~~ about the requests that are made to the bucket, which includes request-type, resources, & time & date of access, turn-around time, & many more.

- S3 Select & Glacier Select :

→ Allows server-side filtering of data using SQL queries.

→ The queries are applied to S3 & only the filtered data is returned back.

- S3 event notifications can be created which gets triggered when a certain action is performed on the bucket like put, delete, update etc. The triggers can be handled by lambda function, SQS, or SNS.

- Byte-range fetch : S3 allows getting small chunks of data from large objects stored in a bucket. No need to fetch the whole object.

- S3 Glacier vault is a way to group archives together. Vault-level access policies can be used to control the access to the vault.
- Vault locks allows you to easily deploy & enforce compliance controls. Once locked, the vault-lock policy becomes immutable. Ex: you can set up a vault lock to protect deletion of an archive for 1 year.

- A pre-signed S3 URL can be used to access an S3 resource without requiring the security credentials & permission. The URL usually last a limited time & have limited permissions on the resources.
- A pre-signed cookie however, gives limited time access to multiple files in S3.

- Data cannot be uploaded to Glacier directly, it has to go to S3 & then transferred to Glacier.

- You can enable Cross-origin resource sharing (CORS) for web application in different domain to interact with resources in the other domains.

VPC

Virtual Private Cloud

- Amazon VPC lets you provision a logically isolated section of the AWS cloud where you can launch AWS resources.
- VPC provides control over the IP addresses, subnets, route tables & other security related features of AWS.
- AWS provides a default VPC where all the resources are launched by default.
- Features:

→ Create a VPC by specifying the private IP address range (CIDR)

→ Expand your VPC by adding secondary IP ranges.

→ Divide your VPC IP to one or more public or private subnets depending on their access.

→ Control inbound & outbound access to your resources by NACL & security groups.

→ Connect your VPC to other VPCs (same account/ other account / on-premise)

→ Privately connect to other ~~se~~ AWS services without using public internet.

→ Use VPC flow logs to log information about network traffic going in & out of the VPC.

- VPC limits
  - → upto 5 non default VPCs per Aws account per region.
  - → upto 4 secondary IP ranges per VPC.
  - → upto 200 subnets per VPC.
  - → upto 5 Elastic IP per Aws account per region.
  - * These limits can be increased by contacting Aws.

- VPC cannot ~~span~~ span across multiple region.
- A subnet is a range of IP addresses in a VPC. You can launch Aws resources in a subnet. A public subnet is used for resources that can be connected to the Internet, while a private subnet is used for resources that won't be connected to the Internet.
- The security of subnets can be controlled by security groups & network access control lists (NACL)
- Subnets are tied to an AZ, cannot span multiple AZs.
- A route table contains a set of rules (routes) that are used to determine where network traffic from your VPC is directed. By default, every subnet is associated to the main route table (created by default in a VPC)

- Each route table specifies the range of IP addresses where the traffic will go from the gateway, network interface, or other connection.
- While creating a new VPC, route to the Internet gateway should be added manually to access the public Internet.

- The resources within a subnet, connect through the public Internet ~~door~~ through an Internet gateway.

- A security group acts as a virtual firewall for your instance to control inbound & outbound traffic. You can assign upto 5 security groups to the instance.
- Security groups are instance level & therefore, each instance in your subnet can be assigned to a different security group.
- A security group is defined by its inbound & outbound rules.
  - → inbound: the traffic that comes from outside to the instance.
  - → outbound: the traffic that ~~comes~~ goes from the instance to the Internet

- A security groups can only give allows rules & not deny rules.
- Security groups are stateful - If the traffic is allowed inside, it will also go outside regardless of the outbound role & vice versa.
- By default, in a new security group, all outbound traffic is allowed but no inbound traffic is allowed.
- security groups are attached to the network interface of your instance.
- A role is defined by giving the source/destination & the protocol & port range.
- All the inbound/or outbound roles are evaluated before allowing the traffic unlike NACL.
- Any addition/deletion to the security group is in effect immediately, without delay or restart.

- A network access control list (NACL) is an optional layer of security for a VPC that acts as a firewall for controlling traffic in & out of multiple subnets.
- A default NACL comes with a ~~def~~ new VPC & allows all inbound & outbound traffic.
- A newly created NACL however, denies all inbound & outbound traffic

- An NACL can be connected to multiple subnets but a subnet can only be connected to to 1 NACL.
- NACL is stateless - for traffic to pass in & out of the NACL, both inbound & outbound rules should agree.
- A rule can either ALLOW or DENY.
- NACL contains a numbered list of rules, the rules are evaluated in order from lowest to highest. As soon as a rule is ~~matched~~ matched, the subsequent rules (having higher numbers) are skipped.
- The NACL rules are applied immediately.

- AWS allows usage of NAT gateway for instances in private subnets to communicate to the internet.
- NAT devices can only ~~work~~ work for IPv4 traffic. For IPv6, use egress-only Internet Gateway
- NAT devices can be used in two ways- NAT gateway & NAT instance. NAT instance needs to be managed by the user & is therefore not a preferred way.
- NAT gateways are fully managed by AWS & are highly available.
- A NAT gateway cannot be shared across VPCs.

- NAT gateways must be in the public subnet with a route to the private subnet.
- For creating a NAT gateway, it should be associated with an Elastic IP.
- NAT gateways do not have a security group.

- Bastion host is an instance which can be used to SSH to the private instances. It should be placed in a public subnet with a route to the private subnet.
- The security group of the bastion host should be toughened to provide only the access only to port 22 to restricted IP addresses.

- VPC peering allows connection b/w two VPC's & can work inter-region & cross account.
- The paths needs to be added to the route table for peering to work.
- VPC peering isn't transitive
- The CIDR blocks of VPCs to be peered should not overlap.

- VPC endpoint enables you to privately connet your VPC to supported AWS services without requiring the public internet.

- VPC endpoints route the request through the AWS backbone network w/o requiring Internet Gateway, NAT devices, or ~~the~~ VPN connection.
- Endpoint types
  - → Gateway endpoint: AWS S3 & DynamoDB
  - → Interface endpoint: they are powered by AWS Private link. Includes API Gateway, Amazon Athena, CloudFormation, CloudTrail, SNS, SQS etc.

- AWS Private Link / VPC Endpoint Services provides a secure way to privately connect to services w/o requiring Internet Gateway or NAT devices.
- It requires a Network load balancer at the service VPC & an Elastic Network Interface at the Customer VPC.
- Private Link can be used by ~~aws~~ 1000s of customer VPCs to connect to your VPC w/o requiring VPC peering.

- AWS Direct Connect links your internal network to an AWS Direct Connect location over a standard Ethernet fibre-optic cable. With this connection, you can create virtual interfaces directly to public AWS services (S3) or to Amazon VPC, bypassing the Internet service

provider in your network path.
- Direct connect location provides access to Aws in the region with which it is associated.

- If you want to setup a direct connect to one or more VPC in many different regions, you must use a Direct Connect Gateway.

- Aws Virtual Private Network is used to establish connection between on-premise network & Aws global network. It is comprised of two services: Aws site-to-site VPN & Aws client VPN.
  - → Site-to-site VPN creates a secure connection b/w your data center & Aws cloud resources.
  - → Client VPN is a fully-managed, elastic VPN service that automatically scales up or down based on user demand.
- Site-to-site VPN connection requires a customer gateway at the client side & VPN gateway attached to the Aws VPC.
- VPN connection is not completely private & shares the Internet service provider network.

- Transit Gateway provide a hub-and-spoke (star) connection b/w thousands of VPC & on-premise connections.

- The resources can work cross-region & cross-account (using Resource Access Manager)
- Route table limits which VPC can talk to other VPC.
- It works with Direct Connect Gateway & VPN connections.

- If you have multiple Aws site-to-site VPN connections, you can provide secure connection b/w sites using Aws VPN CloudHub.
- This enables your remote sites to communicate with each other & not just with the VPC.

- VPC Gateway endpoints (S3, dynamoDB) are not supported outside VPC & therefore, Direct Connect, VPC peering, VPN connection cannot use the endpoint to communicate with the resources.

- Networking cost in Aws.
→ Traffic within an AZ is completely free.
→ Traffic within multiple AZs is chargeable for public IPs & the rates are half for private IPs.
→ traffic b/w different region is chargeable using both public & private IP.

→ therefore, to reduce the cost of network usage, either use private IP b/w AZs or keep the resources within same AZ.

- VPC Flow logs is a feature that enables you to capture information about the IP traffic going to & from network interfaces in your VPC.
- Flow log data can be published to Amazon cloudwatch logs or Amazon S3.

Route 53

- Amazon Route 53 is a highly available & scalable cloud Domain Name System (DNS) web service.
- It is designed to route end users to Internet applications by translating name like www.example.com into numeric IP addresses like 192.0.2.1 that computer can connect to.
- You can use Route 53 to configure DNS health checks to route traffic to healthy endpoints.
- There are various routing policies that you can use.

→ Simple routing : use a single resource that performs a given function for your domain.

→ failover : use when you want to perform active-passive failover. ~~when~~ Traffic is sent to one endpoint, if that fails to pass the health checks, the traffic is switched to the standby endpoint.

→ Geolocation : route traffic based on user's location.

→ latency : route traffic to the destination that gives least latency.

→ geoproximity : ~~route~~ route based on proximity blw user & resource.

→ multivalue : route to multiple endpoints randomly based on their health checks.

→ weighted : route to multiple locations based on the proportion you mention.

- Unlike ELB, Route53 can route traffic inter-region.
- You can register your own domain through Route53
- common records supported by Route53.
  - → A : hostname to IPv4 addresses.
  - → AAAA : hostname to IPv6 addresses.
  - → CNAME : hostname to hostname
    You can use it to redirect traffic from one domain to another. For ex. from acme.example.com to zenith.example.com or to acme.example.org.
  - → ALIAS: hostname to Aws resource.
    It lets you route traffic to selected Aws resources, such as CloudFront or S3.

- Route53 can also be used to create health checks to monitor endpoints or cloudwatch alarms.

- You can configure active-active or active-passive failover configurations.

Database in AWS

- AWS supports numerous types of database for numerous type of use cases. In broader terms, they are categorised as
  - → RDBMS (Relational (SQL) database) : RDS, Aurora.
    They support Online Transaction Processing (OLTP)
  - → NoSQL : DynamoDB (json), ElastiCache (key/value pair), Neptune (graph)
  - → Data Warehouse : Redshift (online Analytics processing), Athena
  - → Search : Elastic Search (json)

- Relational Database Service (RDS) makes it easy to set up, operate, & scale a relational database in cloud.
- It supports PostgreSQL, MySQL, MariaDB, Oracle Database, & SQL Server. (also Aurora)
- RDS is a managed service, you don't need to worry about hardware provisioning, database setup, patching, & backups.
- Supports Multi-AZ deployment for fault tolerance & supports automated fail-over to a secondary ~~data~~ database. For read-heavy workloads, it supports creation of ~~read~~ read replicas.

- The automated backup feature enable point-in-time recovery of database instance.
- They allow database snapshots for user-initiated backups. You can create a new instance from the database snapshot.
- RDS allows encryption at rest using KMS & at transit ~~ssl~~ using SSL & TLS.
- RDS provides Cloudwatch metrics for DB instance at no additional charge for metrics like compute/memory/storage utilization, I/O & instance connections.
- It also allows Enhanced monitoring for OS metrics (no. of threads, child processes etc) at a greater frequency.
- RDS supports General purpose SSDs for standard usage with upto 3000 burst IOPS as well as provisioned IOPS storage for upto 40,000 IOPS.
- In case of ~~disas~~ primary instance going down, the ~~read~~ read replica can be promoted to a primary database.
- RDS is fixed & does not auto-scale.
- The data on RDS is strongly consistent.
- The read/write load on primary database does not affect the replication on the read replica.

- RDS parameter groups allow you to externally provide various parameters like datadir, timeout, cache size etc to be applied to multiple instances during run-time.
- In Multi-Az deployment, the stand by replica is only kept for fault tolerance & cannot be used for read/write operations.
- You can use IAM database authentication with RDS. With this authentication method, you don't need to use a password while connecting to the DB. Instead, you can use a authentication token.
- Each token has a lifetime of 15 min & you don't need to store the user credentials in the DB.

Amazon Aurora

- Aurora is a MySQL & PostgreSQL compatible relational database which is fully-managed by AWS.
- It features a distributed, fault-tolerant, self-healing storage system that auto-scales up to 64TB per DB instance.
- It provides high performance & availability with upto 15 read replicas, point-in-time recovery, continuous backups, & replication across 3 AZs.
- Arora allows custom database endpoints to distribute & load balance workloads across different sets of database instances.
- For globally distributed applications you can use Global Database, where a single Aurora ~~data~~ database is replicated across multiple regions.
- Aurora provides same monitoring as RDS.
- Amazon Aurora Parallel Query is a feature that provides faster analytical queries over your current data.
- Amazon Aurora Serverless is an on-demand, auto-scaling configuration for ~~an~~ Aurora, where DB will automatically start, shut down & scale based on application needs.

- Aurora typically involves a ~~clusted~~ cluster of DB instances instead of a single instance. Each connection is handled by a specific DB instance.
- Using endpoints, you can map each connection to the appropriate instance or group of instances based on your use case.
- Ex: to perform queries, you can ~~connect~~ connect to reader endpoint, with Aurora automatically performing load-balancing blw read replicas.
- Ex: you can connect to custom endpoints associated with different subsets of DB instances.

- ~~Aur~~ During failover, Aurora flips the CNAME of your DB instance to point to a healthy read replica.
- If no read replica is ~~avail~~ available, it attempts to create a new DB instance in the same or different AZ based on the availability.

Elasticache

- Amazon Elasticache allows a scalable in-memory data store in cloud.
- It is a popular choice for caching, session store, gaming, geospatial-services, analytics & queuing.
- It offers fully manged Redis & Memcached for most demanding applications
- It can scale-in & scale-out to meet fluctuating application demands
- Runs on EC2 instances like RDS instances.
- Redis Auth tokens enable Redis to require token (password) before allowing clients to execute commands, ~~theey~~ thereby improving security.

DynamoDB

- Amazon DynamoDB is a key-value (Json based, NoSql) & document database that delivers single-digit millisecond performance at any scale.
- It is fully managed, multiregion, durable DB with built-in security, backup & restore.
- Also, DynamoDB is serverless
- By using DynamoDB global tables, you can replicate your data automatically across multiple ~~AS~~ Aws regions.
- By using DynamoDB streams you can capture item-level modifications & store it for upto 24 hrs. You can use these streams to for applications like sending emails to new users, extracting info from newly ~~added~~ added value in DB etc.
- DynamoDB Accelerator (DAX) is an in-memory cache for DynamoDB tables which provides fast read performance at scale.
- DynamoDB also provides an option for enabling auto-scaling that manages the load based on the no. of requests.
- It works in on-demand & provisioned capacity mode.
- Reads can be eventually consitent or strongly consistent.

- DynamoDB uses primary keys to uniquely identify each item & secondary indexes to provide more querying flexibility.
- You can use additional sort keys with primary key for sorting within a partition.
- DynamoDB can integrate with other AWS services like lambda, Cognito, Redshift, EMR etc.

Athena

- Amazon Athena is an interactive query service that makes it easy to analyze data in S3 using SQL queries.
- Athena is serverless & needs to infrastructure to manage.
- The data never leaves S3.
- You ~~can~~ are charged only for the query you run & the data scanned.
- The results can be stored back to S3.

Neptune

- Amazon Neptune is a fully-managed graph database service to run highly connected datasets.
- Can be used for application like social networking recommendation engines, knowledge graphs etc.

Redshift

- Amazon Redshift is a fully-managed, petabyte-scale data warehouse service in cloud.
- It is a collection of computing resources called nodes which are organized into clusters. Each cluster runs an Amazon Redshift engine & contains one or more databases.
- Redshift is for Online Analytical processing (OLAP)
- Instead of rows, it provides columnar storage of data.
- Redshift Spectrum can be used to directly run queries against S3. The query is submitted to 1000s of redshift spectrum nodes.
- You can use enhanced VPC routing in Redshift. By this, redshift forces all COPY & UNLOAD operations b/w your cluster & data repositories through VPC. You can use standard VPC features like security groups, NACL, endpoints etc.

Storage Gateway

- AWS Storage Gateway is a hybrid cloud storage service that gives you on-prime access to virtually unlimited cloud storage.
- These include moving backups to cloud, using on-premise files shares backed by cloud storage, & providing low-latency access to data in AWS.

File Gateway

- Provides file interface that enables you to store files as objects in S3 using NFS & SMB file protocols.

Tape Gateway

- It presents a virtual tape library (VTL) consisting of virtual tape drives to be stored using iSCSI protocol. Each virtual tape is stored in S3.

Volume Gateway

- It presents your application block storage volumes using the iSCI protocol. Data written to these volumes can be asynchronously backed up as point-in-time snapshots as EBS snapshots & S3

- cached volume gateway store your data in AWS & retain a copy of frequently accessed data locally (on-premise) for fast access.

- stored volume gateway store the entire set of volume data on-premises & store backups in AWS.

Cloudwatch

- Amazon Cloudwatch is a monitoring service from AWS. It collects monitoring & operational data in terms of logs, metrics, and events.
- You can use cloudwatch to detect anomalous behavior in your environments, set alarms, visualize logs & metric side by side, take automated action or troubleshoot.
- Default metrics available in for cloudwatch for EC2 are CPU Utilization, Disk Read Ops, Disk Write Ops, Disk Read Bytes, Disk Write Bytes, Network In/Out, Network Packets In/Out, Meta Data No Token.
- Memory Utilization metric is not available by default & should be pushed by the instance.
- You can use Cloudwatch Agent inside the instance to push custom metrics to Cloudwatch.
- Cloudwatch alarms allow you to set a threshold on metrics & trigger an action. You can use alarms to even trigger actions like send notifications, trigger auto-scaling groups etc.
- By default, EC2 metrics are logged after every 5 min, but by detailed monitoring, you can get data every 1 min.
- To send logs to cloudwatch, make sure the IAM permissions are correct.

CloudTrail

- Aws CloudTrail is a service that enables governance, compliance, operational auditing, and risk auditing of your Aws account.
- CloudTrail provides event history of your Aws account activity, including actions taken through the Aws console, Aws SDK, command line, & other Aws services.
- It is enabled on all Aws accounts by default.
- By default, cloud trail event logs are encrypted using Aws S3 server-side encryption.
- You can identify who or what took which action, what resources were acted upon, when the event occurred, & other details.

Serverless

- Aws serverless is a paradigm in which you don't have to worry about the application ~~are~~ servers running behind a service all the time. Whenever, a request arrives, a virtual server is allocated ~~to it~~ & then, after finishing the request the server are deallocated. You are only charged for the no. of request & execution time.
- Amazon supports the below services in serverless architecture.

→ Aws lambda
→ DynamoDB
→ Aws Cognido
→ API Gateway
→ Aws SQS & SNS
→ Kinesis Data Firehose
→ Aurora Serverless
→ Step functions
→ fargate.

Lambda

- Aws lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume.
- You can set up your code to automatically triggers from other Aws services or call it directly from any web or mobile app.
- There are certain limitations to the lambda compute.

→ memory allocation from 128MB to 3008MB with 64MB increments.
→ timeouts automatically after 15min of execution.
→ Max 4KB of environment variables allowed.

- You can use dead-letter queues in lambda to trigger SNS or SQS in case the lambda function fails during the execution.
- Lambda environment variables allow you to have dynamic variables from within.
- Lambda supports versioning of the code & you can choose which version the lambda should serve.
- Lambda automatically monitors function on you behalf, reporting metric to CloudWatch including invocation request, latency, & error rates.
- You can choose one of the following while deploying new version to lambda.
  - → Canary: traffic is shifted in two increments. by percentage.
  - → linear: traffic is shifted in equal increments.
  - → all at once: all at once.

- Lambda @ Edge: It is a feature of cloudfront which lets you run code closer to the user of your application which improves the latency & performance.

- It is deployed alongside the CF distribution & can filter the requests at the edge location & for faster response.
- Use cases
  - → Security & privacy
  - → Dynamic web app.
  - → Search Engine optimization
  - → Prioritization
  - → Intelligent routing
  - → Bot mitigation
  - → Authentication
  - → Tracking & Analytics.
- Lambda @Edge can access the following 4 requests & responses.
  - → Viewers Request
  - → Viewers Response
  - → Origin Request
  - → Origin Response.
- You can also generate responses to viewers w/o even sending request to the origin.

API Gateway

- Amazon API Gateway is a fully managed service that makes it easy for developers to create, publish, maintain, monitor, & secure APIs at scale.
- APIs act as a front door for applications to access data, business logic, or functionality from backend servers.
- Using API Gateway, you can create RESTful APIs & WebSocket APIs.
- API Gateway console is integrated with CloudWatch to monitor metrics like API calls, latency, and error rates.
- To authorize & verify API requests, you can use signature version 4 by IAM or use lambda to verify JWT tokens or SAML assertions.
- You can set traffic limits to API Gateway say 1000 requests per sec. with burst of 2000 requests per secet.
- You can add caching to API calls by provisioning an API Gateway cache & specifying the size.
- You can use AWS X-Ray to trace & analyse user requests as they travel through your API Gateway.
- API Gateway also supports versioning & environments.
- Use CORS to get resources from different domains.

Cognito

- Amazon Cognito lets you add user sign-up, sign-in, & access control to your web & mobile apps quickly & easily. Cognito scales to millions of users & supports sign-in with social identity providers, such as Facebook, Google, & Amazon, & enterprise identity providers via SAML 2.0.
- Two main components of Cognito are user pool & identity pool. User pool are user directories that provide sign-up & sign-in options. Identity pool enable you to grant access to Aws services to your users.
- There are features to verify the users via email/ phone numbers or MFA.
- Cognito can be integrated with API Gateway for Authentication.

SQS

- Amazon Simple Queue Service (SQS) is a fully managed message queuing service that enables you to decouple your applications.
- You can send, store, & receive messages b/w software components at any volume.
- The messages in the queue can be encrypted for security.
- It scales on demand & there is no limit for no. of messages.
- There are two types of queue.
  - → Standard: The messages can come in any order & may be duplicated.
  - → FIFO: The messages are ordered & are delivered only once.

- The messages can contain data upto 256KB
- The messages can be ~~retained~~ retained in the queue for a maximum of 14 days. Default is 4 days.
- Visibility timeout: The time for which the msg becomes invisible in the queue while an application is processing it. Defaults to 30sec.
- Long polling: when queue is empty, long polling waits up to 20sec. for next msg to arrive. Reduces continously polling of the queue for messages.

- Dead-letter queue: the letter messages which fail to get processed for some no. of times can be sent to DLQ.
- SQS is poll based offering, the applications need to poll the data from the queue themselves.
- The msg inside a queue needs to be deleted explicitly after processing, else they will reappear after the visibility timeout.
- While processing the msg, its visibility can be changed by using the APIs.
- Messages in FIFO queue can be grouped using Message GroupID.

SNS

- Amazon Simple Notification Service is a highly available, durable, secure, fully managed pub/sub messaging service used for decoupling applications.
- SNS provides high-throughput, push-based, many-to-many messaging.
- Fan-out: Send the same message to multiple subscribers
- SNS can be used to fan out notifications to end users using mobile push, SMS, & email.
- For receiving the messages, a subscriber user needs to subscribe the topic for which he has to receive the notification. Whenever an application publishes to that topic, it goes to all subscribers.
- It is a push-based model. & no polling is required.

KINESIS

- Amazon kinesis makes it easy to collect, process, & analyse real-time, streaming data so that you can get timely insights.
- You can ingest realtime data such as video, audio, app logs, website clickstreams & IoT data for ML, analytics, & other applications.
- Kinesis is composed of
  - → kinesis streams : to stream data at scale.
  - → kinesis analytics : to perform real-time analytics on streams using SQL
  - → kinesis Firehose : to load streams into S3, redshift, elastic search, splunk etc.

- Kinesis streams are divided into shards/partitions The more the no. of shards, the more the data can be processed in parallel. Each shard has limit on the data rate hence for increasing the performance, increase the no. of shards.
- The data retention per shard is 1 day by default & can extend upto 7 days.
- Access on kinesis can be controlled by IAM policies. & the data can be encrypted in transit or at rest.

Miscellaneous

Cloudfront

- Amazon Cloudfront is a fast content delivery network (CDN) service that securely delivers data, videos, applications, & APIs to consumers globally with low latency.
- Cloudfront works seamlessly with services including AWS Shield, S3, ELB or EC2 as origin, & Lambda@Edge.
- You don't pay for any data transferred b/w these services & CF.
- CF is highly secure & you can also use AWS Certificate Manager to manage SSL certificates, AWS Shield for DDoS protection, & WAF for detecting malicious activities.
- CF is deployed at Amazon Edge locations & is global.
- CF can speed up the delivery of static content by caching.
- With Signed URLs & cookies you can support limited access to authenticated viewers.
- Through geo-restriction, you can prevent users in specific geographic locations from accessing content.
- With Origin Access Identity (OAI) you can restrict access to S3 bucket to only be accessible from CF. Define bucket policy in S3 to only allow OAI to access the resources

- You can use CF's native origin failover capability to automatically serve your content from a backup origin when your primary origin is unavailable.
- You can change the TTL for cache to 0s for serving dynamic content. The default TTL is 24hrs.
- You can use versioned file names to control which file the CF returns to the users even if he has the older version cached.
- CF can use Server Name Indication (SNI) to serve multiple domain over SSL just like ALB.
- CF is an easy way to make an existing application more scalable & cheaper w/o any changes to the architecture.

Global Accelerator

- Aws Global Accelerator is a service that improves the availability & performance of your application with local or global users.
- It provides static IP addresses that act as a fixed entry point to your application.
- It uses Aws global network to optimize the path from your users to your application.
- GA continually monitors the health of your app endpoint & redirects traffic accordingly.
- GA supports DDoS protection through AWS Shield
- Unlike ~~&~~ Cloudfront which is restricted to HTTP traffic, GA can support TCP & UDP traffic

AWS Shield

- It is a managed Distributed Denial of Service (DDoS) protection service that safeguards applications running on Aws.
- There are two tiers of shield - Standard & Advance.
→ Standard - defends against most common, frequently occuring network & transport layer DDoS attacks
→ Advance - For higher levels of protection. ~~or~~ for apps running on EC2, ELB, CF, GA, S3 etc.

AWS WAF

- AWS web Application Firewall is a web application firewall that helps protect your app or API against common web exploits that may affect the availability, compromise security, or consume excessive resources.
- You can block common attack patters, such as SQL injection or cross-site scripting, and rules that filters out specific traffic (IP).
- You can deploy WAF on Cloud front, ALB or API gateway.

AWS SCT

- Aws Schema Conversion Tool helps convert your existing database schema from one database engine to another.
- You can convert from a relational OLTP sch schema or any warehouse OLAP schema to Amazon RDS or Amazon Redshift.

Compute & Networking

- To provide higher bandwidth, or higher packets per second (PPS) with low latency, use Elastic Network Adapter (upto 100 Gbps).
- Use Elastic Fabric Adapter (EFA) for HPC applications. Only works for Linux.

Aws Batch

- Aws Batch enables developers to easily & efficiently run hundreds of batch computing jobs on Aws.
- It dynamically provisions the optimal quantity & type of computing resources based on volume.

Aws ParallelCluster

- Aws ParallelCluster is an Aws-supported cluster management tool that makes it easy to deploy & manage HPC clusters on Aws.
- It uses a simple text file to model & provision all the resources needed for HPC applications.

Aws codeCommit

- Fully-managed source control service that hosts secure Git-based repositories. It makes it easy for teams to collaborate on code in a secure & highly scalable ecosystem.

Aws codeBuild

- Fully-managed continuous integration service that compiles source code, run tests, & produce software packages that are ready to deploy.

Aws codeDeploy

- Fully-managed deployment service that automates software deployments to a variety of compute services like Ec2, Fargate, Lambda, or on-premise servers.

Aws codePipeline

- Fully-mangged continuous delivery service that helps you automate your release pipelines for fast & reliable application.
- It automates the build, test, & deploy phases. every time there is a code change.

Cloudformation

- Aws Cloudformation provides a common language for you to model & provision Aws & Third party application resources in your cloud environment.
- It allows you to use programming languages or a simple text file to model & provision all the resources needed for your application across all regions or accounts.
- It provides Infrastructure as code.

Elastic Beanstalk

- Aws Elastic Beanstalk is an easy-to-use service for deploying & scaling web applications & services developed with Java, .NET, PHP, Node.js etc on familiar servers such as Apache, Nginx etc.
- You can simply upload your code & Elastic Beanstalk automatically handles the deployment, from capacity provisioning, load balancing, auto scaling to health monitoring.
- You only pay for the resources you use, same as Cloudformation.

Step Functions

- Aws step functions lets you coordinate multiple Aws services into serverless workflows so you can build & update apps quickly.
- You can stitch together services like Aws lambda, fargate, SageMaker into rich feature-rich apps.
- Workflows are made up of a series of steps where the output of one can be the input of the next step.
- It can have a max execution time of 1 year
- Possibility to implement human approval feature which needs to be done manually by the human.

SWF

- Aws Simple Workflow Service is same as Step functions but runs on dedicated EC2 instances & is not serverless.
- Therefore, for newer applications, Step functions are recommended.

EMR

- Amazon Elastic MapReduce is a cloud big data platform for processing vast amounts of data using open source tools like Apache Spark, Hive, HBase, Flink etc.
- EMR can run petabyte-scale analysis at less than half of cost of traditional on-premise solutions.
- clusters can be made up of 100s of EC2 instances & EMR takes care of provisioning & configuration.

Glue

- AWS Glue is a fully-managed extract, transform, & load (ETL) service that makes it easy for customers to prepare & load data for analytics.
- It is serverless & can be run on Aurora, RDS, Redshift or S3.

Opsworks

- For chef & Puppet application to perform server configuration automatically.

Elastic Transcoder

- To convert media files stored in S3 into various formats for tablets, & laptops, TV, mobiles etc.

Workspaces

- It is a managed & secure cloud Desktop

AppSync

- Store & sync data across mobile & web apps in real-time.

Organization

- Hierarchy & centralized management of multiple AWS accounts.

Trusted Advisor

- Aws Trusted Advisor is an online tool that provides you real time guidance to help you provision your resources following Aws best practices.
- It can help optimize your Aws infrastructure, increase security & performance, reduce cost, & monitor service limits.

RAM

- Amazon Resource Access Manager (RAM) is a service that enables you to easily & securely share AWS resources with any AWS account or within your AWS Organization.
- You can share ~~transit~~ Transit Gateway, subnets, Route 53 etc. using RAM.
- Many organization use multiple accounts to create administrative or billing isolation, & to limit the impact of errors. RAM eliminates the need to create duplicate resources in multiple accounts, reducing the operational overhead.

Secrets Manager

- Amazon secrets Manager helps you protect secrets needed to access your applications, services, & IT resources.
- The services enable you to easily rotate, manage, & retrieve database credentials, API keys, & other secrets throughout their lifecycle.
- No need to hard code ~~sec~~ sensitive information in plain text files.

Systems Manager

- Aws System Manager gives you visibility & control of your infrastructure on Aws. It provides a unified user safer interface so you can view operational data from multiple Aws services & allows you to automate operational tasks across Aws account.
- You can group resources like EC2, S3, RDS etc.
- Some of its features include
  - → Run command : provides you safe, secure remote management of your instance w/o using bastion host or SSH.
  - → Parameter store : centralized store to manage your configuration data, whether plan-text data such as database string or secrets such as passwords

WorkDocs

→ Aws workdocs is a simple, fully-managed, secure content creation, storage, & collaboration service. You can easily create, edit, & share content.

certificate Manager

- Aws certificate manager is a service that lets you easily provision, manage, & deploy public & private SSL/TLS certificates for use with Aws services.

Config

- Aws Config is a service that enables you to assess, audit, & evaluate the configuration of your Aws resources.
- It continuously monitors & records your Aws resource configuration & allows you to automate the evaluation of recorded configs against desired configs.

Directory Service

- Aws Directory Service or Microsoft Active Directory enables your directory-aware workloads & Aws resources to use managed Active Directory (AD) in Aws.
- It can easily join EC2 & RDS instances to your domain, & use Amazon Workspaces with Active Directory users & groups.

- You can use Aws Directory Service AD Connector for easily integrating corporate AD with AWS.
- You can use Microsoft AD or LDAP etc with Directory Service.

ECS

- Amazon Elastic Container Service is a fully managed container orchestration service.
- You can deploy the containers on EC2 instances or Fargate which is serverless
- Multiple docker images can run independently on a single machine & can directly integrate with ALB.
- Terminology
  - → ECS cluster: a set of EC2 instances on which the containers will run.
  - → Service: the application definition running on ECS cluster. Multiple services can run in a single ~~EC2~~ EC2 instance.
  - → task definition: containers running to create the application. Can be declared in JSON files.
- Elastic container Registery (ECR) stores, manages & deploys your containers on AWS.

- With Fargate you dont need to provision EC2 clusters. The containers will run in a serverless fashion where you dont need to worry about the underlying hardware.

Disaster Recovery

- Aws provides 4 disaster recovery strategies.

→ Backup & Restore: Periodically take backups of your resources & when the application fails, spin up new resources from the backup & start the application.
This may cause long down times. & loss of data.

→ Pilot light: A small version of app always running on the cloud. Used for critical core of data like RDS instances.
When the application goes down, other resources like instances, can be provisioned to restore the service.
Causes less down time & data loss than backup & restore option.

→ Warm Standby: A full system is running on Aws at a small scale. During failover, the architecture can be scaled to meet the demand.

→ Multi site: Full production scale of Infra running on AWS. Ver-M: No downtime but costly.

Snowball

- AWS snowball is a data migration & storage device that is used to transfer huge amount of data from on-prime data centre to AWS.
- For data greater than 50TB snowball is preferred when transferring data over internet can take months. Snowball can transfer data within a week directly to S3.
- 80TB & 50TB options are available.
- Snowball edge devices are filled with additional CPUs to filter or transform the data before storing in snowball.
- Snowmobile is used for data transfer of petabyte-scale which consists of multiple snowball devices.