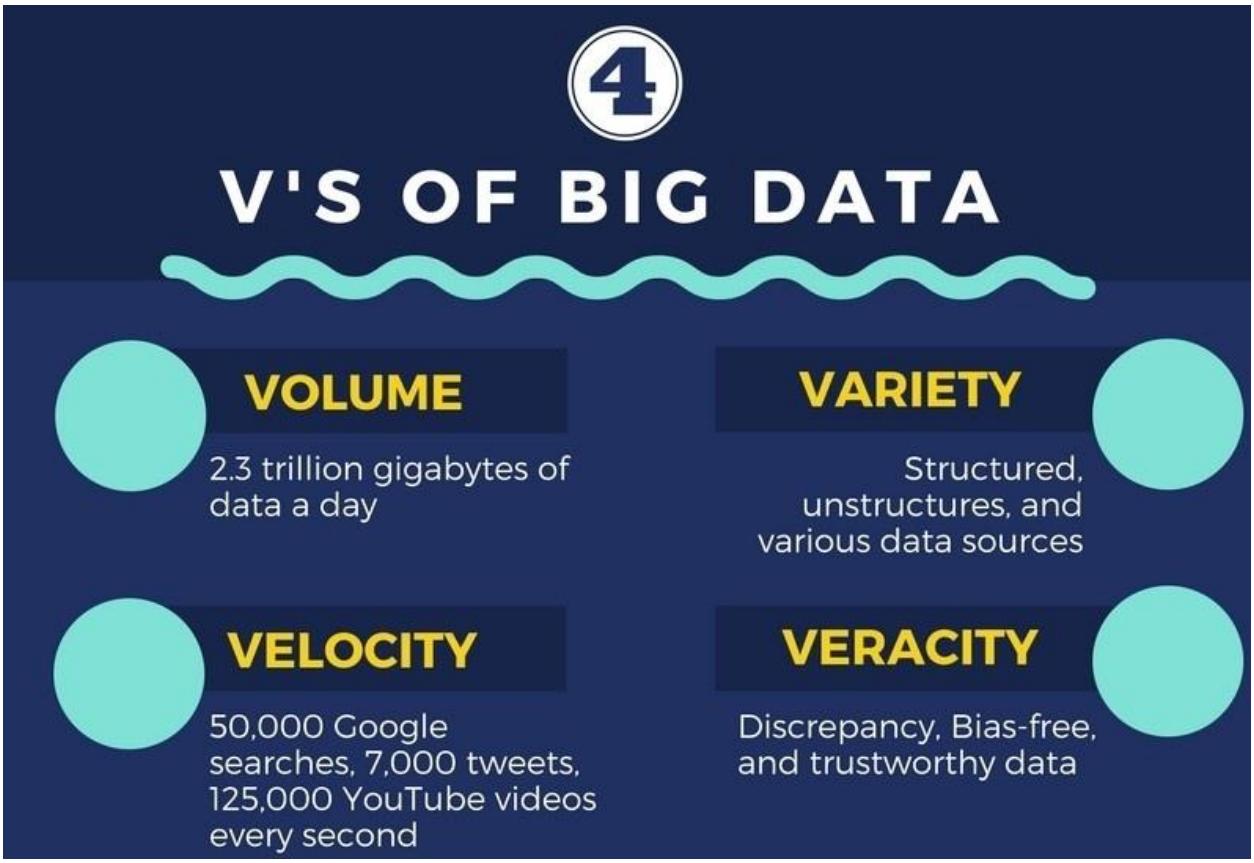
Introduction to Cloud and AWS

Big Data:

- Data that can't be stored and processed by a single computer is referred to as Big Data. There are generally four characteristics that must be part of a dataset to qualify it as big data – Volume, Variety, Velocity and Veracity.
- Volume – It is estimated that we generate 2.3 trillion gigabytes of data every day. Size of the data grows exponentially with Text and WhatsApp messages, photos, videos and many apps ensure that the amount of data increases significantly.
- Velocity – Velocity, or speed refers to the enormous speed at which the data is generated and processed. Data is available in real time and hence Big Data systems should be capable to capture/process data in real time.
- Variety – Data could be in structured, semi-structured or unstructured format. Example – CCTV audio and Video files.
- Veracity - Veracity refers to the quality of the data that is being analyzed. Data quickly becomes outdated and the information shared via the internet and social media does not necessarily have to be correct.



Cloud Computing:

- Cloud Computing is the distribution of on-demand delivery of IT services over the internet("the cloud") with pay as you go pricing model. Instead of buying, owning and maintaining the physical servers, Cloud Computing allows us to access the technology services such as computing power, storage and databases, as you need them from the cloud service providers like Amazon Web Services(AWS), Google Cloud Platform(GCP), Microsoft(Azure) and etc…

- Essentially, "Big Data" refers to the large sets of data collected, while "Cloud Computing" refers to the mechanism that remotely takes this data in and performs any operations specified on that data.

Key terminologies:

On-premise:
On-premise consists of an IT infrastructure comprising systems, hardware applications, and software applications. With an on-premise system, you get complete ownership of your servers, and you can personally oversee the maintenance of your systems. On-premise systems were the traditional method until cloud computing slowly started taking over. One of the major differences between cloud computing and on-premise is the local storage and physical maintenance of systems.

Virtual Machine:
A virtual computer system is known as a "virtual machine" (VM): a tightly isolated software container with an operating system and application inside. Each self-contained VM is completely independent. Putting multiple VMs on a single computer enables several operating systems and applications to run on just one physical server, or "host."

Hypervisor:
A Hypervisor, also known as a virtual machine monitor or VMM, is a software that creates and runs virtual machines (VMs). A hypervisor allows one host computer to support multiple guest VMs by virtually sharing its resources, such as memory and processing.

Types of Hypervisors:

- Bare-Metal (Type1) Hypervisor
  A Type1 hypervisor acts like a lightweight operating system and runs directly on the host's hardware. The most commonly deployed type of hypervisor is the Type 1 or bare-metal hypervisor, where virtualization software is installed directly on the hardware where the operating system is normally installed. Because bare-metal hypervisors are isolated from the attack-prone operating system, they are extremely secure. In addition, they generally perform better and more efficiently than hosted hypervisors. For these reasons, most enterprise companies choose bare-metal hypervisors for data center computing needs.

- Hosted (Type2) Hypervisor
  A Type2 hypervisor runs as a software layer on an operating system, like other computer programs. Hosted hypervisors run on top of the operating system (OS) of the host machine. Although hosted hypervisors run within the OS, additional (and different) operating systems can be installed on top of the hypervisor. The downside of hosted hypervisors is that latency is higher than bare-metal hypervisors. This is because communication between the hardware and the hypervisor must pass through the extra layer of the OS. Hosted hypervisors are sometimes known as client hypervisors because they are most often used with end users and software testing, where higher latency is less of a concern.

Containerisation:

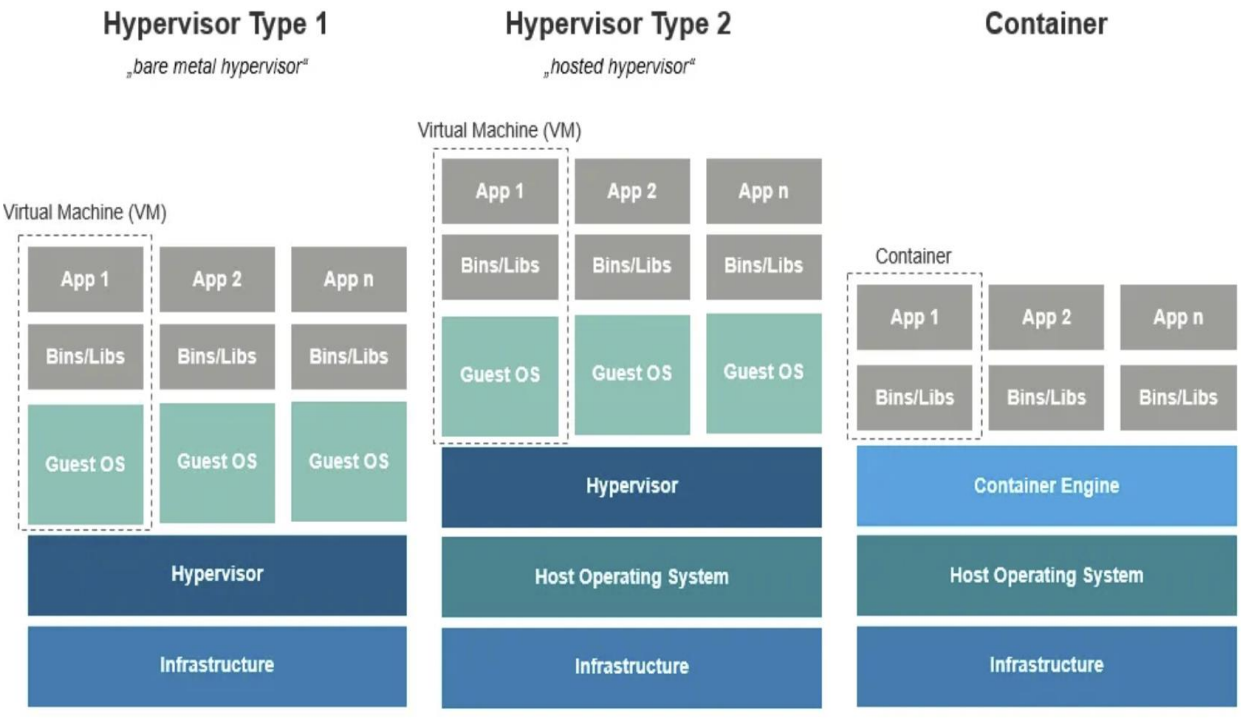
Containerisation is a way to deploy application code to run on any physical or virtual environment without changes. Developers bundle application code with related libraries, configuration files, and other dependencies that the code needs to run. This single package of the software called a container, can run independently on any platform. Containerisation is a type of application virtualisation.

Containers vs Hypervisors
Containers and hypervisors are both involved in making applications faster and more efficient, but they achieve this in different ways.

Hypervisors:
Allow an operating system to run independently from the underlying hardware through the use of virtual machines.
Share virtual computing, storage and memory resources.
Can run multiple operating systems on top of one server (bare-metal hypervisor) or installed on top of one standard operating system and isolated from it (hosted hypervisor).

Containers:
Allow applications to run independently of an operating system.
Can run on any operating system—all they need is a container engine to run.
Are extremely portable since in a container, an application has everything it needs to run

Hypervisors and containers are used for different purposes. Hypervisors are used to create and run virtual machines (VMs), which each have their own complete operating systems, securely isolated from the others. In contrast to VMs, containers package up just an app and its related services. This makes them more lightweight and portable than VMs, so they are often used for fast and flexible application development and movement.



Virtual Machine (VM) vs. Container

Instance:
An instance in cloud computing is a server resource provided by third-party cloud services. While you can manage and maintain physical server resources on premises, it is costly and inefficient to do so. Cloud providers maintain hardware in their data centers and give you virtual access to compute resources in the form of an instance. You can use the cloud instance for running compute-intensive workloads like containers, databases, microservices, and virtual machines.

Instance vs Virtual Machine:
You can run multiple virtual machines on a single computer, but when you run virtual machines in the cloud environment, they are known as instances. Running virtual machines on the cloud allows organizations to benefit from the cost effectiveness of sharing and scaling resources.

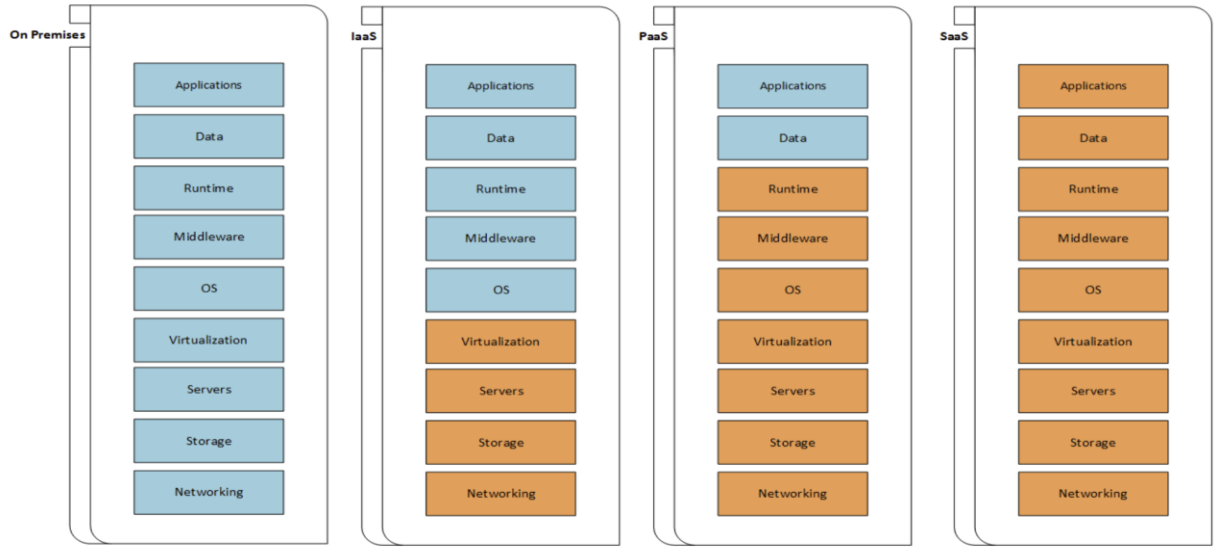
Cluster:
A Cluster is a collection of instances. Each instance in a cluster is called a node.

Types of Cloud Computing
The services hosted on cloud can be broadly classified into Infrastructure As A Service(IAAS), Platform As A Service(PAAS) and Software As A Service(SAAS).

IAAS - For IaaS models, the service provider hosts, maintains, and updates the backend infrastructure, such as compute, storage, networking, and virtualization. You manage everything else including the operating system, middleware, data, and applications. Example: AWS - EC2,S3, Google - Compute Engine, Cloud Storage

PAAS - Like IaaS models, for PaaS models, the service provider delivers and manages the backend infrastructure. However, PaaS models provide all the software features and tools needed for application development. You still have to write the code and manage your apps and data but do not have to worry about managing or maintaining the software development platform.
SAAS - With SaaS service models, the service provider delivers the entire application stack—the complete application and all the infrastructure needed to deliver it. As a customer, all you have to do is connect to the app through the internet—the provider is responsible for everything else.

On-Premise vs IAAS vs PAAS vs SAAS:
Areas that are colored in blue are managed by you and all others are the responsibility of your cloud service provider.

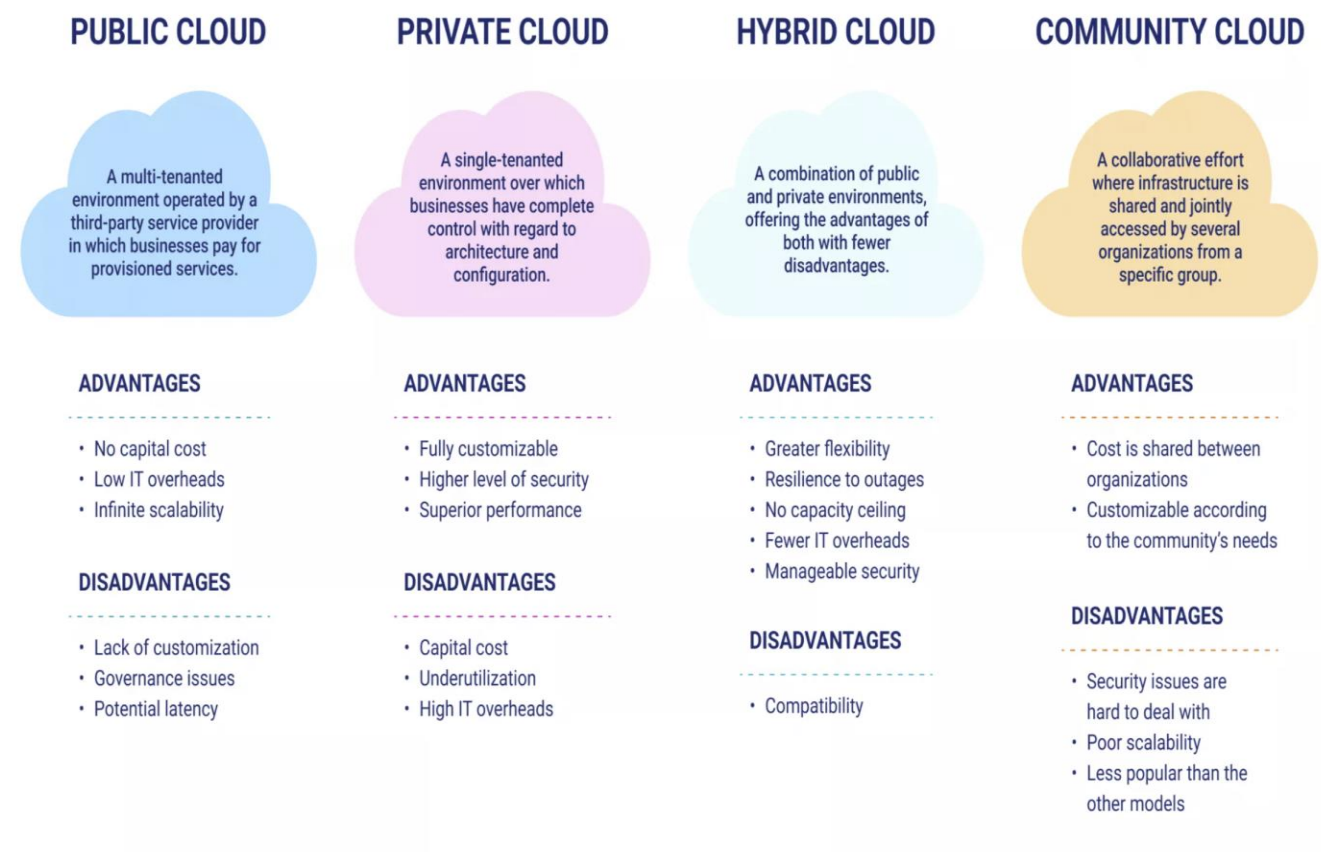


Deployment Models
Based on the deployment model, cloud can be classified into public, private, community and hybrid cloud.
Public - The Public Cloud Model allows the general public to access all the services provided by the cloud. Clouds based on this model are highly scalable and cost-effective, consisting of a huge amount of space.

Private – The Private Cloud allows people within the organization to access all the services provided by the cloud. The Private Cloud is operated only within a single organization. However, it may be managed internally or by a third party.

Community - The Community Cloud allows group of organizations sharing the same interest to access services provided by the cloud. It can be possiblly managed internally or by a third party.

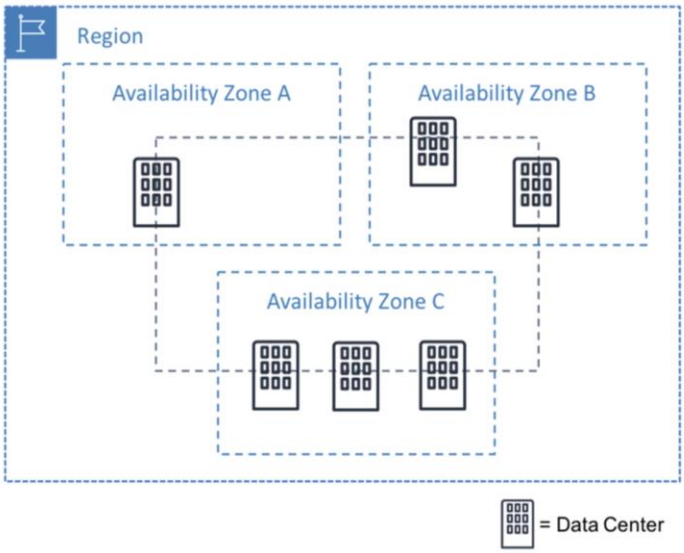
Hybrid - The Hybrid Cloud is a combination of both public and private clouds. In this, Non Critical activities are done through the public cloud while critical activities are done through the private cloud.



Amazon Web Services(AWS) –

AWS is a comprehensive, most adopted cloud service provider. They offer more than 200 services and more features within those services than any other service providers-from infrastructure technologies like compute, storage, and databases–to emerging technologies, such as machine learning and artificial intelligence, data lakes and analytics, and Internet of Things. The AWS Region and Availability Zone model has been recognized by Gartner as the recommended approach for running enterprise applications that require high availability. Some of the companies that use AWS are Instagram, Netflix, LinkedIn, Twitch, Facebook, Pinterest and Dropbox

AWS Region and Availability Zones:
AWS Cloud computing resources are housed in highly available data center facilities. To provide additional scalability and reliability, these data center facilities are located in different physical locations. These locations are categorized by Regions and Availability Zones. Each AWS Region is a separate geographic area. Each AWS Region has multiple, isolated locations known as Availability Zones.



Some of the widely used AWS services are:

IAM – AWS Identity and Access Management (IAM) is a web service that helps you securely control access to AWS resources. With IAM, you can centrally manage permissions that control which AWS resources users can access. You use IAM to control who is authenticated (signed in) and authorized (has permissions) to use resources.

- IAM Root User - When you create an AWS account, you begin with one sign-in identity that has complete access to all AWS services and resources in the account. This identity is called the AWS account root user and is accessed by signing in with the email address and password that you used to create the account.

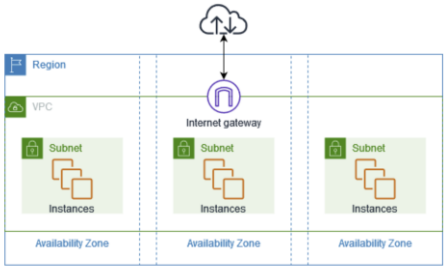
- IAM User - IAM users can be individuals, systems, or applications requiring AWS services A user account comprises a unique name and security credentials such as a password, access key, and/or multi-factor authentication (MFA).

- IAM Policy - IAM policies define permissions for action regardless of the method that you use to operate. For example, suppose a policy allows the GetUser action. In that case, a user with that policy can get user information from the AWS Management Console, the AWS CLI, or the AWS API.

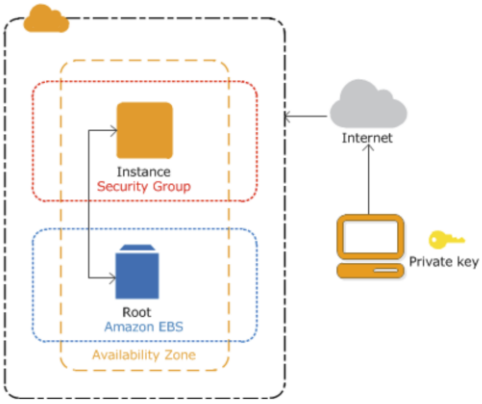
- IAM Group – IAM Groups are a way to assign permissions to your organisation’s logical and functional units. IAM Groups are a tool to help with operational efficiency, bulk permissions management (scalable), and easy permission changes as individuals change teams (portable). A group can have many users, and a user can be a member of multiple groups. Groups cannot be nested; they can only contain users and no other groups.

- IAM Role – An IAM role, like a user, is an AWS identity with permission policies governing what the identity can and cannot do in AWS. For specific access to services, you can authorize roles to be assumed by humans, Amazon EC2 instances, custom code, or other AWS services. Roles do not have standard long-term credentials associated with them, such as a password or access keys; rather, when you assume a role, it provides you with temporary security credentials for your role session.

Virtual Private Cloud(VPC) – A VPC is a virtual network that closely resembles a traditional network that you'd operate in your own data center. After you create a VPC, you can add subnets. A subnet is a range of IP addresses in your VPC. A subnet must reside in a single Availability Zone. After you add subnets, you can deploy AWS resources in your VPC.



Elastic Compute Cloud(EC2) – EC2 is a web service that provides resizable compute capacity-literally, servers in Amazon's data centers, that you use to build and host your software systems. And its an IAAS service.



Here are different types of EC2 Instances:

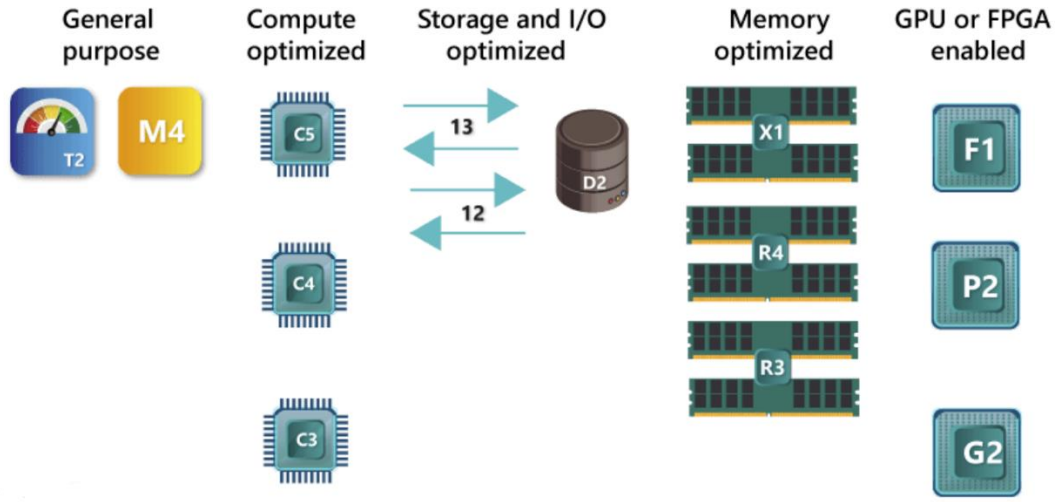
- General Purpose Instances - General Purpose instance offers a wide balance of computing power, memory, and storage., it is suited for a majority of AWS workloads. General-purpose instances are mostly utilized in services related to web servers, mobile or gaming development environments or apps, or enterprise-level applications like ERP or CRM. Another major distinction among general-purpose instances is the use of Fixed EC2 instances and Burstable instances – where you can scale up your overall computing power at an extra cost. Some General Purpose instance types – A1, M4, T2

- Compute Optimized Instances - Compute-optimized instances are used during compute-intensive workloads that can benefit from processors with high computing power. It delivers high performance at a cost-effective price and are typically used in applications like web servers and scientific modeling. Some Compute Optimized instance types – C5/C5n, C6/C6g

- Memory-Optimized Instances – Memory-optimized instances are used for memory-intensive workloads that are required to process large datasets at a fast speed. Examples of memory-intensive applications include Big Data analytics or those running on Hadoop or Apache Spark. Some Memory-Optimized instance types – R5/R5a/R5n, R6g/R6gd, X1/X1e

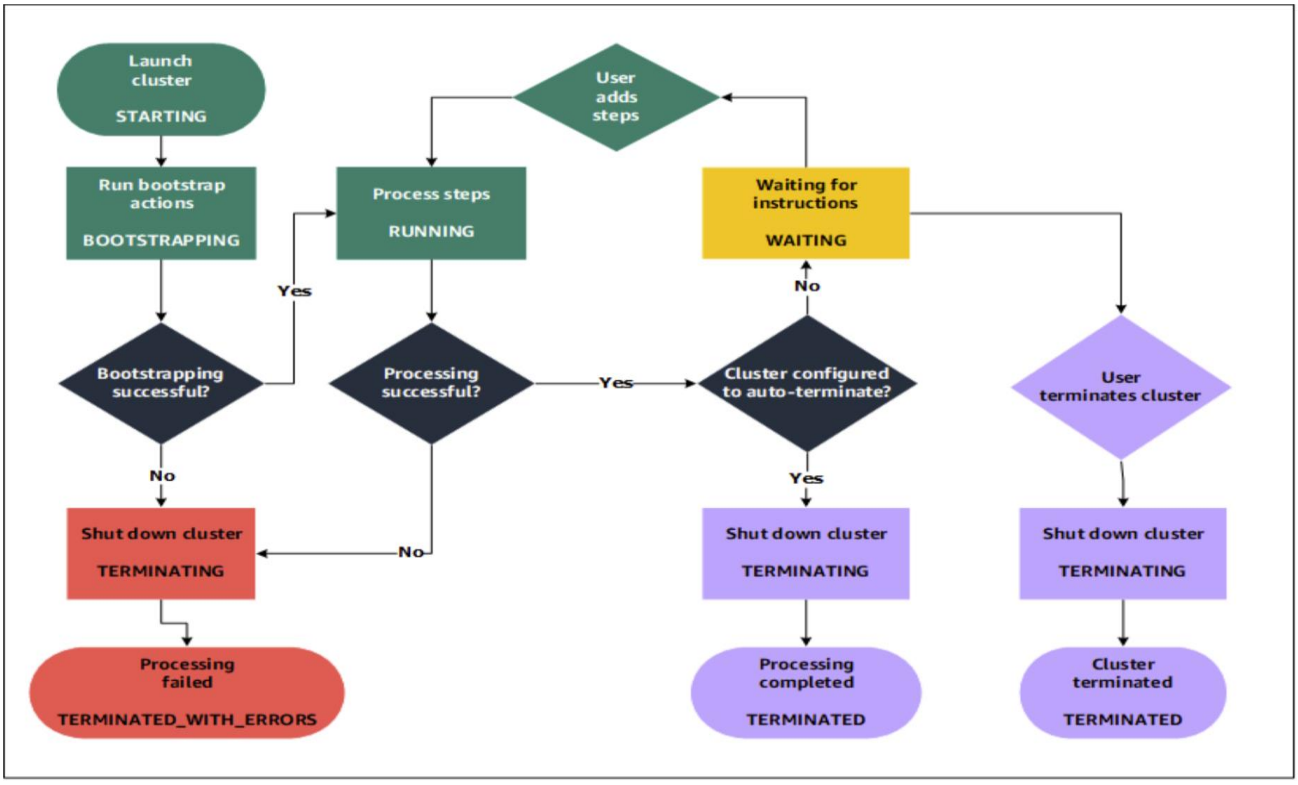
- Accelerated Computing Instances - Accelerated Computing instances use additional hardware accelerators like Graphics Processing Units (or GPUs) and Field Programmable Gate Arrays (or FPGAs) that enable higher throughput in compute-intensive applications with more parallelism. For example, with GPU-powered instances, applications can access NVIDIA GPUs that have thousands of computing cores. Similarly, FPGA-powered instances provide applications with access to large FPGAs with millions of parallel logic cells. This instance type is suitable for applications that require parallel processing. This includes graphic processing, floating-point calculators, and data pattern matching. Some Accelerated Computing instance types – P3, P2, Inf1, G3, G4, F1

- Storage Optimized Instances - Storage-optimized instances are used for applications that have high storage requirements, particularly with sequential read-and-write applications like log processing. Storage-optimized instances are designed to deliver a high number of low latency and random I/O operations each second (or IOPS). Storage-optimized instances are also suitable for cloud-running applications that run high transaction and low latency workloads in use cases such as in-memory databases, data warehousing, and data analytics. Some storage optimized instance types – D2, H1, I3/I3en



Elastic MapReduce(EMR) – Amazon EMR is a web service that makes it easy to process vast amounts of data efficiently using Apache Hadoop and services offered by Amazon Web Services. Amazon EMR processes big data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). The Elastic in EMR's name refers to its dynamic resizing ability, which enables administrators to increase or reduce resources, depending on their current needs.

Life Cycle of an EMR Cluster-

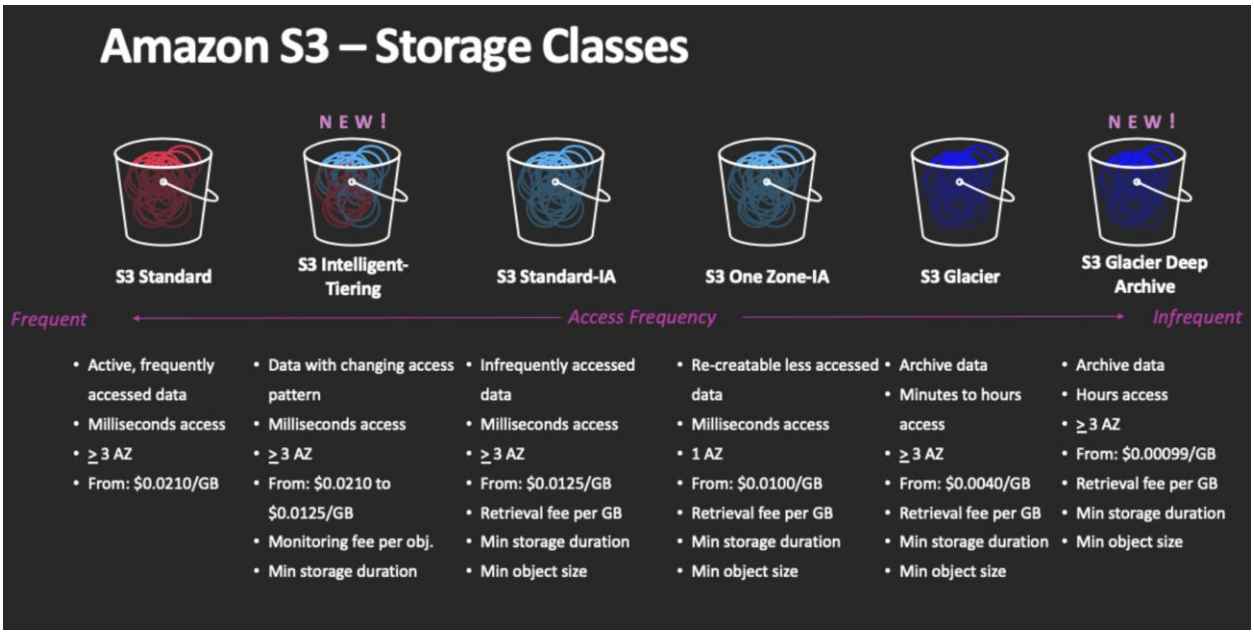


S3 – Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance. S3 could store and protect any amount of data for a range of use cases, such as data lakes, websites, cloud-native applications, backups, archive, machine learning, and analytics.Amazon S3 is designed for 99.999999999% (11 9's) of durability, and stores data for millions of customers all around the world.

Amazon S3 offers a range of storage classes that you can choose from based on the data access, resiliency, and cost requirements of your workloads. S3 storage classes are purpose-built to provide the lowest cost storage for different access patterns. S3 storage classes are ideal for virtually any use case, including those with demanding performance needs, data residency requirements, unknown or changing access patterns, or archival storage.

The S3 storage classes include S3 Intelligent-Tiering for automatic cost savings for data with unknown or changing access patterns, S3 Standard for frequently accessed data, S3 Standard-Infrequent Access (S3 Standard-IA) and S3 One Zone-Infrequent Access (S3 One Zone-IA) for less frequently accessed data, S3 Glacier Instant Retrieval for archive data that needs immediate access, S3 Glacier Flexible Retrieval (formerly S3 Glacier) for rarely accessed long-term data that does not require immediate access, and Amazon S3 Glacier Deep Archive (S3 Glacier Deep Archive) for long-term archive and digital preservation with retrieval in hours at the lowest cost storage in the cloud. If you have data residency requirements that

can’t be met by an existing AWS Region, you can use the S3 Outposts storage class to store your S3 data on premises. Amazon S3 also offers capabilities to manage your data throughout its lifecycle. Once an S3 Lifecycle policy is set, your data will automatically transfer to a different storage class without any changes to your application.



RDS - Amazon Relational Database Service (Amazon RDS) is a web service that makes it easier to set up, operate, and scale a relational database in the AWS Cloud. It provides cost-efficient, resizable capacity for an industry-standard relational database and manages common database administration tasks. We can choose from seven popular engines — Amazon Aurora with MySQL compatibility, Amazon Aurora with PostgreSQL compatibility, MySQL, MariaDB, PostgreSQL, Oracle, and SQL Server — and deploy on-premises with Amazon RDS on AWS Outposts.

Lambda - AWS Lambda is a serverless, event-driven compute service that lets you run code for virtually any type of application or backend service without provisioning or managing servers. You can trigger Lambda from over 200 AWS services and software as a service (SaaS) applications, and only pay for what you use.

Sample Case Study - File Processing with Lambda

Accessing AWS services:

AWS services can be accessed through AWS Management console, AWS Command Line Interface(CLI) or AWS Software Development Kit(SDK).

AWS Management Console - is a web application that comprises and refers to a broad collection of service consoles for managing AWS resources. When you first sign in, you see the console home page. The home page provides access to each service console and offers a single place to access the information you need to perform your AWS related tasks. It also lets you customize the Console Home experience by adding, removing, and rearranging widgets such as Recently visited, AWS Health, Trusted Advisor, and more.

AWS CLI – AWS CLI is an open-source command-based application developed by AWS for easy management of your AWS cloud service. It provides different commands for managing AWS cloud services. It is compatible with the windows command line, Linux/Mac shell programs, and even remote terminals like putty. You can run commands in the terminal of your operating system. For windows, you can run commands in the command prompt. For Linux and Mac, you can use any shell programs like Bash, Zsh, or Tcsh. You can even run AWS CLI commands on remote EC2 instances through putty or SSH. Just like you would set up and manage different AWS services through its console, you can perform all the functions through the CLI tool.

AWS SDK – The AWS SDK for Python (Boto3) provides a Python API for AWS infrastructure services. Using the SDK for Python, you can build applications on top of Amazon S3, Amazon EC2, Amazon DynamoDB, and more.