AWS

CLOUD COMPUTING:

on-demand delivery of compute power, database, storage, applications, and other IT resources through a cloud services platform through the internet

Cloud-based deployment:

- Run all parts of the application in the cloud.
- Migrate existing applications to the cloud.
- Design and build new applications in the cloud.

In a **cloud-based deployment** model, you can migrate existing applications to the cloud, or you can design and build new applications in the cloud. You can build those applications on low-level infrastructure that requires your IT staff to manage them. Alternatively, you can build them using higher-level services that reduce the management, architecting, and scaling requirements of the core infrastructure.

For example, a company might create an application consisting of virtual servers, databases, and networking components that are fully based in the cloud.

On Premises Deployment:

- Deploy resources by using virtualization and resource management tools.
- Increase resource utilization by using application management and virtualization technologies.

On-premises deployment is also known as a private cloud deployment. In this model, resources are deployed on premises by using virtualization and resource management tools.
For example, you might have applications that run on technology that is fully kept in your on-premises data center. Though this model is much like legacy IT infrastructure, its incorporation of application management and virtualization technologies helps to increase resource utilization.

Hybrid Deployment*:*

- Connect cloud-based resources to on-premises infrastructure.
- Integrate cloud-based resources with legacy IT applications.

In a hybrid deployment, cloud-based resources are connected to on-premises infrastructure. You might want to use this approach in a number of situations. For example, you have legacy applications that are better maintained on premises, or government regulations require your business to keep certain records on premises.
For example, suppose that a company wants to use cloud services that can automate batch data processing and analytics. However, the company has several legacy applications that are more suitable on premises and will not be migrated to the cloud. With a hybrid deployment, the company would be able to keep the legacy applications on premises while benefiting from the data and analytics services that run in the cloud.

BENEFITS OF CLOUD COMPUTING:

- **Trade upfront expenses for variable expenses**:- Upfront expense refers to data centers, physical servers, and other resources that you would need to invest in before using them. Variable expense means you only pay for computing resources you consume instead of investing heavily in data centers and servers before you know how you're going to use them.By taking a cloud computing approach that offers the benefit of variable expense, companies can implement innovative solutions while saving on costs.
- **Stop sending money to run and maintain data centers.:-** Computing in data centers often requires you to spend more money and time managing infrastructure and servers.
- **Stop guessing capacity**.:- With cloud computing, you don't have to predict how much infrastructure capacity you will need before deploying an application. For example, you can launch Amazon EC2 instances when needed, and pay only for the compute time you use. Instead of paying for unused resources or having to deal with limited capacity, you can access only the capacity that you need. You can also scale in or scale out in response to demand
- **Benefit from massive economies of scale.:-** By using cloud computing, you can achieve a lower variable cost than you can get on your own. Because usage from hundreds of thousands of customers can aggregate in the cloud, providers, such as AWS, can achieve higher economies of scale. The economy of scale translates into lower pay-as-you-go prices.
- **Increase speed and agility;-** The flexibility of cloud computing makes it easier for you to develop and deploy applications.This flexibility provides you with more time to experiment and innovate. When computing in data centers, it may take weeks to obtain new resources that you need. By comparison, cloud computing enables you to access new resources within minutes.
- **Go global in minutes:-** The global footprint of the AWS Cloud enables you to deploy applications to customers around the world quickly, while providing them with low latency. This means that even if you are located in a different part of the world than your customers, customers are able to access your applications with minimal delays

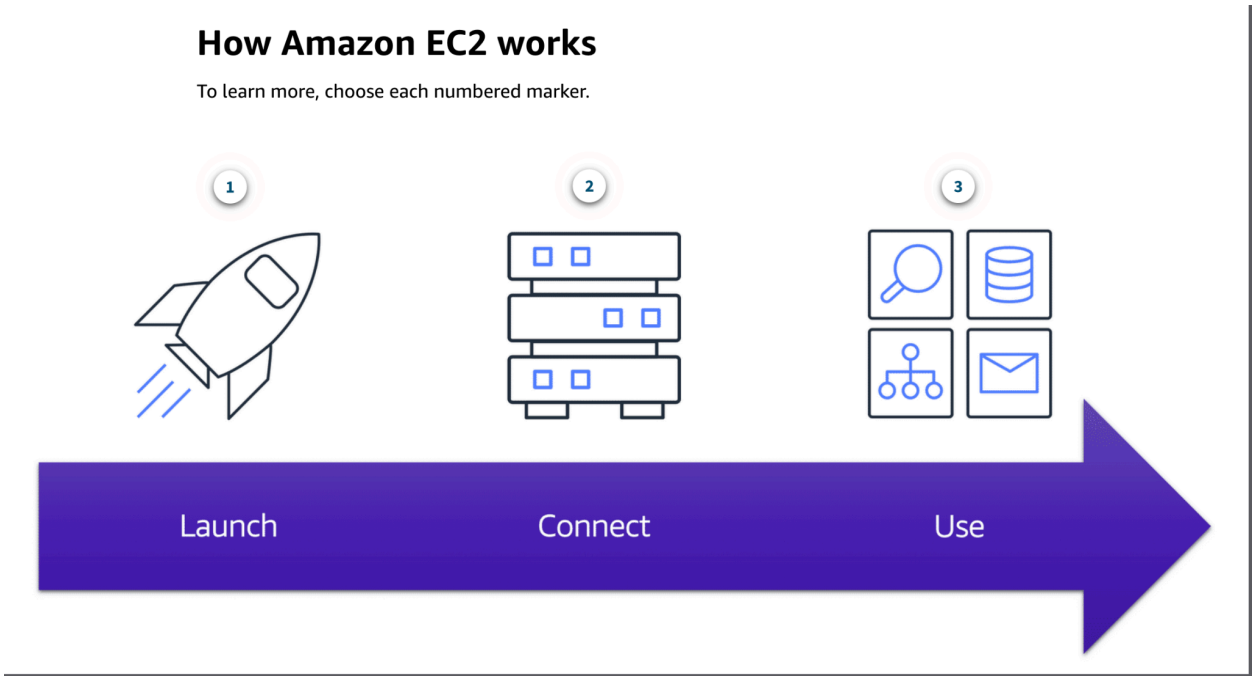
EC2

EC2 runs on top of physical host machines managed by AWS using virtualization technology. When you spin up an EC2 instance, you aren't necessarily taking an entire host to yourself. Instead, you are sharing the host with multiple other instances, otherwise known as virtual machines.And a hypervisor running on the host machine is

responsible for sharing the underlying physical resources between the virtual machines. This idea of sharing underlying hardware is called multitenancy. The hypervisor is responsible for coordinating this multitenancy and it is managed by AWS.

Multitenancy: Sharing underlying hardware between virtual machines
EC2 runs on top of physical host machines managed by AWS using virtualization technology.

EC2 instance, you can choose the operating system based on either Windows or Linux. You can provision thousands of EC2 instances on demand. With a blend of operating systems and configurations to power your business' different applications.
Beyond the OS, you also configure what software you want running on the instance. Whether it's your own internal business applications, simple web apps, or complex web apps, databases or third party software like enterprise software packages, you have complete control over what happens on that instance. EC2 instances are also resizable



Amazon Elastic Compute Cloud (Amazon EC2) provides secure, resizable compute capacity in the cloud as Amazon EC2 instances.

Different types of EC2 instances

- **General purpose**: General purpose instances provide a balance of compute, memory, and networking resources. You can use them for a variety of workloads, such as:application servers, gaming servers, backend servers for enterprise applications, small and medium databases

  Suppose that you have an application in which the resource needs for compute, memory, and networking are roughly equivalent. You might consider running it on a general purpose instance because the application does not require optimization in any single resource area.

  eg:**balances compute, memory, and networking resources**

- **Compute Optimised**: are ideal for compute-bound applications that benefit from high-performance processors. Like general purpose instances, you can use compute optimized instances for workloads such as web, application, and gaming servers.

  However, the difference is compute optimized applications are ideal for high-performance web servers, compute-intensive applications servers, and dedicated gaming servers. You can also use compute optimized instances for **batch processing workloads** that require processing many transactions in a single group.

  Eg: **high-performance processors**

- **Memory Optimised**: are designed to deliver fast performance for workloads that process large datasets in memory. In computing, memory is a temporary storage area. It holds all the data and instructions that a central processing unit (CPU) needs to be able to complete actions. Before a computer program or application is able to run, it is loaded from storage into memory. This preloading process gives the CPU direct access to the computer program.

  Suppose that you have a workload that requires large amounts of data to be preloaded before running an application. This scenario might be a high-performance database or a workload that involves performing real-time processing of a large amount of unstructured data. In these types of use cases, consider using a memory optimized instance. Memory optimized instances enable you to run workloads with high memory needs and receive great performance.

  Eg: **high-performance databases**

- **Accelerated Computing**: use hardware accelerators, or coprocessors, to perform some functions more efficiently than is possible in software running on CPUs. Examples of these functions include floating-point number calculations, graphics processing, and data pattern matching.

In computing, a hardware accelerator is a component that can expedite data processing. Accelerated computing instances are ideal for workloads such as graphics applications, game streaming, and application streaming.

- **Storage Optimised**:are designed for workloads that require high, sequential read and write access to large datasets on local storage. Examples of workloads suitable for storage optimized instances include distributed file systems, data warehousing applications, and high-frequency online transaction processing (OLTP) systems.

  In computing, the term input/output operations per second (IOPS) is a metric that measures the performance of a storage device. It indicates how many different input or output operations a device can perform in one second. Storage optimized instances are designed to deliver tens of thousands of low-latency, random IOPS to applications.

  Eg: **data warehousing applications**

EC2 PRICING

- **On Demand**:- On-Demand Instances are ideal for short-term, irregular workloads that cannot be interrupted. No upfront costs or minimum contracts apply. The instances run continuously until you stop them, and you pay for only the compute time you use.

  Sample use cases for On-Demand Instances include developing and testing applications and running applications that have unpredictable usage patterns. On-Demand Instances are not recommended for workloads that last a year or longer because these workloads can experience greater cost savings using Reserved Instances.
- **Reserved instances**:These are suited for steady-state workloads or ones with predictable usage and offer you up to a 75% discount versus On-Demand pricing.Amazon EC2 pricing option provides a discount when you specify a number of EC2 instances to run a specific OS, instance family and size, and tenancy in one Region.
- **Savings Plan**:Savings Plan offers low prices on EC2 usage in exchange for a commitment to a consistent amount of usage measured in dollars per hour for a one or three-year term. This flexible pricing model can therefore provide savings of up to 72% on your AWS compute usage. This can lower prices on your EC2 usage, regardless of instance family, size, OS, tenancy, or AWS region. This also applies to AWS Fargate and AWS Lambda usage
  EC2 pricing option provides a discount when you make an hourly spend commitment to an instance family and Region for a 1-year or 3-year term

- **Spot Instances:** Spot Instances are ideal for workloads with flexible start and end times, or that can withstand interruptions. Spot Instances use unused Amazon EC2 computing capacity and offer you cost savings at up to 90% off of On-Demand prices.eg: batch workloads.
- **Dedicated Hosts:** Dedicated Hosts are physical servers with Amazon EC2 instance capacity that is fully dedicated to your use. You can use your existing per-socket, per-core, or per-VM software licenses to help maintain license compliance. You can purchase On-Demand Dedicated Hosts and Dedicated Hosts Reservations. Of all the Amazon EC2 options that were covered, Dedicated Hosts are the most expensive.

EC2 SCALING

Amazon EC2 Auto Scaling enables you to automatically add or remove Amazon EC2 instances in response to changing application demand. By automatically scaling your instances in and out as needed, you can maintain a greater sense of application availability.

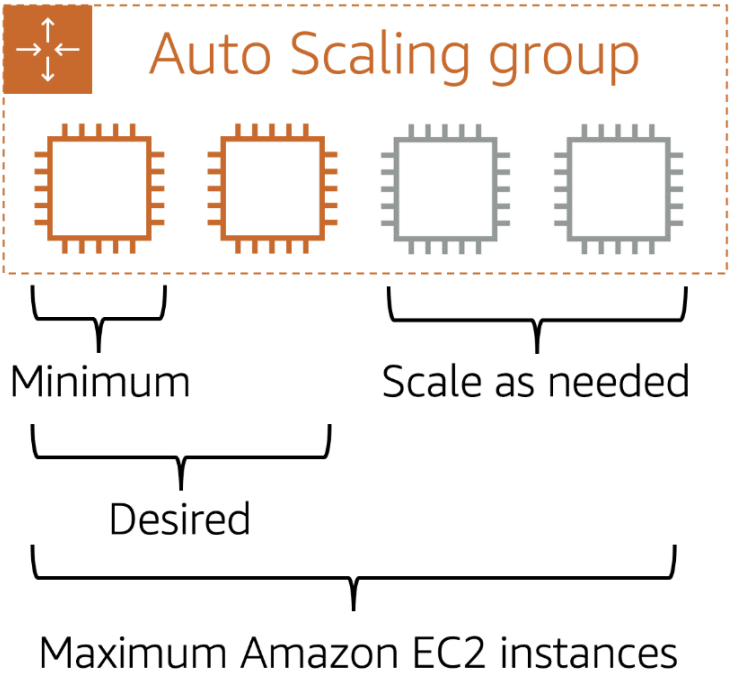
Within Amazon EC2 Auto Scaling, you can use two approaches: dynamic scaling and predictive scaling.

- Dynamic scaling responds to changing demand.
- Predictive scaling automatically schedules the right number of Amazon EC2 instances based on predicted demand.

Suppose that you are preparing to launch an application on Amazon EC2 instances. When configuring the size of your Auto Scaling group, you might set the minimum number of Amazon EC2 instances at one. This means that at all times, there must be at least one Amazon EC2 instance running.

When you create an Auto Scaling group, you can set the minimum number of Amazon EC2 instances. The **minimum capacity** is the number of Amazon EC2 instances that launch immediately after you have created the Auto Scaling group. In this example, the Auto Scaling group has a minimum capacity of one Amazon EC2 instance.

Next, you can set the desired capacity at two Amazon EC2 instances even though your application needs a minimum of a single Amazon EC2 instance to run.



The third configuration that you can set in an Auto Scaling group is the maximum capacity. For example, you might configure the Auto Scaling group to scale out in response to increased demand, but only to a maximum of four Amazon EC2 instances

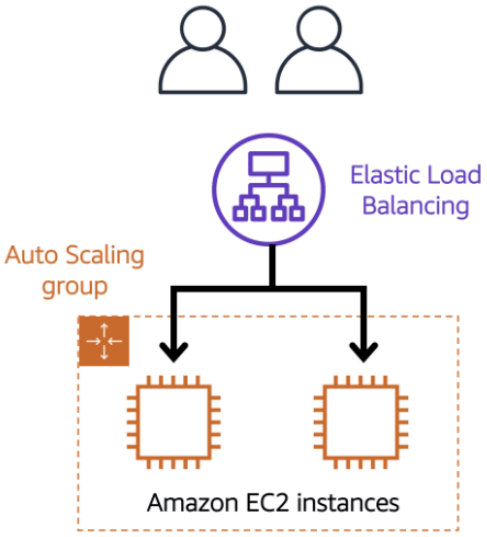
Directing Traffic with Elastic Load Balancing

-ELASTIC LOAD BALANCING

Elastic Load Balancing is the AWS service that automatically distributes incoming application traffic across multiple resources, such as Amazon EC2 instances.

A load balancer acts as a single point of contact for all incoming web traffic to your Auto Scaling group. This means that as you add or remove Amazon EC2 instances in response to the amount of incoming traffic, these requests route to the load balancer first. Then, the requests spread across multiple resources that will handle them. For example, if you have multiple Amazon EC2 instances, Elastic Load Balancing distributes the workload across the multiple instances so that no single instance has to carry the bulk of it.

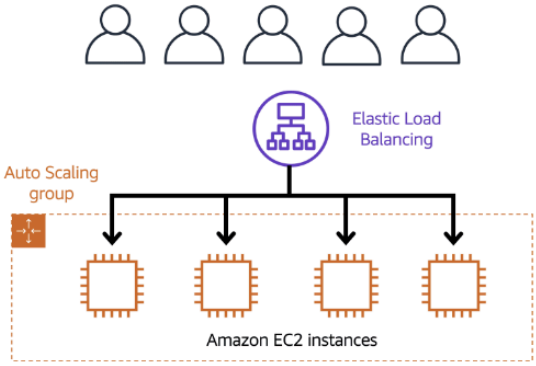
Example: Elastic Load Balancing



Low-demand period

Here's an example of how Elastic Load Balancing works. Suppose that a few customers have come to the coffee shop and are ready to place their orders.

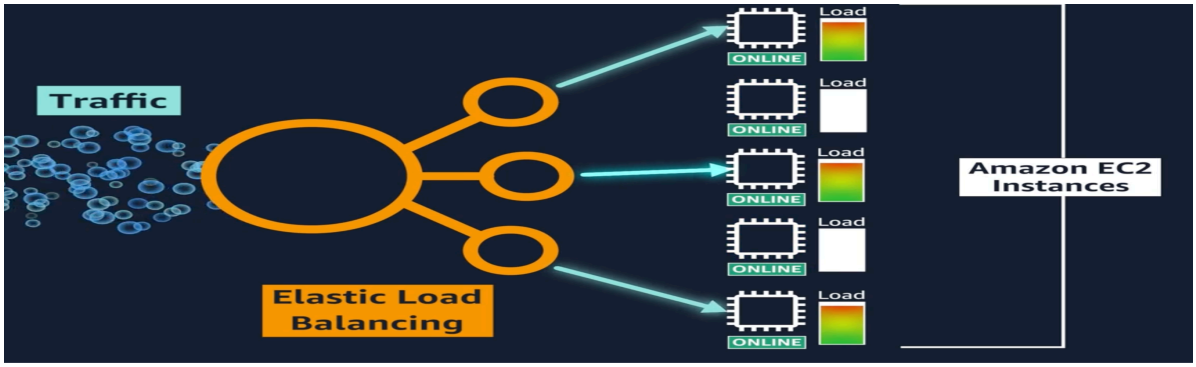
If only a few registers are open, this matches the demand of customers who need service. The coffee shop is less likely to have open registers with no customers. In this example, you can think of the registers as Amazon EC2 instances.



High-demand period

Throughout the day, as the number of customers increases, the coffee shop opens more registers to accommodate them.

Additionally, a coffee shop employee directs customers to the most appropriate register so that the number of requests can evenly distribute across the open registers. You can think of this coffee shop employee as a load balancer.



Messaging and Queuing

Loosely coupled architecture, A single failure won't cause cascading failures. Message queue ie buffer between two applications.

AMAZON SIMPLE QUEUE SERVICE(Amazon SQS) & AMAZON SIMPLE NOTIFICATION SERVICE(Amazon SNS)

Monolithic applications and microservices

Suppose that you have an application with tightly coupled components. These components might include databases, servers, the user interface, business logic, and so on. This type of architecture can be considered a **monolithic application.**
In this approach to application architecture, if a single component fails, other components fail, and possibly the entire application fails.

To help maintain application availability when a single component fails, you can design your application through a microservices approach.

In a **microservices approach**, application components are loosely coupled. In this case, if a single component fails, the other components continue to work because they are communicating with each other. The loose coupling prevents the entire application from failing.

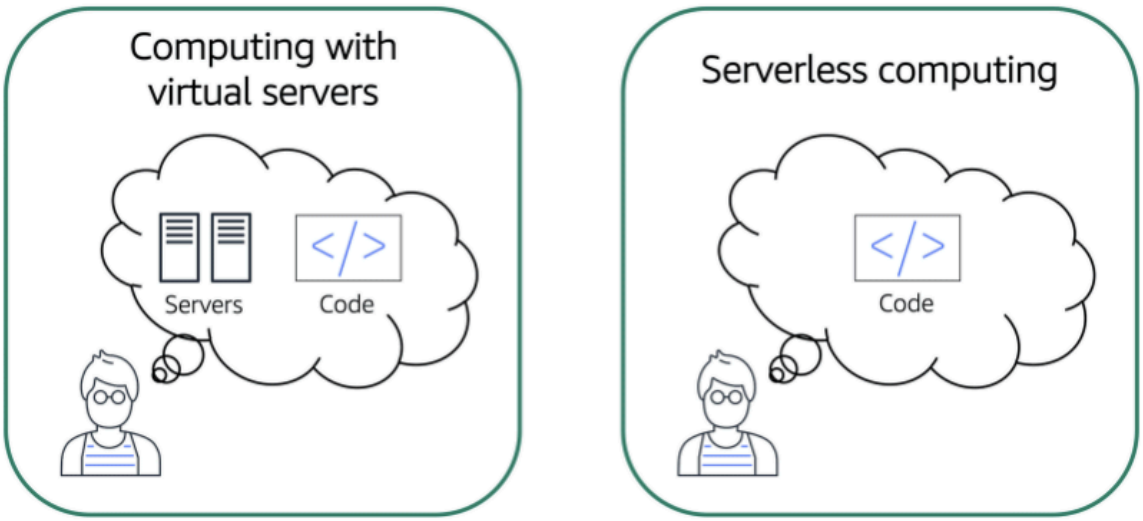
In **Amazon SNS** is a publish/subscribe service. Using Amazon SNS topics, a publisher publishes messages to subscribers**.** subscribers can be web servers, email addresses, AWS Lambda functions, or several other options.

Using **Amazon SQS**, you can send, store, and receive messages between software components, without losing messages or requiring other services to be available.

OTHER SERVICES:

Serverless: you cannot see or access underlying infrastructure or instances that are hosting the application.Managing, scaling, provisioning, high availability and maintaining

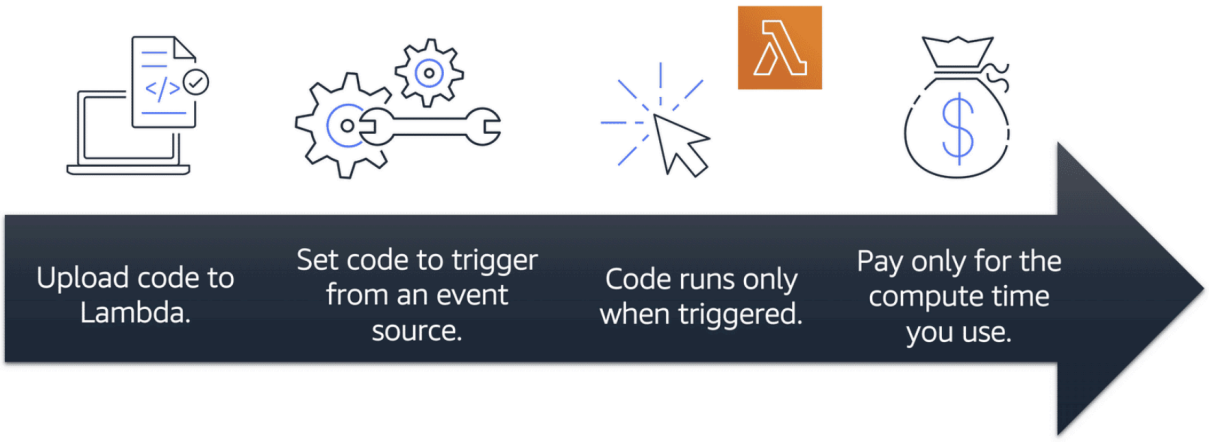
are done.



AWS Lambda

Lambda's a service that allows you to upload your code into what's called a Lambda function. Configure a trigger and from there, the service waits for the trigger. When the trigger is detected, the code is automatically run in a managed environment, an environment you do not need to worry too much about because it is automatically scalable, highly available and all of the maintenance in the environment itself is done by AWS

How AWS Lambda works



***Amazon elastic container service(amazon ECS)**
Amazon Elastic Container Service (Amazon ECS) is a highly scalable, high-performance container management system that enables you to run and scale containerized applications on AWS.

Amazon ECS supports Docker containers. Docker is a software platform that enables you to build, test, and deploy applications quickly. AWS supports the use of open-source Docker Community Edition and subscription-based Docker Enterprise Edition. With Amazon ECS, you can use API calls to launch and stop Docker-enabled applications.

***Amazon elastic kubernetes service (amazon EKS)** these are container orchestration tools, here container is docker container. Container is a package of code.
Amazon Elastic Kubernetes Service (Amazon EKS) is a fully managed service that you can use to run Kubernetes on AWS.

Kubernetes is open-source software that enables you to deploy and manage containerized applications at scale. A large community of volunteers maintains Kubernetes, and AWS actively works together with the Kubernetes community. As new features and functionalities release for Kubernetes applications, you can easily apply these updates to your applications managed by Amazon EKS.

***AWS Fargate**
AWS Fargate is a serverless compute engine for containers. It works with both Amazon ECS and Amazon EKS.

When using AWS Fargate, you do not need to provision or manage servers. AWS Fargate manages your server infrastructure for you. You can focus more on innovating and developing your applications, and you pay only for the resources that are required to run your containers.

if you don't want to even think about using EC2s to host your containers because you either don't need access to the underlying OS or you don't want to manage those EC2 instances, you can use a compute platform called **AWS Fargate**

If you are trying to host traditional applications and want full access to the underlying operating system like Linux or Windows, you are going to want to use EC2. If you are looking to host short running functions, service-oriented or event driven applications and you don't want to manage the underlying environment at all, look into the serverless AWS Lambda. If you are looking to run Docker container-based workloads on AWS, you first need to choose your orchestration tool. Do you want to use Amazon ECS or Amazon

EKS? After you choose your tool, you then need to choose your platform. Do you want to run your containers on EC2 instances that you manage or in a serverless environment like AWS Fargate that is managed for you

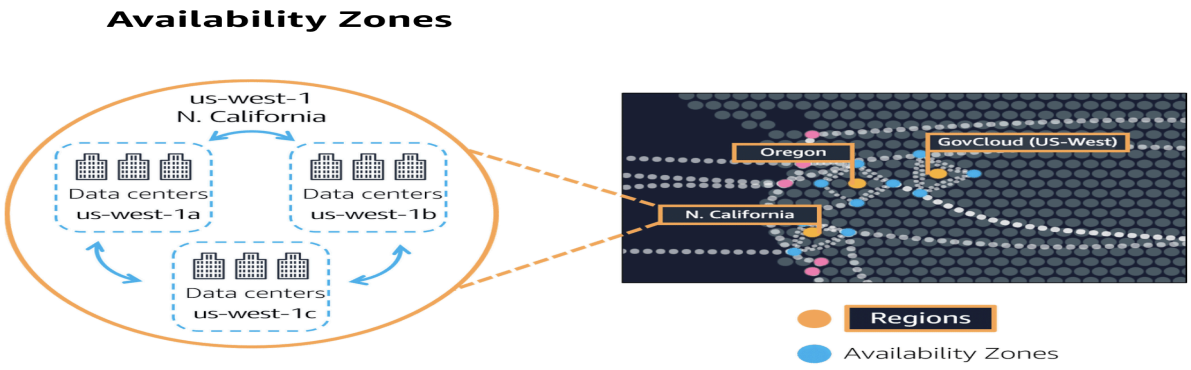
AWS REGIONS(global infrastructure)

Many data centres instead of a single centre in many regions.Each region connected to the other regions through high speed fibre networks. Regions does not share data unless you give the permission to do it. 4 factors in choosing region:

- Compliance: client requirements
- Proximity: choose region where most of the customers are ie customer base
- Feature availability or Available services
- Pricing

AVAILABILITY ZONES:

AWS calls a single data centre or a group of data centres, an Availability Zone or AZ. Each Availability Zone is one or more discrete data centres with redundant power, networking, and connectivity. When you launch an Amazon EC2 instance, it launches a virtual machine on a physical hardware that is installed in an Availability Zone. This means each AWS Region consists of multiple isolated and physically separate Availability Zones within a geographic Region.



A best practice is to run applications across at least two Availability Zones in a Region.
A Region is a geographical area that contains AWS resources.

CDN

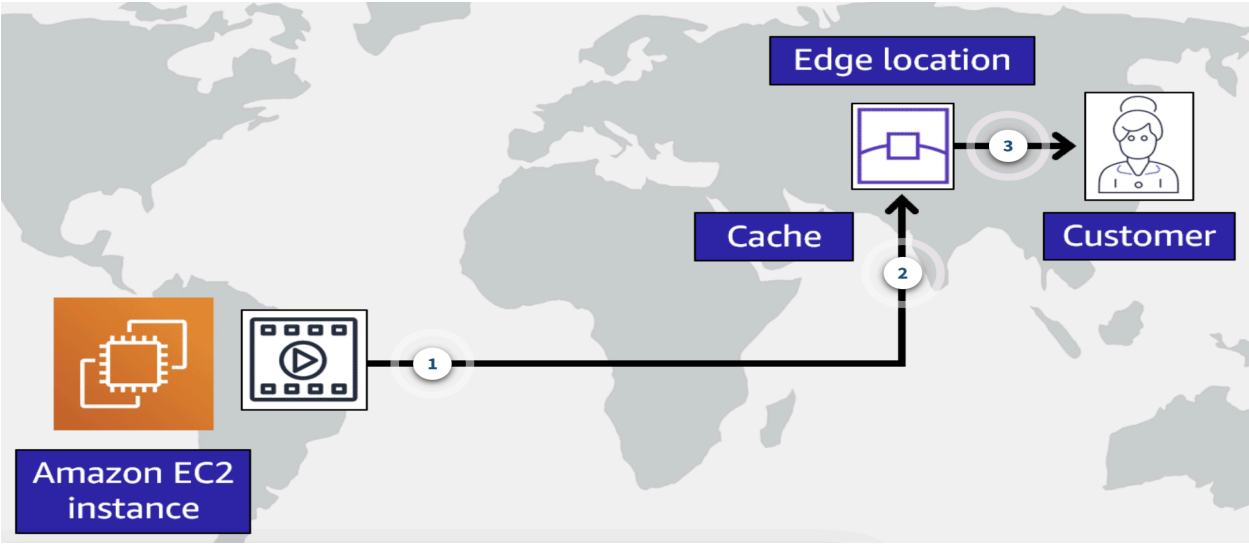
Caching copies of data closer to the customers all around the world uses the concept of CONTENT DELIVERY NETWORKS, or CDNs.

CDNs are commonly used, and on AWS, we call our **CDN Amazon CloudFront**. Amazon CloudFront is a service that helps deliver data, video, applications, and APIs to customers around the world with low latency and high transfer speeds. Amazon CloudFront uses what are called **Edge locations**, all around the world, to help accelerate communication with users, no matter where they are.
Edge locations are separate from Regions, so you can push content from inside a Region to a collection of Edge locations around the world, in order to accelerate communication and content delivery. AWS Edge locations also run more than just CloudFront. They run a domain name service, or DNS, known as **Amazon Route 53.**

AWS outposts
AWS services inside their own building? Well sure. AWS can do that for you. Introducing AWS Outposts, where AWS will basically install a fully operational mini Region, right inside your own data center. That's owned and operated by AWS, using 100% of AWS functionality, but isolated within your own building.

- Regions are geographically isolated areas, where you can access services needed to run your enterprise
- Regions contain Availability Zones that allow you to run across physically separated buildings, tens of miles of separation, while keeping your application logically unified.
- AWS Edge locations run Amazon CloudFront to help get content closer to your customers, no matter where they are in the world

AWS Outposts, where AWS will basically install a fully operational mini Region, right inside your own data centre



INTERACTING WITH AWS

Everything is an api call.

- Aws management console :- test environments, view bills, view monitoring, work with non technical resources.
- Aws command line interface CLI
- Aws software development kit: through various programming languages.

AWS Elastic Beanstalk

AWS Elastic Beanstalk is a service that helps you provision Amazon EC2-based environments. Instead of clicking around the console or writing multiple commands to build out your network, EC2 instances, scaling and Elastic Load Balancers, you can instead provide your application code and desired configurations to the AWS Elastic Beanstalk service, which then takes that information and builds out your environment for you. AWS Elastic Beanstalk also makes it easy to save environment configurations, so they can be deployed again easily. AWS Elastic Beanstalk gives you the convenience of not having to provision and manage all of these pieces separately, while still giving you the visibility and control of the underlying resources. You get to focus on your business application, not the infrastructure.
With AWS Elastic Beanstalk, you provide code and configuration settings, and Elastic Beanstalk deploys the resources necessary to perform the following tasks:

- Adjust capacity
- Load balancing
- Automatic scaling
- Application health monitoring

AWS CloudFormation

Amazon CloudFront is a content delivery service.AWS CloudFormation is an infrastructure as code tool that allows you to define a wide variety of AWS resources in a declarative way using JSON or YAML text-based documents called CloudFormation templates.A declarative format like this allows you to define what you want to build without specifying the details of exactly how to build it.CloudFormation supports many different AWS resources from storage, databases, analytics, machine learning, and more. Once you define your resources in a CloudFormation template, CloudFormation will parse the template and begin provisioning all the resources you defined in parallel. CloudFormation manages all the calls to the backend AWS APIs for you. You can run

the same CloudFormation template in multiple accounts or multiple regions, and it will create identical environments across them.

With AWS CloudFormation, you can treat your infrastructure as code. This means that you can build an environment by writing lines of code instead of using the AWS Management Console to individually provision resources. AWS CloudFormation provisions your resources in a safe, repeatable manner, enabling you to frequently build your infrastructure and applications without having to perform manual actions. It determines the right operations to perform when managing your stack and rolls back changes automatically if it detects errors.

NETWORKING

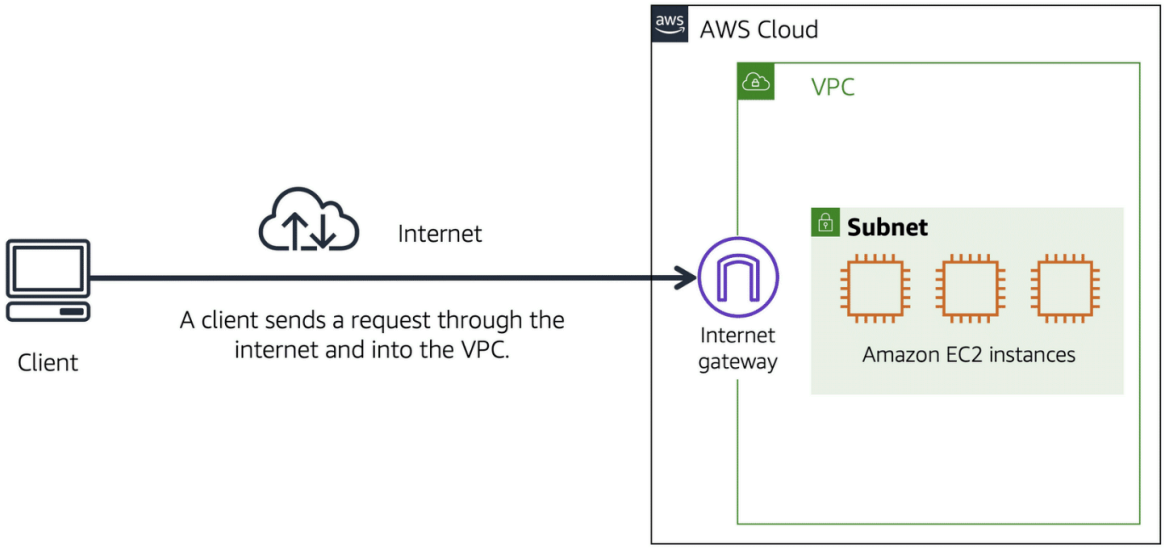
AMAZON VIRTUAL PRIVATE CLOUD:

A VPC lets you provision a logically isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define. These resources can be public facing so they have access to the internet, or private with no internet access, usually for backend services like databases or application servers. The public and private grouping of resources are known as subnets and they are ranges of IP addresses in your VPC.

Connectivity to AWS

Amazon Virtual Private Cloud:
A networking service that you can use to establish boundaries around your AWS resources is **Amazon Virtual Private Cloud (Amazon VPC).**
Amazon VPC enables you to provision an isolated section of the AWS Cloud. In this isolated section, you can launch resources in a virtual network that you define. Within a virtual private cloud (VPC), you can organize your resources into subnets. A subnet is a section of a VPC that can contain resources such as Amazon EC2 instances.

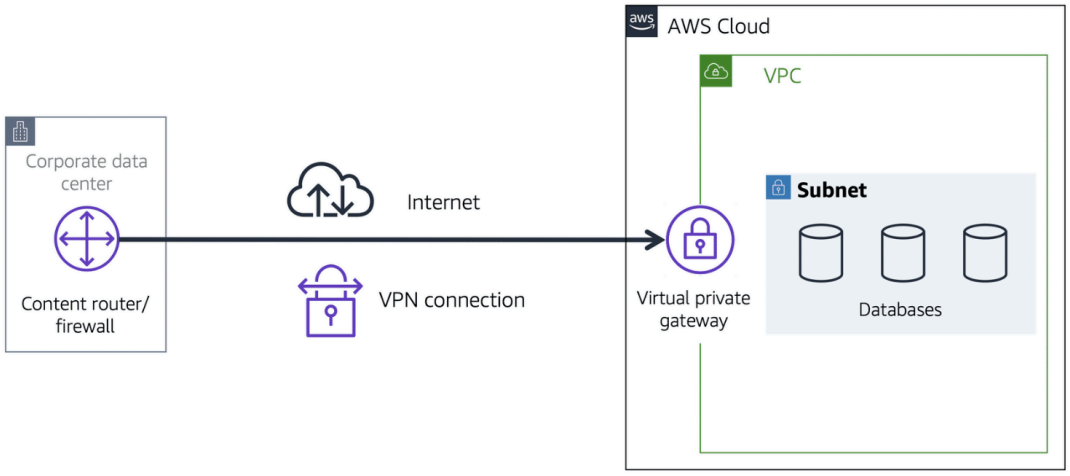
Internet Gateway
To allow public traffic from the internet to access your VPC, you attach an **internet gateway** to the VPC.An internet gateway is a connection between a VPC and the internet. You can think of an internet gateway as being similar to a doorway that customers use to enter the coffee shop. Without an internet gateway, no one can access the resources within your VPC.



VPC that includes only private resources

Virtual private gateway

To access private resources in a VPC, you can use a virtual private gateway.

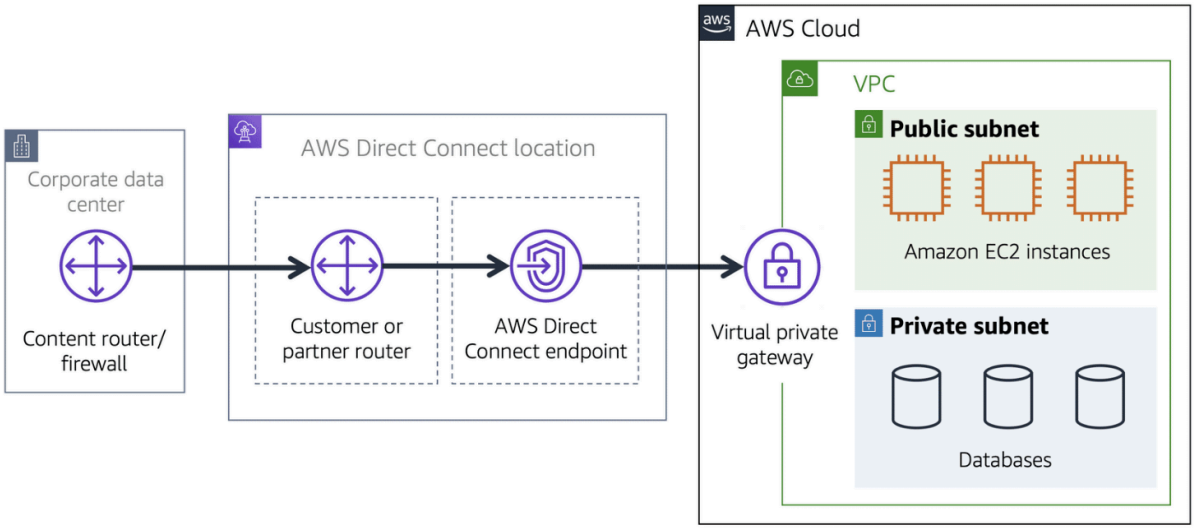


A virtual private gateway enables you to establish a virtual private network (VPN) connection between your VPC and a private network, such as an on-premises data center or internal corporate network. A virtual private gateway allows traffic into the VPC only if it is coming from an approved network.

AWS Direct Connect

AWS Direct Connect is a service that lets you to establish a dedicated private connection between your data center and a VPC.The private connection that AWS

Direct Connect provides helps you to reduce network costs and increase the amount of bandwidth that can travel through your network.



Subnets and network access control list

AWS has a wide range of tools that cover every layer of security: network hardening, application security, user identity, authentication and authorization, distributed denial-of-service or DDoS prevention, data integrity, encryption, much more

packets are messages from the internet, and every packet that crosses the subnet boundaries gets checked against something called a network access control list or network ACL. This check is to see if the packet has permissions to either leave or enter the subnet based on who it was sent from and how it's trying to communicate.

N**etwork ACLs** check traffic going into and leaving a subnet, just like passport control. network ACL only gets to evaluate a packet if it crosses a subnet boundary, in or out. It doesn't evaluate if a packet can reach a specific EC2 instance or not.

you need instance level network security as well. To solve instance level access questions, we introduce **security groups**

Every EC2 instance, when it's launched, automatically comes with a security group. And by default, the security group does not allow any traffic into the instance at all.

Security Group : Stateful
Network ACL: Stateless

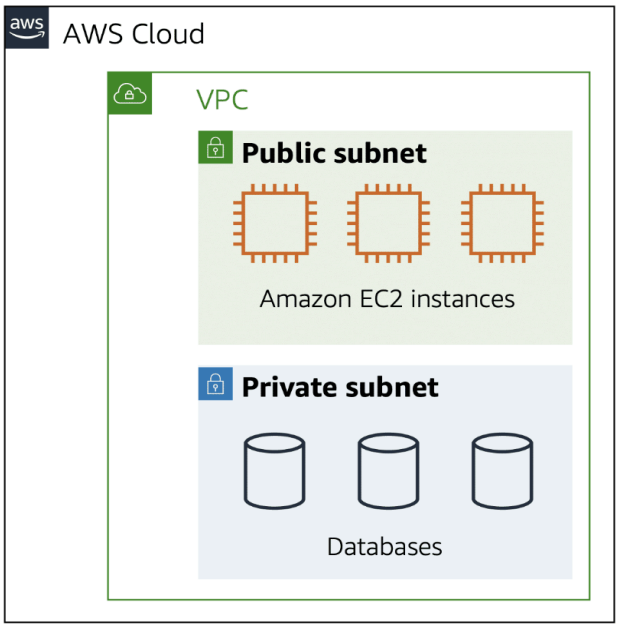
Network ACL is for subnet region and Security groups are for ec2 instances inside it.

Subnets:
A subnet is a section of a VPC in which you can group resources based on security or operational needs. Subnets can be public or private.

Public subnets contain resources that need to be accessible by the public, such as an online store's website.

Private subnets contain resources that should be accessible only through your private network, such as a database that contains customers' personal information and order histories.

In a VPC, subnets can communicate with each other. For example, you might have an application that involves Amazon EC2 instances in a public subnet communicating with databases that are located in a private subnet.



Network Traffic in VPC
When a customer requests data from an application hosted in the AWS Cloud, this request is sent as a packet. A packet is a unit of data sent over the internet or a network. It enters into a VPC through an internet gateway. Before a packet can enter

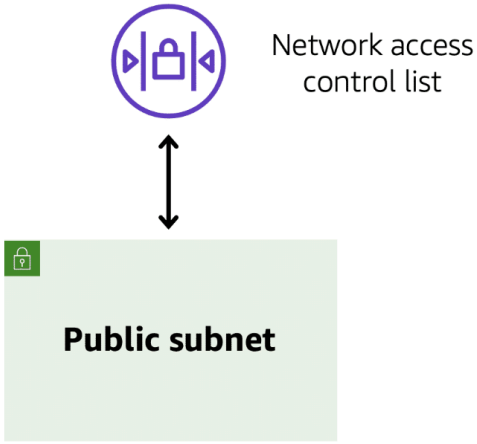
into a subnet or exit from a subnet, it checks for permissions. These permissions indicate who sent the packet and how the packet is trying to communicate with the resources in a subnet.The VPC component that checks packet permissions for subnets is a network access control list (ACL).

Network ACL

A network ACL is a virtual firewall that controls inbound and outbound traffic at the subnet level.

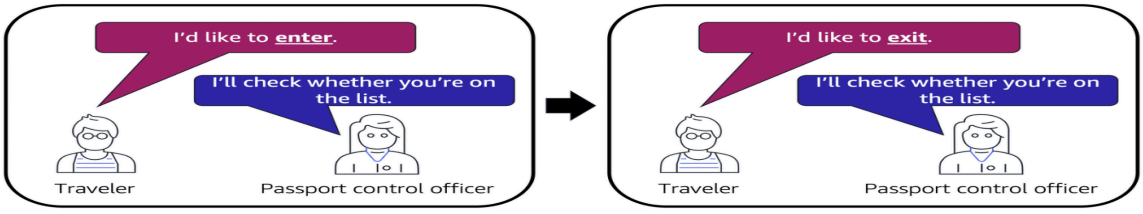
Each AWS account includes a default network ACL. When configuring your VPC, you can use your account's default network ACL or create custom network ACLs.

By default, your account's default network ACL **allows** all inbound and outbound traffic, but you can modify it by adding your own rules. For custom network ACLs, all inbound and outbound traffic is denied until you add rules to specify which traffic to allow. Additionally, all network ACLs have an explicit deny rule. This rule ensures that if a packet doesn't match any of the other rules on the list, the packet is denied.



Stateless Packet Filtering

Network ACLs perform stateless packet filtering. They remember nothing and check packets that cross the subnet border each way: inbound and outbound.

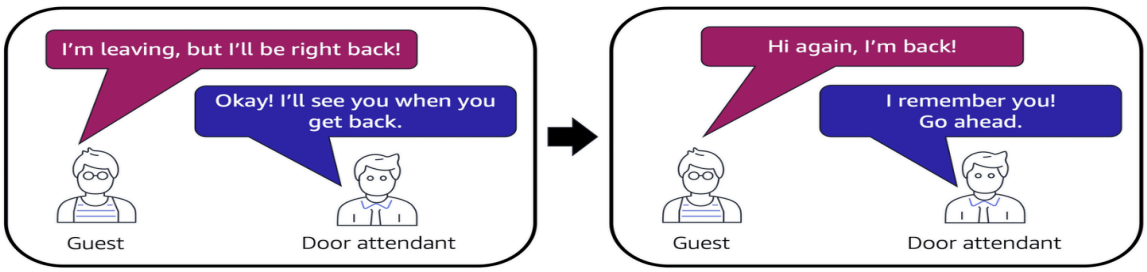


After a packet has entered a subnet, it must have its permissions evaluated for resources within the subnet, such as Amazon EC2 instances.

The VPC component that checks packet permissions for an Amazon EC2 instance is a **security group.**

Security Group
A security group is a virtual firewall that controls inbound and outbound traffic for an Amazon EC2 instance.
a security group **denies** all inbound traffic and allows all outbound traffic

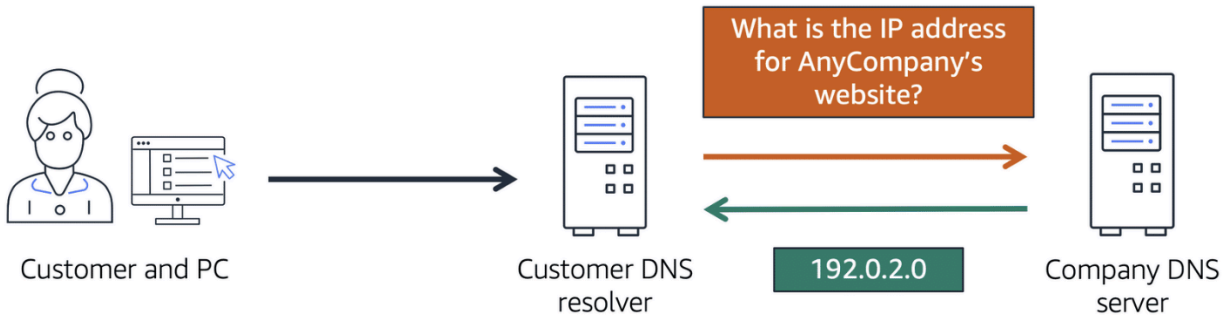
Stateful Packet Filtering
Security groups perform stateful packet filtering. They remember previous decisions made for incoming packets.



Private Subnet: isolate database containing customer's personal informations
Virtual Private Gateway: Create VPN connection between VPC and internal corporate network
Public Subnet: Support customer facing website
AWS Direct Connect : Establish a dedicated connection between on-premises data center and VPC

GLOBAL NETWORKING

Domain Name System



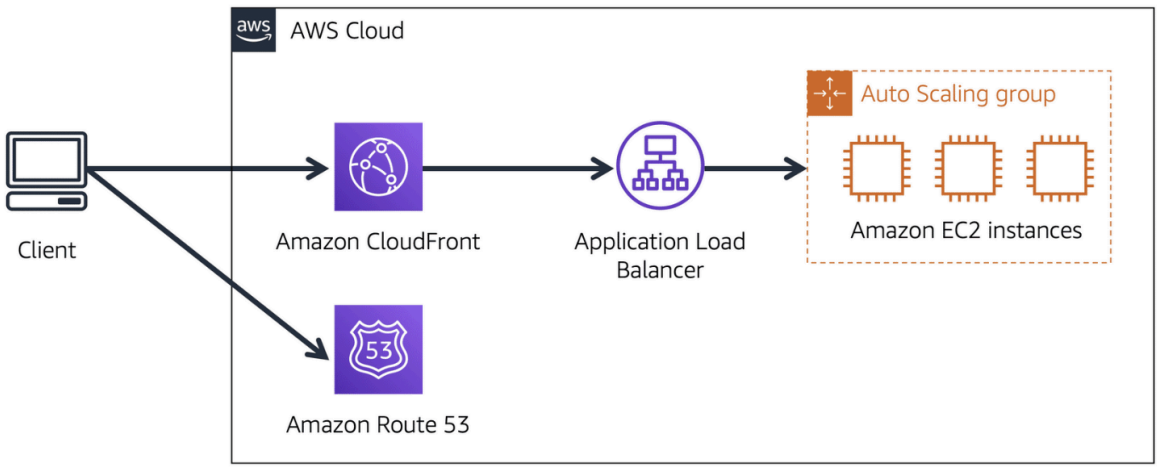
Amazon Route 53

Amazon Route 53(opens in a new tab) is a DNS web service. It gives developers and businesses a reliable way to route end users to internet applications hosted in AWS.

Amazon Route 53 connects user requests to infrastructure running in AWS (such as Amazon EC2 instances and load balancers). It can route users to infrastructure outside of AWS.

Another feature of Route 53 is the ability to manage the DNS records for domain names. You can register new domain names directly in Route 53. You can also transfer DNS records for existing domain names managed by other domain registrar

Example: **How Amazon Route 53 and Amazon CloudFront deliver content**



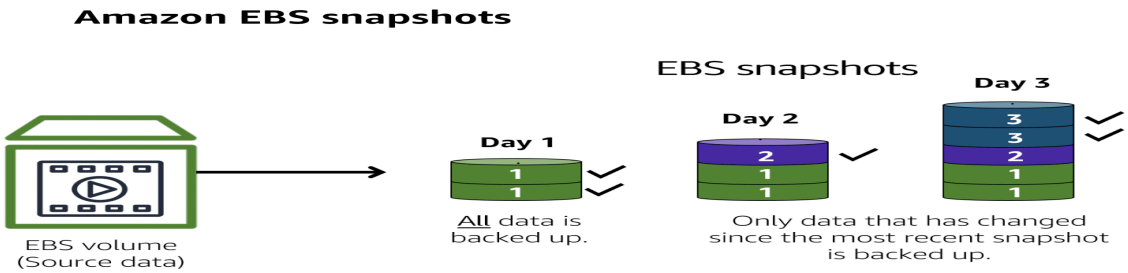
- A customer requests data from the application by going to AnyCompany's website.
- Amazon Route 53 uses DNS resolution to identify AnyCompany.com's corresponding IP address, 192.0.2.0. This information is sent back to the customer.
- The customer's request is sent to the nearest edge location through Amazon CloudFront.
- Amazon CloudFront connects to the Application Load Balancer, which sends the incoming packet to an Amazon EC2 instance.

STORAGE AND DATABASES

Instance stores and Amazon Elastic Block Store(Amazon EBS)

An **instance store** provides temporary block-level storage for an Amazon EC2 instance. An instance store is disk storage that is physically attached to the host computer for an EC2 instance, and therefore has the same lifespan as the instance. When the instance is terminated, you lose any data in the instance store.

Amazon Elastic Block Store (Amazon EBS) is a service that provides block-level storage volumes that you can use with Amazon EC2 instances. If you stop or terminate an Amazon EC2 instance, all the data on the attached EBS volume remains available. To create an EBS volume, you define the configuration (such as volume size and type) and provision it. After you create an EBS volume, it can attach to an Amazon EC2 instance.Because EBS volumes are for data that needs to persist, it's important to back up the data. You can take incremental backups of EBS volumes by creating Amazon EBS snapshots.

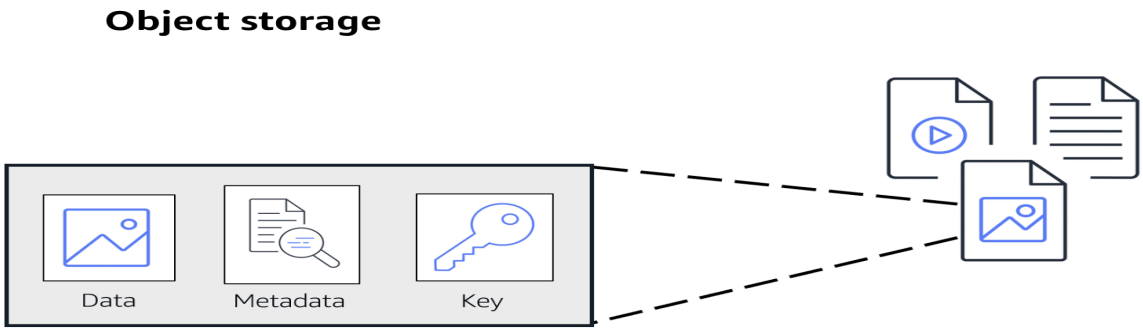


An **EBS snapshot** is an incremental backup. This means that the first backup taken of a volume copies all the data. For subsequent backups, only the blocks of data that have changed since the most recent snapshot are saved.

Incremental backups are different from full backups, in which all the data in a storage volume copies each time a backup occurs. The full backup includes data that has not changed since the most recent backup.

Amazon Simple Storage Service(Amazon s3)

Store and retrieve unlimited amount of data.Store data as **objects**. Objects stored in **buckets**.Maximum upload object size is **5TB**. Can create multiple buckets.



In **object storage** each object contains data metadata and a key.The data might be an image, video, text document, or any other type of file. Metadata contains information about what the data is, how it is used, the object size, and so on. An object's key is its unique identifier.

Amazon Simple Storage Service (Amazon S3) is a service that provides object-level storage. Amazon S3 stores data as objects in buckets.The maximum file size for an object in Amazon S3 is 5 TB. When you upload a file to Amazon S3, you can set permissions to control visibility and access to it. You can also use the Amazon S3 versioning feature to track changes to your objects over time.

Amazon S3 storage classes

- **S3 standard** : Designed for **frequently accessed data**, Stores data in a minimum of three Availability Zones. Amazon S3 Standard provides high availability for objects. This makes it a good choice for a wide range of use cases, such as websites, content distribution, and data analytics. Amazon S3 Standard has a higher cost than other storage classes intended for infrequently accessed data and archival storage

- **S3 standard infrequent Access(s3 standard IA):** Ideal for i**nfrequently accessed data** ,Similar to Amazon S3 Standard but has a lower storage price and higher retrieval price. Amazon S3 Standard-IA is ideal for data infrequently accessed but requires high availability when needed. Both Amazon S3 Standard and Amazon S3 Standard-IA store data in a minimum of three Availability Zones. Amazon S3 Standard-IA provides the same level of availability as Amazon S3 Standard but with a lower storage price and a higher retrieval price.
- **S3 one zone-infrequent access(s3 onezone IA):** Stores data in a single Availability Zone, Has a lower storage price than Amazon S3 Standard-IA. One Zone-IA stores data in a single Availability Zone. This makes it a good storage class to consider if the following conditions apply: You want to save costs on storage.You can easily reproduce your data in the event of an Availability Zone failure.
- **S3 intelligent tiering:** Ideal for data with unknown or c**hanging access patterns** Requires a small monthly monitoring and automation fee per object. In the S3 Intelligent-Tiering storage class, Amazon S3 monitors objects' access patterns. If you haven't accessed an object for 30 consecutive days, Amazon S3 automatically moves it to the infrequent access tier, S3 Standard-IA. If you access an object in the infrequent access tier, Amazon S3 automatically moves it to the frequent access tier, S3 Standard.
- **S3 glacier instant retrieval:** Works well for **archived data** that requires immediate access. Can retrieve objects within a few milliseconds. When you decide between the options for archival storage, consider how quickly you must retrieve the archived objects. You can retrieve objects stored in the S3 Glacier Instant Retrieval storage class within milliseconds, with the same performance as S3 Standard.
- **S3 glacier flexible retrieval:** Low-cost storage designed for **data archiving**. Able to retrieve objects within a few minutes to hours. S3 Glacier Flexible Retrieval is a low-cost storage class that is ideal for data archiving. For example, you might use this storage class to store archived customer records or older photos and video files. You can retrieve your data from S3 Glacier Flexible Retrieval from 1 minute to 12 hours.
- **S3 glacier deep archive:** Lowest-cost object storage class ideal for **archiving**. Able to retrieve objects within 12 hours. S3 Deep Archive supports long-term retention and digital preservation for data that might be accessed once or twice in a year. This storage class is the lowest-cost storage in the AWS Cloud, with data retrieval from 12 to 48 hours. All objects from this storage class are replicated and stored across at least three geographically dispersed Availability Zones.
- **S3 Outposts:** Creates S3 buckets on Amazon S3 Outposts. Makes it easier to retrieve, store, and access data on AWS Outposts. Amazon S3 Outposts delivers object storage to your on-premises AWS Outposts environment. Amazon S3 Outposts is designed to store data durably and redundantly across multiple devices and servers on your Outposts. It works well for workloads with local data residency requirements that must satisfy demanding performance needs by keeping data close to on-premises applications

Amazon EBS vs Amazon S3

Amazon EBS:

- Size up to 16tb
- Survive termination of their ec2 instance.
- Solid state by default.
- HDD options.

Amazon s3:

- Unlimited storage.
- Individual objects up to 5tb
- Write once read many
- 99.99999999% durable.

This means, if you are using complete objects or only occasional changes, S3 is victorious. If you are doing complex read, write, change functions, then, absolutely, EBS is your knockout winner.

Amazon Elastic File System(Amazon EFS)

Amazon Elastic File System (Amazon EFS)(opens in a new tab) is a scalable file system used with AWS Cloud services and on-premises resources. As you add and remove files, Amazon EFS grows and shrinks automatically. It can scale on demand to petabytes without disrupting applications.

Amazon EBS vs Amazon EFS

EBS:
An Amazon EBS volume stores data in a single Availability Zone.
To attach an Amazon EC2 instance to an EBS volume, both the Amazon EC2 instance and the EBS volume must reside within the same Availability Zone.
Volumes attached to EC2 instance
Availability zone level resource
Need to be in the same availability zone to attach ec2 instance.
Volumes do not automatically scale.

EFS:
Amazon EFS is a regional service. It stores data in and across multiple Availability Zones.

The duplicate storage enables you to access data concurrently from all the Availability Zones in the Region where a file system is located. Additionally, on-premises servers can access Amazon EFS using AWS Direct Connect.
Multiples instances reading and writing simultaneously.
Regional resource.
Automatically scales.

Amazon Relational Database Services(Amazon RDS)

Amazon RDS is a managed service that automates tasks such as hardware provisioning, database setup, patching, and backups. With these capabilities, you can spend less time completing administrative tasks and more time using data to innovate your applications. You can integrate Amazon RDS with other services to fulfil your business and operational needs, such as using AWS Lambda to query your database from a serverless application
Amazon RDS provides a number of different security options. Many Amazon RDS database engines offer encryption at rest (protecting data while it is stored) and encryption in transit (protecting data while it is being sent and received).

Amazon RDS Engines

- Amazon Aurora
- PostgreSQL
- MySQL
- MariaDB
- Oracle Database
- Microsoft SQL Server

Amazon Aurora

Amazon Aurora is an **enterprise-class relational database**. It is compatible with MySQL and PostgreSQL relational databases. It is up to five times faster than standard MySQL databases and up to three times faster than standard PostgreSQL databases. Amazon Aurora helps to reduce your database costs by reducing unnecessary input/output (I/O) operations, while ensuring that your database resources remain reliable and available.
Consider Amazon Aurora if your workloads require high availability. It replicates six copies of your data across three Availability Zones and continuously backs up your data to Amazon S3.

Amazon DynamoDB

A serverless database.A non relational database. Millisecond response time.
Amazon DynamoDB is a key-value database service.DynamoDB is serverless, which means that you do not have to provision, patch, or manage servers. You also do not have to install, maintain, or operate software.DynamoDB automatically scales to adjust for changes in capacity while maintaining consistent performance. This makes it a suitable choice for use cases that require high performance while scaling.

RDS:

- Automatic high availability. Recovery provided.
- Customer ownership of data.
- Customer ownership of schema.
- Customer control of the network.

DynamoDB:

- Key value
- Massive throughput capabilities
- PB(peta Byte) size potential
- Granular API access.

Amazon RedShift

Amazon Redshift is a **data warehousing** service that you can use for big data analytics. It offers the ability to collect data from many sources and helps you to understand relationships and trends across your data. Mainly business intelligence.

AWS Data Migration Services

AWS Database Migration Service (AWS DMS) enables you to migrate relational databases, non relational databases, and other types of data stores.
With AWS DMS, you move data between a source database and a target database. The source and target databases can be of the same type or different types. During the migration, your source database remains operational, reducing downtime for any applications that rely on the database.
For example, suppose that you have a MySQL database that is stored on premises in an Amazon EC2 instance or in Amazon RDS. Consider the MySQL database to be your source database. Using AWS DMS, you could migrate your data to a target database, such as an Amazon Aurora database.
Homogenous : mysql to amazon RDS for mysql
Heterogeneous : 2 steps, we first need to convert them using the AWS Schema Conversion Tool. This will convert the source schema and code to match that of the target database. The next step is then to use DMS to migrate data from the source database to the target database.

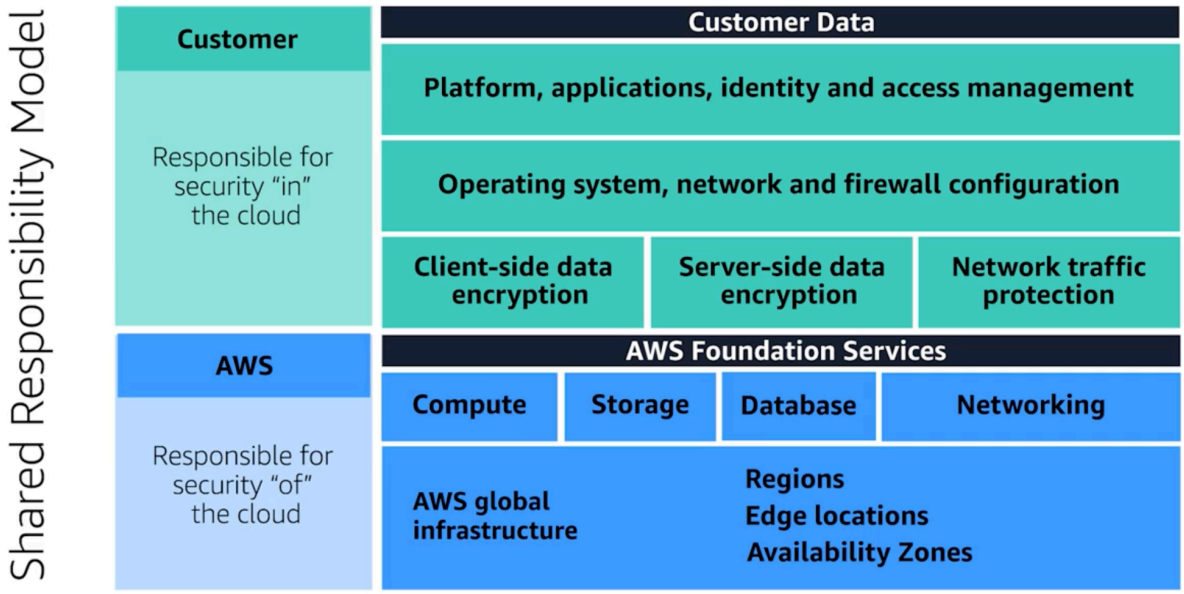
But these are not the only use cases for DMS. Others include d**evelopment and test database migrations, database consolidation, and even continuous database replication.**
Development and test migration is when you want to develop this to test against production data, but without affecting production users. In this case, you use DMS to migrate a copy of your production database to your dev or test environments, either once-off or continuously.
Database consolidation is when you have several databases and want to consolidate them into one central database.
Finally, **continuous replication** is when you use DMS to perform continuous data replication. This could be for disaster recovery or because of geographic separation.

Additional Database Services

- **Amazon DocumentDB** : Amazon DocumentDB is a document database service that supports MongoDB workloads. (MongoDB is a document database program.)
- **Amazon Neptune** : Amazon Neptune is a graph database service. You can use Amazon Neptune to build and run applications that work with highly connected datasets, such as recommendation engines, fraud detection, and knowledge graphs.
- **Amazon Quantum Ledger Databases(Amazon QLDB)**: Amazon Quantum Ledger Database (Amazon QLDB) is a ledger database service. You can use Amazon QLDB to review a complete history of all the changes that have been made to your application data.
- **Amazon Managed Blockchain** : Amazon Managed Blockchain is a service that you can use to create and manage blockchain networks with open-source frameworks. Blockchain is a distributed ledger system that lets multiple parties run transactions and share data without a central authority.
- **Amazon ElastiCache** : Amazon ElastiCache is a service that adds caching layers on top of your databases to help improve the read times of common requests. It supports two types of data stores: Redis and Memcached.
- **Amazon DynamoDB Accelerator**: Amazon DynamoDB Accelerator (DAX) is an in-memory cache for DynamoDB. It helps improve response times from single-digit milliseconds to microseconds.

SECURITY

AWS Shared Responsibility Model



Customers: Security in the cloud
Customers are responsible for the security of everything that they create and put in the AWS Cloud.
When using AWS services, you, the customer, maintain complete control over your content. You are responsible for managing security requirements for your content, including which content you choose to store on AWS, which AWS services you use, and who has access to that content. You also control how access rights are granted, managed, and revoked.
The security steps that you take will depend on factors such as the services that you use, the complexity of your systems, and your company's specific operational and security needs. Steps include selecting, configuring, and patching the operating systems that will run on Amazon EC2 instances, configuring security groups, and managing user accounts.

AWS: Security of the cloud
AWS is responsible for security of the cloud.
AWS operates, manages, and controls the components at all layers of infrastructure. This includes areas such as the host operating system, the virtualization layer, and even the physical security of the data centres from which services operate.

AWS is responsible for protecting the global infrastructure that runs all of the services offered in the AWS Cloud. This infrastructure includes AWS Regions, Availability Zones, and edge locations

AWS manages the security of the cloud, specifically the physical infrastructure that hosts your resources, which include:

- Physical security of data centres
- Hardware and software infrastructure
- Network infrastructure
- Virtualization infrastructure

Although you cannot visit AWS data centres to see this protection firsthand, AWS provides several reports from third-party auditors. These auditors have verified its compliance with a variety of computer security standards and regulations.

USER PERMISSIONS AND ACCESS

AWS IDENTITY AND ACCESS MANAGEMENT(IAM)

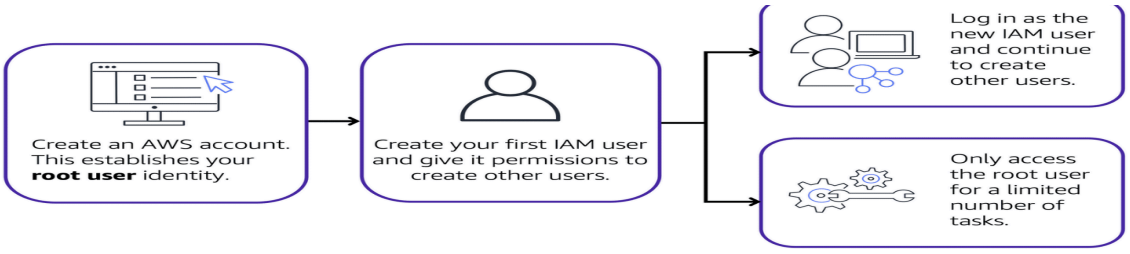
AWS Identity and Access Management (IAM) enables you to manage access to AWS services and resources securely.

IAM gives you the flexibility to configure access based on your company's specific operational and security needs. You do this by using a combination of IAM features, which are explored in detail in this lesson:

- IAM users, groups, and roles
- IAM policies
- Multi-factor authentication

AWS Root User

When you first create an AWS account, you begin with an identity known as the root user. The root user is accessed by signing in with the email address and password that you used to create your AWS account. It has **complete access** to all the AWS services and resources in the account.



Instead, use the root user to create your first IAM user and assign it permissions to create other users.Then, continue to create other IAM users, and access those

identities for performing regular tasks throughout AWS. Only use the root user when you need to perform a limited number of tasks that are only available to the root user.

IAM Users

An IAM user is an identity that you create in AWS. It represents the person or application that interacts with AWS services and resources. It consists of a name and credentials.
By default, when you create a new IAM user in AWS, it **has no permissions associated with it**. To allow the IAM user to perform specific actions in AWS, such as launching an Amazon EC2 instance or creating an Amazon S3 bucket, you must grant the IAM user the necessary permissions.

We recommend that you create individual IAM users for each person who needs to access AWS. Even if you have multiple employees who require the same level of access, you should create individual IAM users for each of them. This provides additional security by allowing each IAM user to have a unique set of security credentials.

IAM Policies

An IAM policy is a document that allows or denies permissions to AWS services and resources.
IAM policies enable you to customise users' levels of access to resources. For example, you can allow users to access all of the Amazon S3 buckets within your AWS account, or only a specific bucket.

Follow the security **principle of least privilege** when granting permissions.

Example: IAM policy

Here's an example of how IAM policies work. Suppose that the coffee shop owner has to create an IAM user for a newly hired cashier. The cashier needs access to the receipts kept in an Amazon S3 bucket with the ID: AWSDOC-EXAMPLE-BUCKET.

```
{
  "Version": "2012-10-17",
  "Statement": {
    "Effect": "Allow",
    "Action": "s3:ListObject",
    "Resource": "arn:aws:s3:::
AWSDOC-EXAMPLE-BUCKET"
  }
}
```

This example IAM policy allows permission to access the objects in the Amazon S3 bucket with ID: *AWSDOC-EXAMPLE-BUCKET*.

In this example, the IAM policy is allowing a specific action within Amazon S3: ListObject. The policy also mentions a specific bucket ID: AWSDOC-EXAMPLE-BUCKET. When the owner attaches this policy to the cashier's IAM user, it will allow the cashier to view all of the objects in the AWSDOC-EXAMPLE-BUCKET bucket.

If the owner wants the cashier to be able to access other services and perform other actions in AWS, the owner must attach additional policies to specify these services and actions.

IAM Groups

An IAM group is a collection of IAM users. When you assign an IAM policy to a group, all users in the group are granted permissions specified by the policy.

Assigning IAM policies at the group level also makes it easier to adjust permissions when an employee transfers to a different job. For example, if a cashier becomes an inventory specialist, the coffee shop owner removes them from the “Cashiers” IAM group and adds them into the “Inventory Specialists” IAM group. This ensures that employees have only the permissions that are required for their current role.

IAM Roles

An IAM role is an identity that you can assume to gain temporary access to permissions.

When the employee needs to switch to a different task, they give up their access to one workstation and gain access to the next workstation

Before an IAM user, application, or service can assume an IAM role, they must be granted permissions to switch to the role. When someone assumes an IAM role, they abandon all previous permissions that they had under a previous role and assume the permissions of the new role.

IAM roles are ideal for situations in which access to services or resources needs to be granted temporarily, instead of long-term.

AWS ORGANISATIONS

- Centralised Management
- Consolidated Billing
- Hierarchical grouping of accounts
- AWS service and API action access control

You can use AWS Organizations to consolidate and manage multiple AWS accounts within a central location.

When you create an organisation, AWS Organizations automatically creates a root, which is the parent container for all the accounts in your organisation.

In AWS Organizations, you can centrally control permissions for the accounts in your organisation by using **service control policies (SCPs)**. SCPs enable you to place restrictions on the AWS services, resources, and individual API actions that users and roles in each account can access..

In AWS Organizations, you can apply service control policies (SCPs) to the organisation root, an individual member account, or an OU. An SCP affects all IAM users, groups, and roles within an account, including the AWS account root user.

You can apply IAM policies to IAM users, groups, or roles. You cannot apply an IAM policy to the AWS account root user

Consolidated billing is another feature of AWS Organizations.

Organisational Units

In AWS Organizations, you can group accounts into organisational units (OUs) to make it easier to manage accounts with similar business or security requirements. When you apply a policy to an OU, all the accounts in the OU automatically inherit the permissions specified in the policy.

By organising separate accounts into OUs, you can more easily isolate workloads or applications that have specific security requirements. For instance, if your company has accounts that can access only the AWS services that meet certain regulatory requirements, you can put these accounts into one OU. Then, you can attach a policy to the OU that blocks access to all other AWS services that do not meet the regulatory requirements.

By grouping your accounts into OUs, you can easily give them access to the services and resources that they need. You also prevent them from accessing any services or resources that they do not need.

COMPLIANCE

AWS Artifact

AWS Artifact is a service that provides on-demand access to AWS security and compliance reports and select online agreements. AWS Artifact consists of two main sections: **AWS Artifact Agreements** and **AWS Artifact Reports.**

AWS Artifact Agreements:

In AWS Artifact Agreements, you can review, accept, and manage agreements for an individual account and for all your accounts in AWS Organizations.

AWS Artifact Reports:

AWS Artifact Reports provide compliance reports from third-party auditors. These auditors have tested and verified that AWS is compliant with a variety of global, regional, and industry-specific security standards and regulations. AWS Artifact Reports remains up to date with the latest reports released. You can provide the AWS audit artifacts to your auditors or regulators as evidence of AWS security controls.

Customer Compliance Centre

In the Customer Compliance Center, you can read customer compliance stories to discover how companies in regulated industries have solved various compliance, governance, and audit challenges.

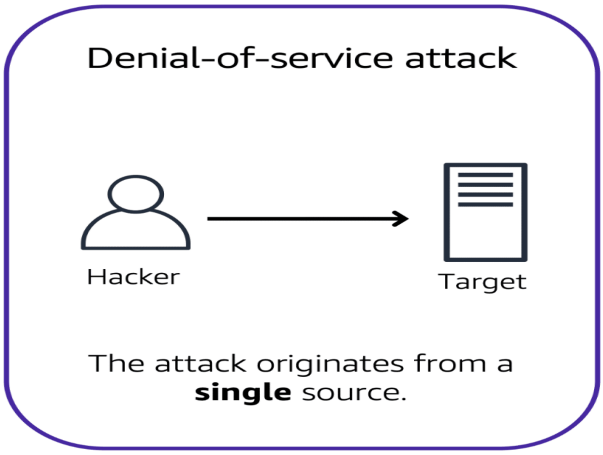
You can also access compliance white papers and documentation on topics such as:

- AWS answers to key compliance questions
- An overview of AWS risk and compliance
- An auditing security checklist

Additionally, the Customer Compliance Center includes an auditor learning path. This learning path is designed for individuals in auditing, compliance, and legal roles who want to learn more about how their internal operations can demonstrate compliance using the AWS Cloud.

Denial of Service Attack

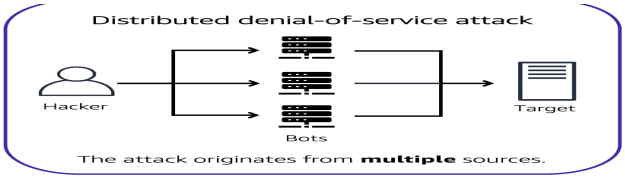
A denial-of-service (DoS) attack is a deliberate attempt to make a website or application unavailable to users.



For example, an attacker might flood a website or application with excessive network traffic until the targeted website or application becomes overloaded and is no longer able to respond. If the website or application becomes unavailable, this denies service to users who are trying to make legitimate requests.

Distributed denial-of-service attacks

In a distributed denial-of-service (DDoS) attack, multiple sources are used to start an attack that aims to make a website or application unavailable. This can come from a group of attackers, or even a single attacker. The single attacker can use multiple infected computers (also known as “bots”) to send excessive traffic to a website or application.



AWS Shield

AWS Shield is a service that protects applications against DDoS attacks. AWS Shield provides two levels of protection: Standard and Advanced.

AWS Shield Standard:AWS Shield Standard automatically protects all AWS customers at no cost. It protects your AWS resources from the most common, frequently occurring types of DDoS attacks. As network traffic comes into your applications, AWS Shield Standard uses a variety of analysis techniques to detect malicious traffic in real time and automatically mitigates it.

AWS Shield Advanced:AWS Shield Advanced is a paid service that provides detailed attack diagnostics and the ability to detect and mitigate sophisticated DDoS attacks. It also integrates with other services such as Amazon CloudFront, Amazon Route 53, and Elastic Load Balancing. Additionally, you can integrate AWS Shield with **AWS WAF(web application firewall)** by writing custom rules to mitigate complex DDoS attacks.

ADDITIONAL SECURITY SERVICES

AWS Key Management Services(AWS KMS)

you must ensure that your applications' data is secure while in storage (encryption at rest) and while it is transmitted, known as **encryption in transi**t.

AWS Key Management Service (AWS KMS)enables you to perform encryption operations through the use of c**ryptographic keys**. A cryptographic key is a random string of digits used for locking (encrypting) and unlocking (decrypting) data. You can use AWS KMS to create, manage, and use cryptographic keys. You can also control the use of keys across a wide range of services and in your applications.

With AWS KMS, you can choose the specific levels of access control that you need for your keys. For example, you can specify which IAM users and roles are able to manage keys. Alternatively, you can temporarily disable keys so that they are no longer in use by anyone. Your keys never leave AWS KMS, and you are always in control of them.

AWS Web Application Firewall(AWS WAF)

AWS WAF is a web application firewall that lets you monitor network requests that come into your web applications.

AWS WAF works together with Amazon CloudFront and an Application Load Balancer. Recall the network access control lists that you learned about in an earlier module. AWS WAF works in a similar way to block or allow traffic. However, it does this by using a web access control list (ACL) to protect your AWS resources.

AWS Inspector

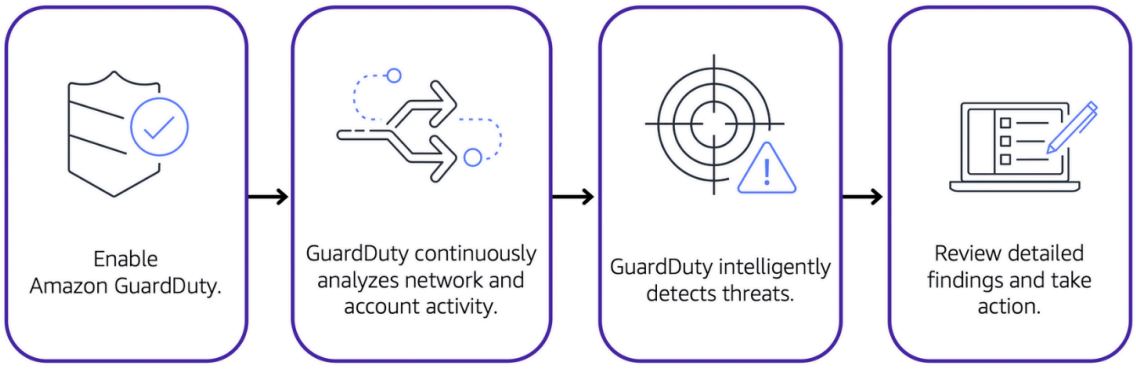
To perform automated security assessments, we can use Amazon Inspector.

Amazon Inspector helps to improve the security and compliance of applications by running automated security assessments. It checks applications for security vulnerabilities and deviations from security best practices, such as open access to Amazon EC2 instances and installations of vulnerable software versions.

After Amazon Inspector has performed an assessment, it provides you with a list of security findings. The list prioritises by severity level, including a detailed description of each security issue and a recommendation for how to fix it

Amazon GuardDuty

Amazon GuardDuty is a service that provides intelligent threat detection for your AWS infrastructure and resources. It identifies threats by continuously monitoring the network activity and account behaviour within your AWS environment.
GuardDuty then continuously analyses data from multiple AWS sources, including VPC Flow Logs and DNS logs.
If GuardDuty detects any threats, you can review detailed findings about them from the AWS Management Console. Findings include recommended steps for remediation. You can also configure AWS Lambda functions to take remediation steps automatically in response to Guard Duties security findings.



MONITORING AND ANALYTICS

Amazon CloudWatch

Amazon CloudWatch is a web service that enables you to monitor and manage various metrics and configure alarm actions based on data from those metrics.
CloudWatch uses **metrics** to represent the data points for your resources. AWS services send metrics to CloudWatch. CloudWatch then uses these metrics to create graphs automatically that show how performance has changed over time.

- Access all your metrics from a central location.
- Gain visibility into your applications, infrastructure and services.
- you can reduce mean time to resolution, or MTTR, and improve total cost of ownership, or TCO.
- Drive insights to optimise applications and operational resources.

CloudWatch Alarms

create alarms that automatically perform actions if the value of your metric has gone above or below a predefined threshold.
CloudWatch alarm that automatically stops an Amazon EC2 instance when the CPU utilisation percentage has remained below a certain threshold for a specified period. When configuring the alarm, you can specify to receive a notification whenever this alarm is triggered

CloudWatch Dashboard

The CloudWatch dashboard feature enables you to access all the metrics for your resources from a single location. For example, you can use a CloudWatch dashboard to monitor the CPU utilisation of an Amazon EC2 instance, the total number of requests made to an Amazon S3 bucket, and more. You can even customise separate dashboards for different business purposes, applications, or resources.

AWS CloudTrail

AWS CloudTrail records API calls for your account. The recorded information includes the identity of the API caller, the time of the API call, the source IP address of the API caller, and more. You can think of CloudTrail as a “trail” of breadcrumbs (or a log of actions) that someone has left behind them.
Recall that you can use API calls to provision, manage, and configure your AWS resources. With CloudTrail, you can view a complete history of user activity and API calls for your applications and resources.

Events are typically updated in CloudTrail within 15 minutes after an API call. You can filter events by specifying the time and date that an API call occurred, the user who requested the action, the type of resource that was involved in the API call, and more.

CloudTrail Insights
Within CloudTrail, you can also enable CloudTrail Insights. This optional feature allows CloudTrail to automatically detect unusual API activities in your AWS account.

AWS Trusted Advisor

Five pillars

- Cost optimisation
- Performance
- Security
- Fault tolerance
- Service limits

AWS Trusted Advisor is a web service that inspects your AWS environment and provides real-time recommendations in accordance with AWS best practices.
Trusted Advisor compares its findings to AWS best practices in five categories: cost optimization, performance, security, fault tolerance, and service limits. For the checks in each category, Trusted Advisor offers a list of recommended actions and additional resources to learn more about AWS best practices.

AWS Trusted Advisor dashboard

- **The green check indicates the number of items for which it detected no problems.**
- **The orange triangle represents the number of recommended investigations.**
- **The red circle represents the number of recommended actions.**

PRICING AND SUPPORT

AWS Free Tier

The AWS Free Tier(opens in a new tab) enables you to begin using certain services without having to worry about incurring costs for the specified period.
Three types of offers are available:

- **Always Free**:- These offers do not expire and are available to all AWS customers.

For example, **AWS Lambda** allows 1 million free requests and up to 3.2 million seconds of compute time per month. Amazon DynamoDB allows 25 GB of free storage per month.

- **12 Months Free:-** These offers are free for 12 months following your initial sign-up date to AWS.
  Examples include specific amounts of **Amazon S3 Standard Storage**, thresholds for monthly hours of Amazon EC2 compute time, and amounts of Amazon CloudFront data transfer out.
- **Trials :-** Short-term free trial offers start from the date you activate a particular service. The length of each trial might vary by number of days or the amount of usage in the service.
  For example, **Amazon Inspector** offers a 90-day free trial. **Amazon Lightsail** (a service that enables you to run virtual private servers) offers 750 free hours of usage over a 30-day period.

AWS Pricing Concepts

AWS offers a range of cloud computing services with pay-as-you-go pricing.

- **Pay for what you use;-**For each service, you pay for exactly the amount of resources that you actually use, without requiring long-term contracts or complex licensing.
- **Pay less when you reserve;-** Some services offer reservation options that provide a significant discount compared to On-Demand Instance pricing.
  For example, suppose that your company is using Amazon EC2 instances for a workload that needs to run continuously. You might choose to run this workload on Amazon EC2 Instance Savings Plans, because the plan allows you to save up to 72% over the equivalent On-Demand Instance capacity.
- **Pay less with volume based discounts when you use more;-** Some services offer tiered pricing, so the per-unit cost is incrementally lower with increased usage.
  For example, the more Amazon S3 storage space you use, the less you pay for it per GB.

AWS Pricing Calculator

The AWS Pricing Calculator lets you explore AWS services and create an estimate for the cost of your use cases on AWS. You can organise your AWS estimates by groups that you define. A group can reflect how your company is organised, such as providing estimates by cost centre.

Suppose that your company is interested in using Amazon EC2. However, you are not yet sure which AWS Region or instance type would be the most cost-efficient for your use case. In the AWS Pricing Calculator, you can enter details, such as the kind of operating system you need, memory requirements, and input/output (I/O) requirements. By using the AWS Pricing Calculator, you can review an estimated comparison of different EC2 instance types across AWS Regions.

AWS LAMBDA PRICING

- For AWS Lambda, you are charged based on the number of requests for your functions and the time that it takes for them to run.
- AWS Lambda allows 1 million free requests and up to 3.2 million seconds of compute time per month.
- You can save on AWS Lambda costs by signing up for a Compute Savings Plan. A Compute Savings Plan offers lower compute costs in exchange for committing to a consistent amount of usage over a 1-year or 3-year term. This is an example of paying less when you reserve.

AMAZON EC2 PRICING

- With Amazon EC2, **you pay for only the compute time** that you use while your instances are running.
- For some workloads, you can significantly reduce Amazon EC2 costs by using Spot Instances. For example, suppose that you are running a batch processing job that is able to withstand interruptions. Using a Spot Instance would provide you with up to 90% cost savings while still meeting the availability requirements of your workload.
- You can find additional cost savings for Amazon EC2 by considering Savings Plans and Reserved Instances.

AMAZON S3

- **Storage** - You pay for only the storage that you use. You are charged the rate to store objects in your Amazon S3 buckets based on your objects' sizes, storage classes, and how long you have stored each object during the month.
- **Requests and data retrievals** - You pay for requests made to your Amazon S3 objects and buckets. For example, suppose that you are storing photo files in Amazon S3 buckets and hosting them on a website. Every time a visitor requests the website that includes these photo files, this counts towards requests you must pay for.
- **Data transfer** - There is no cost to transfer data between different Amazon S3 buckets or from Amazon S3 to other services within the same AWS Region. However, you pay for data that you transfer into and out of Amazon S3, with a

few exceptions. There is no cost for data transferred into Amazon S3 from the internet or out to Amazon CloudFront. There is also no cost for data transferred out to an Amazon EC2 instance in the same AWS Region as the Amazon S3 bucket.

- **Management and replication** - You pay for the storage management features that you have enabled on your account's Amazon S3 buckets. These features include Amazon S3 inventory, analytics, and object tagging.

BILLING DASHBOARD

Use the AWS Billing & Cost Management dashboard to pay your AWS bill, monitor your usage, and analyse and control your costs.

- Compare your current month-to-date balance with the previous month, and get a forecast of the next month based on current usage.
- View month-to-date spend by service.
- View Free Tier usage by service.
- Access Cost Explorer and create budgets.
- Purchase and manage Savings Plans.
- Publish AWS Cost and Usage Reports.

CONSOLIDATED BILLING

AWS Organizations, a service that enables you to manage multiple AWS accounts from a central location. **AWS Organizations** also provides the option for consolidated billing. The consolidated billing feature of AWS Organizations enables you to receive a single bill for all AWS accounts in your organisation. By consolidating, you can easily track the combined costs of all the linked accounts in your organisation. The default maximum number of accounts allowed for an organisation is 4, but you can contact AWS Support to increase your quota, if needed.
Another benefit of consolidated billing is the ability to share bulk discount pricing, Savings Plans, and Reserved Instances across the accounts in your organization

AWS BUDGETS

In AWS Budgets, you can create budgets to plan your service usage, service costs, and instance reservations.
The information in AWS Budgets updates three times a day. This helps you to accurately determine how close your usage is to your budgeted amounts or to the AWS Free Tier limits.
In AWS Budgets, you can also set custom alerts when your usage exceeds (or is forecasted to exceed) the budgeted amount.

AWS COST EXPLORER

AWS Cost Explorer is a tool that lets you visualise, understand, and manage your AWS costs and usage over time.
AWS Cost Explorer includes a default report of the costs and usage for your top five cost-accruing AWS services. You can apply custom filters and groups to analyse your data. For example, you can view resource usage at the hourly level.
By analysing your AWS costs over time, you can make informed decisions about future costs and how to plan your budgets.

AWS SUPPORT PLANS

AWS offers four different Support plans(opens in a new tab) to help you troubleshoot issues, lower costs, and efficiently use AWS services.
You can choose from the following Support plans to meet your company's needs:

- Basic
- Developer
- Business
- Enterprise On-Ramp
- Enterprise

BASIC SUPPORT

Basic Support is free for all AWS customers. It includes access to whitepapers, documentation, and support communities. With Basic Support, you can also contact AWS for billing questions and service limit increases.
With Basic Support, you have access to a limited selection of AWS Trusted Advisor checks. Additionally, you can use the **AWS Personal Health Dashboard**, a tool that provides alerts and remediation guidance when AWS is experiencing events that may affect you.
If your company needs support beyond the Basic level, you could consider purchasing Developer, Business, Enterprise On-Ramp, and Enterprise Support.

DEVELOPER SUPPORT

Customers in the Developer Support plan have access to features such as:

- Best practice guidance
- Client-side diagnostic tools
- Building-block architecture support, which consists of guidance for how to use AWS offerings, features, and services together

For example, suppose that your company is exploring AWS services. You've heard about a few different AWS services. However, you're unsure of how to potentially use them together to build applications that can address your company's needs. In this scenario, the building-block architecture support that is included with the Developer

Support plan could help you to identify opportunities for combining specific services and features.

BUSINESS SUPPORT

Customers with a Business Support plan have access to additional features, including:

- Use-case guidance to identify AWS offerings, features, and services that can best support your specific needs
- **All AWS Trusted Advisor checks**
- Limited support for third-party software, such as common operating systems and application stack components

Suppose that your company has the Business Support plan and wants to install a common third-party operating system onto your Amazon EC2 instances. You could contact AWS Support for assistance with installing, configuring, and troubleshooting the operating system. For advanced topics such as optimising performance, using custom scripts, or resolving security issues, you may need to contact the third-party software provider directly.

ENTERPRISE ON-RAMP SUPPORT

Enterprise On-Ramp Support plan have access to:

- A pool of Technical Account Managers to provide proactive guidance and coordinate access to programs and AWS experts
- A Cost Optimization workshop (one per year)
- A Concierge support team for billing and account assistance
- Tools to monitor costs and performance through Trusted Advisor and Health API/Dashboard

Enterprise On-Ramp Support plan also provides access to a specific set of proactive support services, which are provided by a pool of Technical Account Managers.

- Consultative review and architecture guidance (one per year)
- Infrastructure Event Management support (one per year)
- Support automation workflows
- 30 minutes or less response time for business-critical issues

ENTERPRISE SUPPORT

In addition to all features included in the Basic, Developer, Business, and Enterprise On-Ramp support plans, customers with Enterprise Support have access to:

- A designated Technical Account Manager to provide proactive guidance and coordinate access to programs and AWS experts
- A Concierge support team for billing and account assistance
- Operations Reviews and tools to monitor health

- Training and Game Days to drive innovation
- Tools to monitor costs and performance through Trusted Advisor and Health API/Dashboard

The Enterprise plan also provides full access to proactive services, which are provided by a designated Technical Account Manager:

- Consultative review and architecture guidance
- Infrastructure Event Management support
- Cost Optimization Workshop and tools
- Support automation workflows
- 15 minutes or less response time for business-critical issues

TECHNICAL ACCOUNT MANAGER

The Enterprise On-Ramp and Enterprise Support plans include access to a Technical Account Manager (TAM).

The TAM is your primary point of contact at AWS. If your company subscribes to Enterprise Support or Enterprise On-Ramp, your TAM educates, empowers, and evolves your cloud journey across the full range of AWS services. TAMs provide expert engineering guidance, help you design solutions that efficiently integrate AWS services, assist with cost-effective and resilient architectures, and provide direct access to AWS programs and a broad community of experts.

AWS MARKETPLACE

AWS Marketplace is a digital catalogue that includes thousands of software listings from independent software vendors. You can use AWS Marketplace to find, test, and buy software that runs on AWS.

AWS Marketplace offers products in several categories, such as Infrastructure Software, DevOps, Data Products, Professional Services, Business Applications, Machine Learning, Industries, and Internet of Things (IoT).

Within each category, you can narrow your search by browsing through product listings in subcategories. For example, subcategories in the DevOps category include areas such as Application Development, Monitoring, and Testing.

MIGRATION AND INNOVATION

AWS CLOUD ADOPTION FRAMEWORK(AWS CAF)

At the highest level, the AWS Cloud Adoption Framework (AWS CAF) organises guidance into six areas of focus, called Perspectives. Each Perspective addresses distinct responsibilities. The planning process helps the right people across the organisation prepare for the changes ahead.

In general, the Business, People, and Governance Perspectives focus on business capabilities, whereas the Platform, Security, and Operations Perspectives focus on technical capabilities.

BUSINESS PERSPECTIVE

The Business Perspective ensures that IT aligns with business needs and that IT investments link to key business results.

Use the Business Perspective to create a **strong business case for cloud adoption and prioritise cloud adoption initiatives**. Ensure that your business strategies and goals align with your IT strategies and goals.

Common roles in the Business Perspective include:

- Business managers
- Finance managers
- Budget owners
- Strategy stakeholders

PEOPLE PERSPECTIVE

The People Perspective **supports development of an organisation-wide change management strategy for successful cloud adoption.**

Use the People Perspective to evaluate organisational structures and roles, new skill and process requirements, and identify gaps. This helps prioritise training, staffing, and organisational changes.

Common roles in the People Perspective include:

- Human resources
- Staffing
- People managers

GOVERNANCE PERSPECTIVE

The Governance Perspective focuses on the skills and processes to align IT strategy with business strategy. This ensures that you **maximise the business value and minimise risks.**

Use the Governance Perspective to understand how to update the staff skills and processes necessary to ensure business governance in the cloud. Manage and measure cloud investments to evaluate business outcomes.

Common roles in the Governance Perspective include:

- Chief Information Officer (CIO)
- Program managers
- Enterprise architects
- Business analysts
- Portfolio managers

PLATFORM PERSPECTIVE

The Platform Perspective includes principles and patterns for i**mplementing new solutions on the cloud, and migrating on-premises workloads to the cloud.**

Use a variety of architectural models to understand and communicate the structure of IT systems and their relationships. Describe the architecture of the target state environment in detail.

Helps you design,implement, and optimise aws infrastructure.

Common roles in the Platform Perspective include:

- Chief Technology Officer (CTO)
- IT managers
- Solutions architects

SECURITY PERSPECTIVE

The Security Perspective ensures that the organisation **meets security objectives** for visibility, auditability, control, and agility.

Use the AWS CAF to structure the selection and implementation of security controls that meet the organisation's needs.

Common roles in the Security Perspective include:

- Chief Information Security Officer (CISO)
- IT security managers
- IT security analysts

OPERATIONS PERSPECTIVE

The Operations Perspective h**elps you to enable, run, use, operate, and recover IT workloads to the level agreed upon with your business stakeholders.**

Define how day-to-day, quarter-to-quarter, and year-to-year business is conducted. Align with and support the operations of the business. The AWS CAF helps these stakeholders define current operating procedures and identify the process changes and training needed to implement successful cloud adoption.

Common roles in the Operations Perspective include:

- IT operations managers
- IT support managers

MIGRATION STRATEGIES

six of the most common migration strategies that you can implement are:

- Rehosting

- Replatforming
- Refactoring/re-architecting
- Repurchasing
- Retaining
- Retiring

Rehosting
Rehosting also known as “lift-and-shift” involves moving applications without changes. In the scenario of a large legacy migration, in which the company is looking to implement its migration and scale quickly to meet a business case, the majority of applications are rehosted.

Replatforming
Replatforming, also known as “lift, tinker, and shift,” involves making a few cloud optimizations to realise a tangible benefit. Optimization is achieved without changing the core architecture of the application.

Refactoring or Re-Architecting
Refactoring (also known as re-architecting) involves reimagining how an application is architected and developed by using cloud-native features. Refactoring is driven by a strong business need to add features, scale, or performance that would otherwise be difficult to achieve in the application’s existing environment.

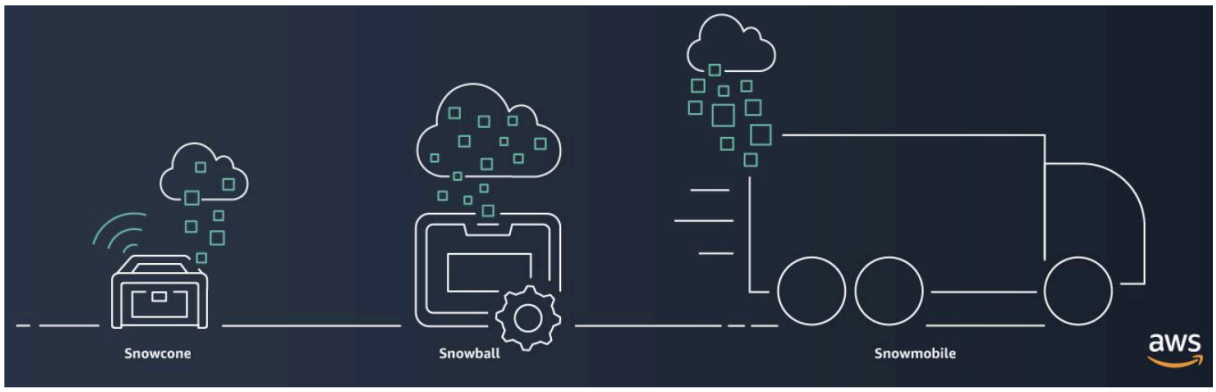
Repurchasing
Repurchasing involves moving from a traditional licence to a software-as-a-service model.This migration strategy involves moving to a different product.
For example, a business might choose to implement the repurchasing strategy by migrating from a customer relationship management (CRM) system to Salesforce.com.

Retaining
Retaining consists of keeping applications that are critical for the business in the source environment. This might include applications that require major refactoring before they can be migrated, or, work that can be postponed until a later time.

Retiring
Retiring is the process of removing applications that are no longer needed.

AWS SNOW FAMILY

The AWS Snow Family is a collection of physical devices that help to physically transport up to exabytes of data into and out of AWS.
AWS Snow Family is composed of AWS Snowcone, AWS Snowball, and AWS Snowmobile.
These devices offer different capacity points, and most include built-in computing capabilities. AWS owns and manages the Snow Family devices and integrates with AWS security, monitoring, storage management, and computing capabilities.



AWS Snowcone

AWS Snowcone is a small, rugged, and secure edge computing and data transfer device. It features 2 CPUs, 4 GB of memory, and up to 14 TB of usable storage.

AWS Snowball

AWS Snowball offers two types of devices:
Snowball Edge Storage Optimised devices are well suited for large-scale data migrations and recurring transfer workflows, in addition to local computing with higher capacity needs.

- Storage: 80 TB of hard disk drive (HDD) capacity for block volumes and Amazon S3 compatible object storage, and 1 TB of SATA solid state drive (SSD) for block volumes.
- Compute: 40 vCPUs, and 80 GiB of memory to support Amazon EC2 sbe1 instances (equivalent to C5).

Snowball Edge Compute Optimised provides powerful computing resources for use cases such as machine learning, full motion video analysis, analytics, and local computing stacks.

- Storage: 80-TB usable HDD capacity for Amazon S3 compatible object storage or Amazon EBS compatible block volumes and 28 TB of usable NVMe SSD capacity for Amazon EBS compatible block volumes.

- Compute: 104 vCPUs, 416 GiB of memory, and an optional NVIDIA Tesla V100 GPU. Devices run Amazon EC2 sbe-c and sbe-g instances, which are equivalent to C5, M5a, G3, and P3 instances.

INNOVATION WITH AWS

When examining how to use AWS services, it is important to focus on the desired outcomes. You are properly equipped to drive innovation in the cloud if you can clearly articulate the following conditions:

- The current state
- The desired state
- The problems you are trying to solve

Serverless Applications

With AWS, serverless refers to applications that don't require you to provision, maintain, or administer servers. You don't need to worry about fault tolerance or availability. AWS handles these capabilities for you.

AWS Lambda is an example of a service that you can use to run serverless applications. If you design your architecture to trigger Lambda functions to run your code, you can bypass the need to manage a fleet of servers.

Building your architecture with serverless applications enables your developers to focus on their core product instead of managing and operating servers.

Artificial Intelligence

AWS offers a variety of services powered by artificial intelligence (AI).

For example, you can perform the following tasks:

- Convert speech to text with **Amazon Transcribe**.
- Discover patterns in text with **Amazon Comprehend**.
- Identify potentially fraudulent online activities with **Amazon Fraud Detector.**
- Build voice and text chatbots with **Amazon Lex**.

Machine Learning

Traditional machine learning (ML) development is complex, expensive, time consuming, and error prone. AWS offers Amazon SageMaker to remove the difficult work from the process and empower you to build, train, and deploy ML models quickly.

You can use ML to analyse data, solve complex problems, and predict outcomes before they happen.

Amazon SageMaker: Quickly build, train, and deploy machine learning models at scale.

Tools like **Amazon SageMaker and Amazon Augmented AI, or Amazon A2I**, provide a machine learning platform that any business can build upon without needing PhD level expertise in-house. Or perhaps, ready-to-go AI solutions like **Amazon Lex**, the heart of Alexa.

Amazon Textract. Extracting text and data from documents to make them more usable for your enterprise instead of them just being locked away in a repository

AWS DeepRacer is an autonomous 1/18 scale race car that you can use to test reinforcement learning models.

CLOUD JOURNEY

AWS WELL-ARCHITECTED FRAMEWORK

The AWS Well-Architected Framework helps you understand how to design and operate reliable, secure, efficient, and cost-effective systems in the AWS Cloud. It provides a way for you to consistently measure your architecture against best practices and design principles and identify areas for improvement.

A well architected framework has 6 pillars:

- Operational Excellence
- Security
- Reliability
- Performance efficiency
- Cost optimisation
- sustainability

Operational Excellence:
Operational excellence is the ability to run and monitor systems to deliver business value and to continually improve supporting processes and procedures.
Design principles for operational excellence in the cloud include performing operations as code, annotating documentation, anticipating failure, and frequently making small, reversible changes.

Security:
The Security pillar is the ability to protect information, systems, and assets while delivering business value through risk assessments and mitigation strategies.
When considering the security of your architecture, apply these best practices:

- Automate security best practices when possible.
- Apply security at all layers.
- Protect data in transit and at rest.

Reliability:
Reliability is the ability of a system to do the following:
- Recover from infrastructure or service disruptions
- Dynamically acquire computing resources to meet demand
- Mitigate disruptions such as misconfigurations or transient network issues

Reliability includes testing recovery procedures, scaling horizontally to increase aggregate system availability, and automatically recovering from failure.

Performance Efficiency:
Performance efficiency is the ability to use computing resources efficiently to meet system requirements and to maintain that efficiency as demand changes and technologies evolve.
Evaluating the performance efficiency of your architecture includes experimenting more often, using serverless architectures, and designing systems to be able to go global in minutes.

Cost Optimisation:
Cost optimization is the ability to run systems to deliver business value at the lowest price point.
Cost optimization includes adopting a consumption model, analysing and attributing expenditure, and using managed services to reduce the cost of ownership.

Sustainability:
Sustainability is the ability to continually improve sustainability impacts by reducing energy consumption and increasing efficiency across all components of a workload by maximising the benefits from the provisioned resources and minimising the total resources required.
To facilitate good design for sustainability:
- Understand your impact
- Establish sustainability goals
- Maximise utilisation
- Anticipate and adopt new, more efficient hardware and software offerings
- Use managed services
- Reduce the downstream impact of your cloud workloads

Benefits of AWS Cloud

Operating in the AWS Cloud offers many benefits over computing in on-premises or hybrid environments. The main 6 advantages are:

- Trade upfront expense for variable expense.
- Benefits from massive economies of scale.
- Stop guessing capacity.
- Increase speed and agility.
- Stop spending money running and maintaining data centres.
- Go global in minutes.

Trade upfront expense for variable expense:

Upfront expenses include data centres, physical servers, and other resources that you would need to invest in before using computing resources.
Instead of investing heavily in data centres and servers before you know how you're going to use them, you can pay only when you consume computing resources.

Benefits from massive economies of scale:

By using cloud computing, you can achieve a lower variable cost than you can get on your own.
Because usage from hundreds of thousands of customers aggregates in the cloud, providers such as AWS can achieve higher economies of scale. Economies of scale translate into lower pay-as-you-go prices.

Stop Guessing Capacity:

With cloud computing, you don't have to predict how much infrastructure capacity you will need before deploying an application.
For example, you can launch Amazon Elastic Compute Cloud (Amazon EC2) instances when needed and pay only for the compute time you use. Instead of paying for resources that are unused or dealing with limited capacity, you can access only the capacity that you need, and scale in or out in response to demand.

Increase speed and agility:

The flexibility of cloud computing makes it easier for you to develop and deploy applications.
This flexibility also provides your development teams with more time to experiment and innovate.

Stop spending money on running and maintaining data centres:

Cloud computing in data centres often requires you to spend more money and time managing infrastructure and servers.

A benefit of cloud computing is the ability to focus less on these tasks and more on your applications and customers.

Go global in minutes:
The AWS Cloud global footprint enables you to quickly deploy applications to customers around the world, while providing them with low latency.