AWS CLF-C02 Notes

Cloud Concepts

Cloud Computing

The practice of using a network of remote servers hosted on the internet to store & manage, & process data rather than a local server or a personal computer.

Evolution of Cloud Hosting

1. Dedicated Server
    - One machine for a single business.
    - *Very Expensive, High Maintenance, High Security.*

2. Virtual Private Server
    - One physical machine divided into submachines for single business.
    - *Better Utilisation & Isolation of Resources.*

3. Shared Hosting
    - One physical machine, shared by hundreds of businesses.
    - *Very Cheap, Limited Functionality, Poor Isolation.*

4. Cloud Hosting
    - Multiple physical machines acting as a single system.
    - *Flexible, Scalable, Secure, Cost-Effective, High Configurability.*

What is Amazon

- American multinational computer technology corporation.
- Headquarters in Seattle, Washington.
- Founded by Jeff Bezos in 1994.

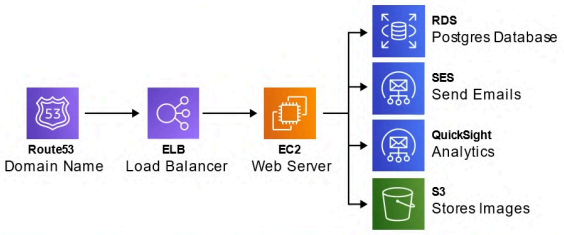
What is AWS

- Amazon's Cloud Provider Service. Commonly referred to as AWS.
- Launched in 2006 & is the leading cloud service provider in the world.

What is Cloud Service Provider (CSP)

A **Cloud Service Provider (CSP)** is a company which

- provides multiple Cloud Services e.g. tens to hundreds of services
- those Cloud Services **can be chained together** to create cloud architectures
- those Cloud Services are accessible **via Single Unified API** eg. AWS API
- those Cloud Services utilized **metered billing** based on usage e.g. per second, per hour
- those Cloud Services have rich monitoring built in eg. AWS CloudTrail
- those Cloud Services have an Infrastructure as a Service (IaaS) offering
- Those Cloud Services offers **automation** via Infrastructure as Code (IaC)



If a company offers multiple cloud services under a single UI but do not meet most of or all of these requirements, it would be referred to as a Cloud Platform e.g. Twilio, HashiCorp, Databricks

If a company offers multiple cloud services under a single UI but does not meet most of or all of these requirements, it would be referred to as a Cloud Platform.

Landscape of CSP's

Tier 1 (Top Tier) : Early to market & well recognised in the industry

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)
- Alibaba Cloud

Tier 2 (Mid Tier) : Backed by well known tech companies

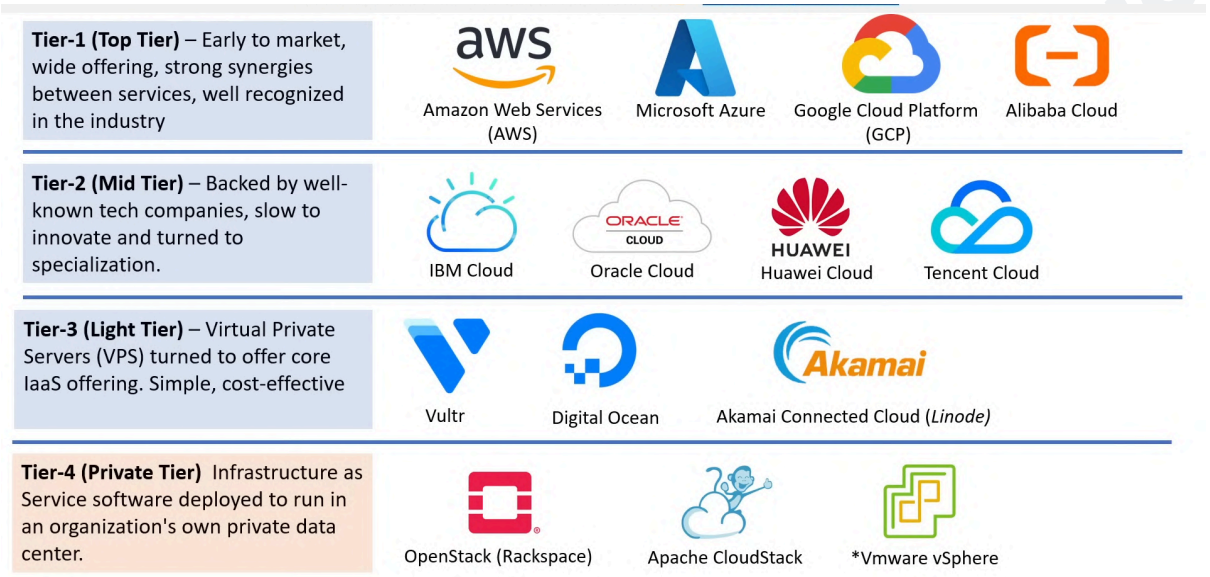
- IBM Cloud *(AI, Enterprise, managed infrastructure specialised services)*
- Oracle Cloud

Tier 3 (Light Tier) : VPS turned to core IaaS offering. Simple, Cost-effective

- Vultr
- Digital Ocean
- Linode

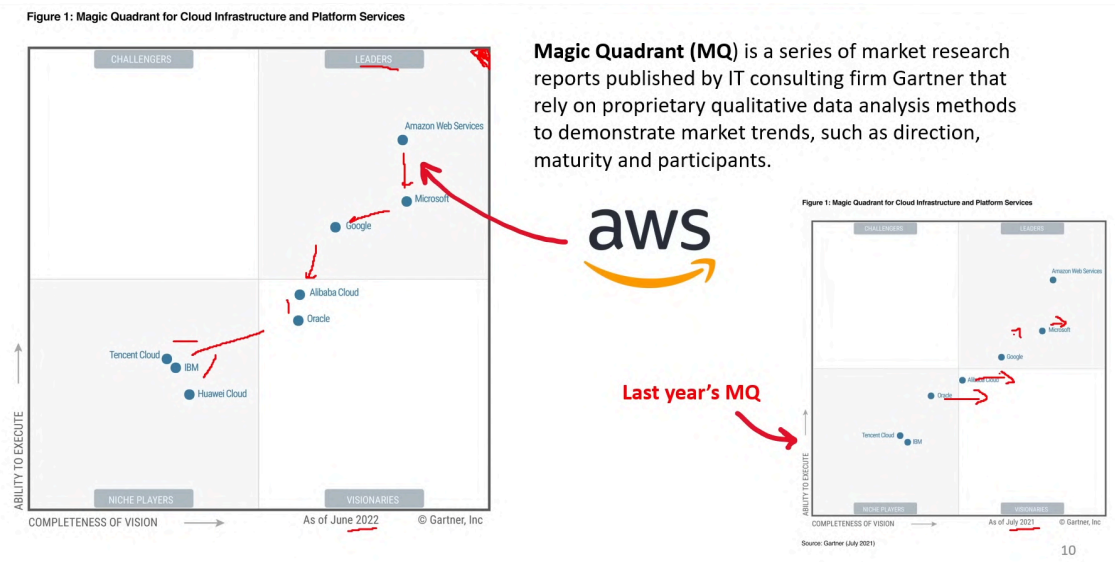
Tier 4 (Private Tier) : Infrastructure as Service Software deployed to run in an organisation's own private data centre.

- OpenStack(Rackspace)
- Apache CloudStack
- VMWare vSphere



Gartner Magic Quadrant for Cloud

Magic Quadrant (MQ) is a series of market research reports published by IT Consulting firm Gartner that rely on proprietary qualitative data analysis methods to demonstrate market trends, such as direction, maturity & participants.

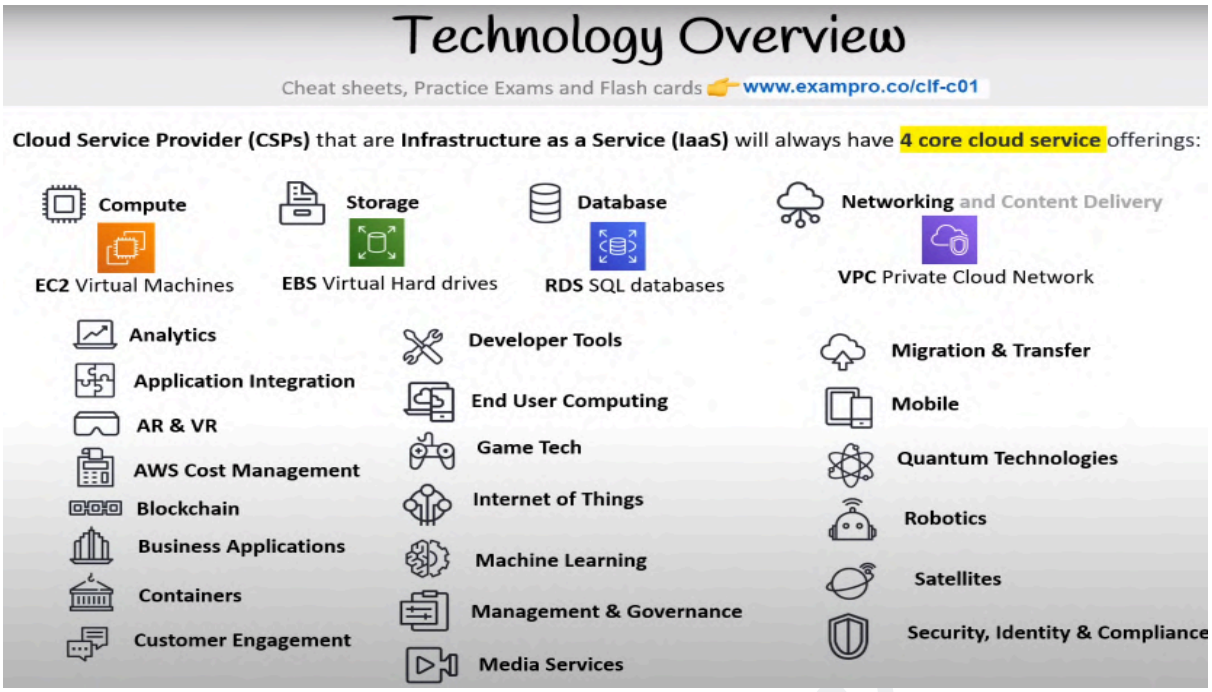


Common Cloud Services

A CSP can have hundreds of cloud services that are grouped into various types of services.

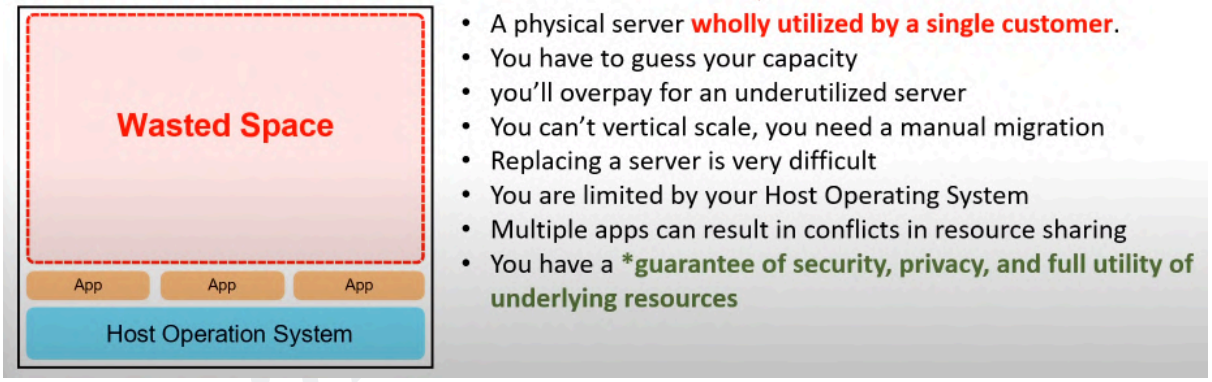
1. Compute :
   A virtual computer than can run applications, programs & code.
2. Storage :
   A virtual hard-drive that can store files.
3. Networking :
   A virtual network defining internet connection or network isolations between services or outbound to the internet.
4. Databases :
   A virtual database for storing reporting data or a database for general purpose web-application.

The term “Cloud Computing” can be used to refer to all categories.

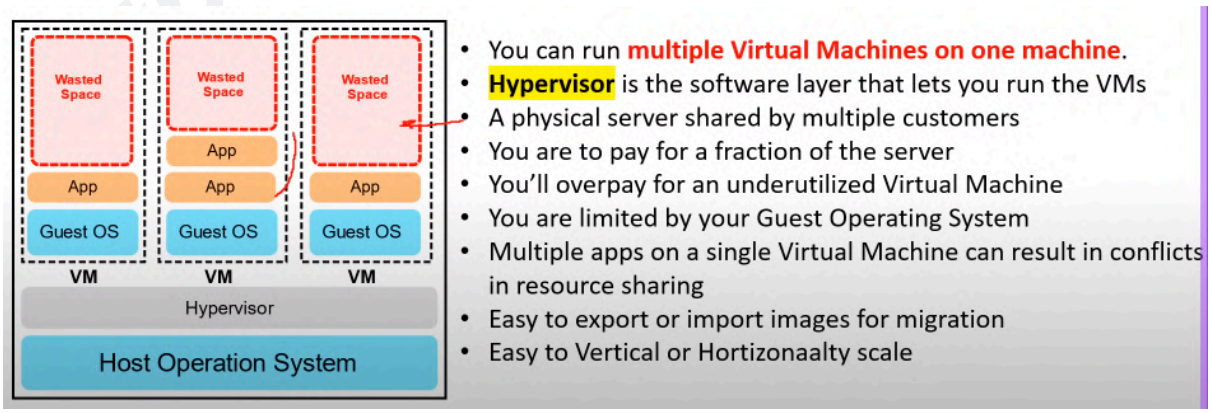


Evolution of Computing

Dedicated :

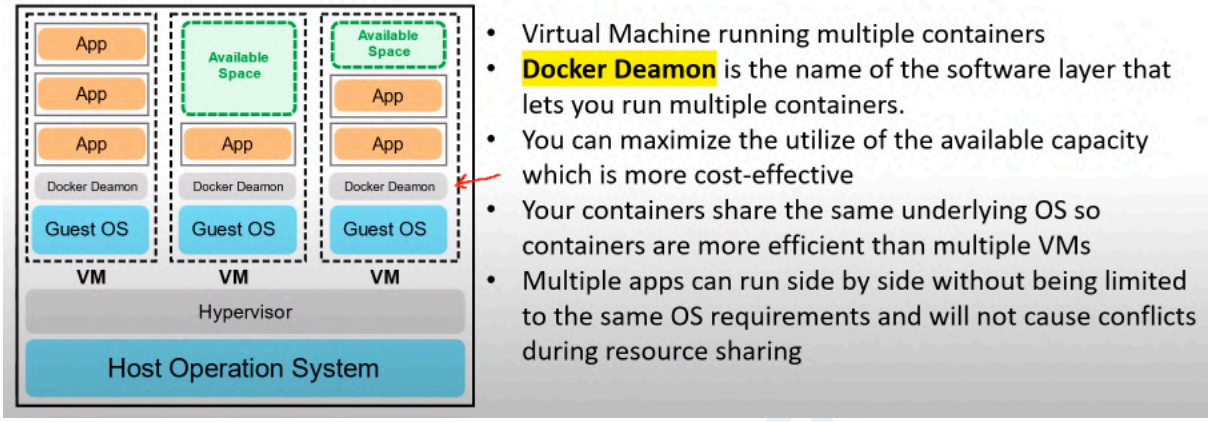


VMs :



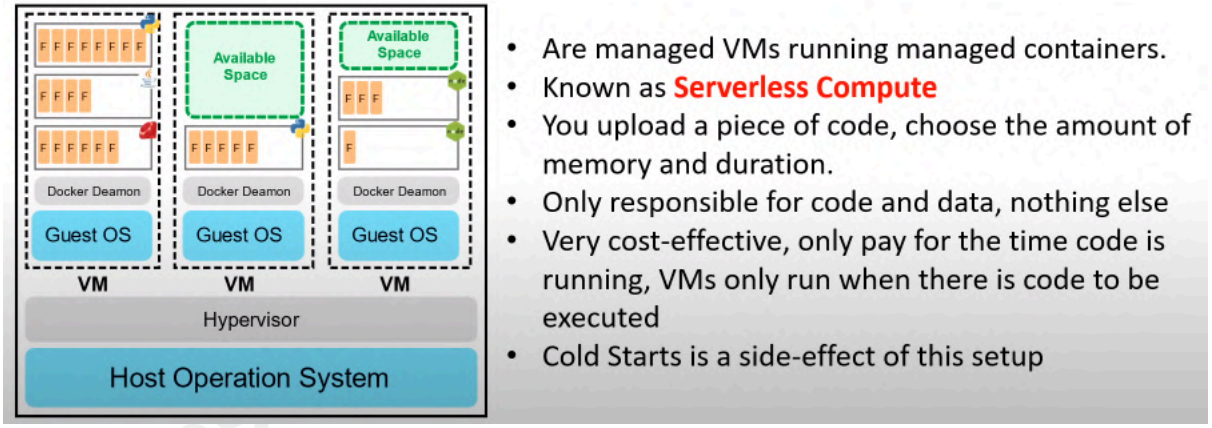
Hypervisor is the software layer that lets you run the Virtual Machines.

Containers :



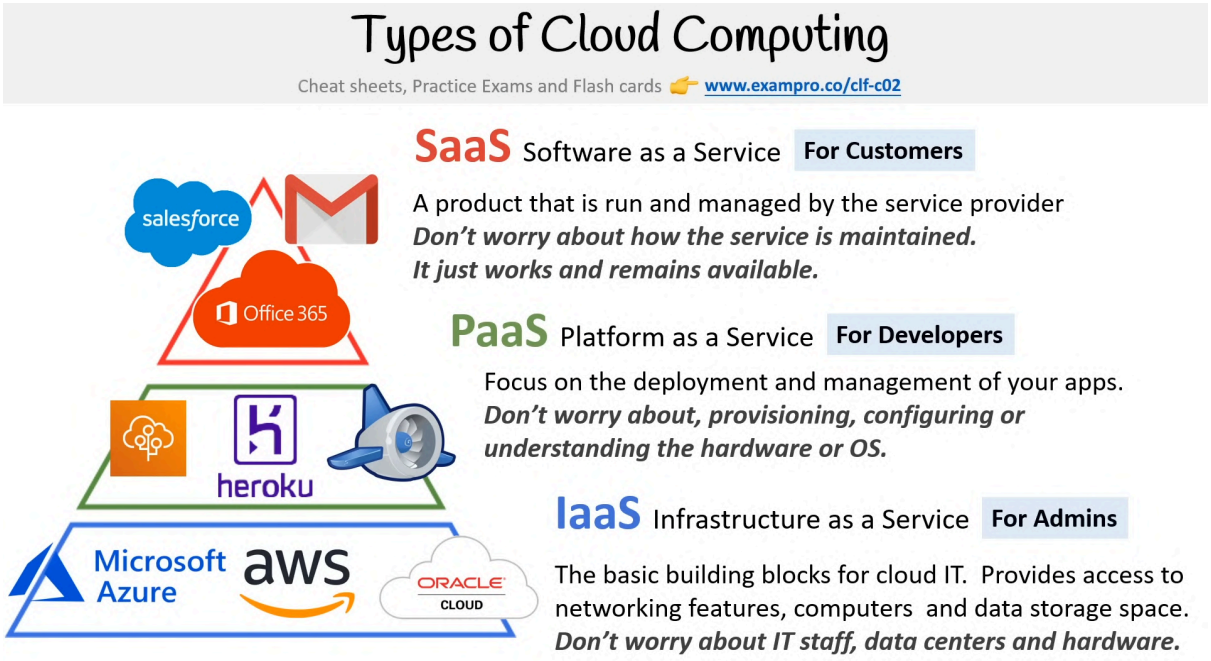
Docker Daemon is the software that lets you run multiple containers.

Functions :



Types of Cloud Computing

1. **SaaS :** Software as a Service. (For Customers)
2. **PaaS :** Platform as a Service. (For Developers)
3. **IaaS :** Infrastructure as a Service. (For Administrators)



Cloud Computing Deployment Model

1. Public Cloud :
   Everything is built on CSP. Also known as Cloud-Native or Cloud First.
2. Private Cloud :
   Everything built on the company's data centres. Also known a On-Premise.
3. Hybrid :
   Using both On-Premise & A Cloud Service Provider.
4. Cross Cloud :
   Using multiple Public Clouds.

Deployment Model Use Cases

Cloud Computing Deployment Models

Cheat sheets, Practice Exams and Flash cards 👉 www.exampro.co/clf-c01

Cloud

Fully utilizing cloud computing

Companies that are starting out today, or are small enough to make the leap from a VPS to a CSP.

- Startups
- SaaS offerings
- New projects and companies

Hybrid

Using both Cloud and On-Premise

Organizations that started with their own datacenter, can't fully move to cloud due to effort of migration or security compliance

- Banks
- FinTech, Investment Management
- Large Professional Service providers
- Legacy on-premise

On-Premise

Deploying resources on-premises, using virtualization and resource management tools, is sometimes called "private cloud".

Organizations that cannot run on cloud due to strict regulatory compliance or the sheer size of their organization

- Public Sector eg. Government
- Super Sensitive Data eg. Hospitals
- Large Enterprise with heavy regulation eg. Insurance Companies

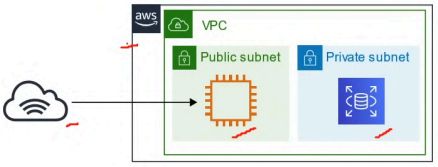
There really isn't reason to **be fully on-premise**

Cloud Computing Deployment Models

Cheat sheets, Practice Exams and Flash cards 👉 www.exampro.co/clf-c02

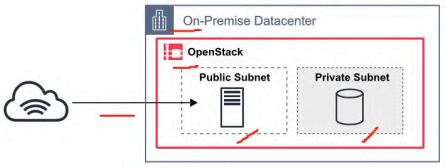
Public Cloud

Everything (the workload or project) is built on the CSP
Also known as: *Cloud-Native or Cloud First



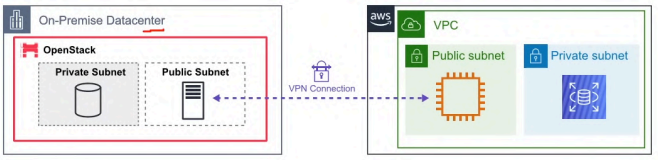
Private Cloud

Everything built on company's datacenters
Also known as **On-Premise**
The cloud could be **OpenStack**



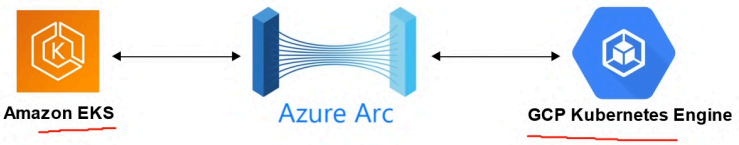
Hybrid

Using both **On-Premise** and
A **Cloud Service Provider**



Cross-Cloud

Using **Multiple Cloud Providers**
Aka multi-cloud, ~~"hybrid-cloud"~~



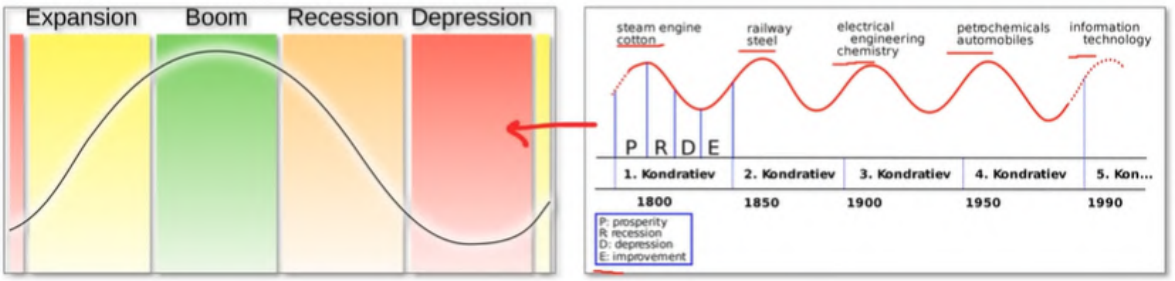
Anthos is GCP's offering for a control plane for compute across multiple CSPs and On-premise environments

Digital Transformation

Innovation Waves

Kondratiev Waves (Innovation Waves or K-Waves) are hypothesised cycle-like phenomena in the global world economy. The phenomenon is closely connected with technology life cycles.Each wave irreversibly changes society on a global scale.

Latest wave being Cloud Technology

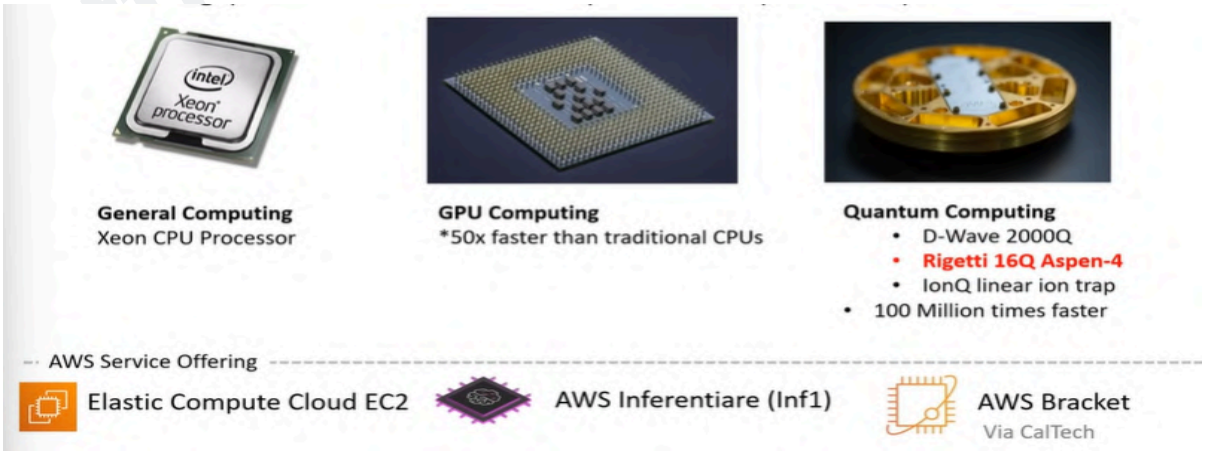


Burning Platform

It is a term used *when a company abandons old technology for new technology* with the uncertainty of success & can be motivated by the fear that the organisation's future survival hinges on its digital transformation.

Evolution of Computing Power

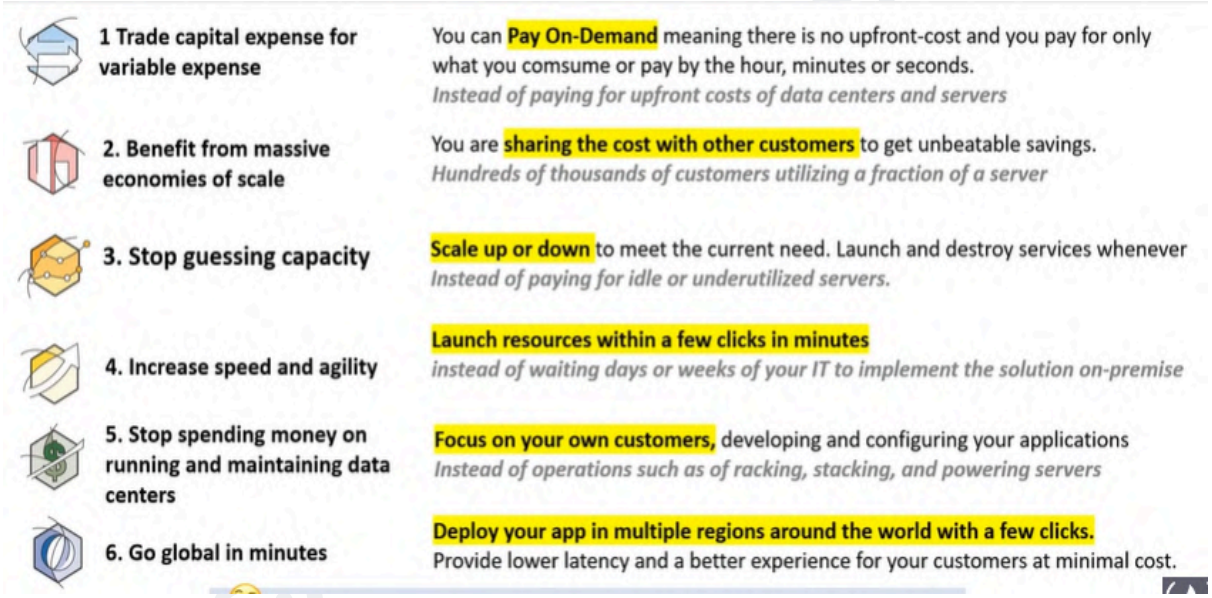
Computing Power is the throughput measured at which a computer can complete a computational task.



Benefits of the Cloud

Benefits of Cloud

- Agility : Increase speed & agility.
- Pay-as-you go pricing : Trade capital expense for variable expense.
- Economy of Scale : Benefit from massive economies of scale.
- Global Reach : Go global in minutes.
- Reliability : Stop spending money on running & maintaining data centres.
- Scalability : Benefit from massive economies of scale.
- Security.
- High Availability.
- Elasticity.
- Fault Tolerance.
- Disaster Recovery.



Six Benefits Version by Amazon

Cost-effective	You **pay for what you consume**, no up-front cost. On-demand pricing or Pay-as-you-go (PAYG) with thousands of customers sharing the cost of the resources
Global	Launch workloads **anywhere in the world**, Just choose a region
Secure	Cloud provider takes care of physical security. **Cloud services can be secure by default** or you have the ability to configure access down to a granular level.
Reliable	Data backup, disaster recovery, data replication, and fault tolerance
Scalable	Increase or decrease resources and services based on demand
Elastic	**Automate** scaling during spikes and drop in demand
Current	The underlying hardware and managed software is patched, upgraded and replaced by the cloud provider without interruption to you.

Seven Benefits Version by Tutor

AWS Global Infrastructure

AWS Global Infrastructure

The AWS Global Infrastructure is globally distributed hardware & datacenters that are physically networked together to act as one large resource for the end customer.
Millions of Active customers & tens of thousands of partners globally.

It comprises of these resources :

- 32 Launched Regions.
- 102 Availability Zones.
- 115 Direct Connection Locations.
- 550+ Points of Presence.
- 35 Local Zone.
- 29 Wavelength Zones.

Regions

- They are geographically distinct locations consisting of one or more Availability Zones.
- Every region is physically isolated from & independent of every other region in terms of location, power & water supply.
- US-EAST 1 (Northern Virginia) is the first AWS Region setup in 2006.
- All billing information appears in **US-EAST-1** (North Virginia)
- New Services are available first in US-EAST.
- Each region has generally three Availability Zones.
- Not all AWS Services are available in all regions.

4 factors required to consider while choosing a region :

1. What Regulatory Compliance does this region meet?
2. What is the cost of AWS Services in this region?
3. What AWS Services are available in this region?
4. What is the distance of latency to my end-users?

Regional vs Global Services

Regional Services

AWS scopes their AWS Management Console on a selection Region.
This will determine where an AWS Service will be launched & what will be seen within an AWS Service's console.
You generally don't explicitly set the Region for a service at the time of creation.

Global Services

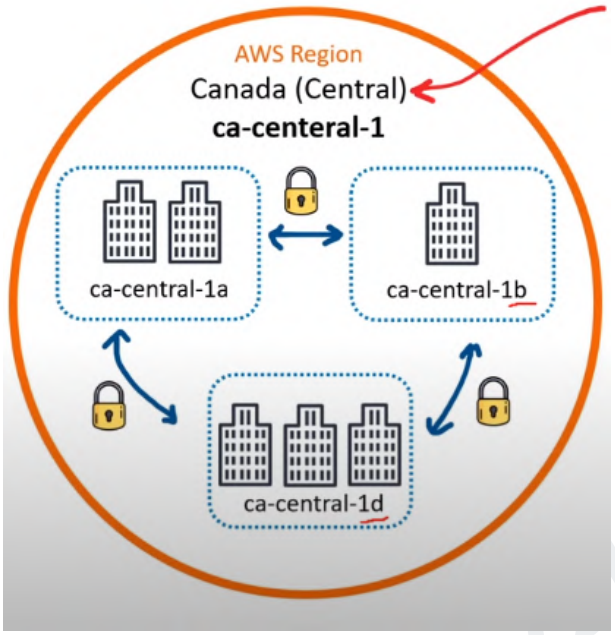
Some AWS Services operate across multiple regions & the region will be fixed to "Global"
E.g. : Amazon S3, CloudFront, Route53, IAM

For these global services at the time of creation :

- There's no concept of region eg. IAM User.
- A single region must be explicitly chosen eg. S3 Bucket.
- A group of regions are chosen eg. CloudFront Distribution.

Availability Zones (AZ)

- An Availability Zone (AZ) is a physical location made up of one or more data centres.
- A data centre is a secured building that contains hundreds of thousands of computers.
- A region will generally contain 3 Availability Zones.
    a) Data centres within a region will be isolated from each other (different buildings). But then will be close enough to provide low-latency (<10ms).
    b) It's common practice to run workloads in at least 3 AZs to ensure services remain available in case one or two data centres fail. (High Availability)
    c) AZs are represented by a Region Code, followed by a letter identifier. Eg. **us-east-1a.**
- su
- You never choose the AZ when launching resources. You choose the Subnet which is associated with AZ.
- The US-EAST-1 Region has 6 AZs which is the most of any region.
- Example of an architectural diagram, representing two AZs, the Subnets associated with those AZs, & EC2 instances (Virtual Machines) launched are those subnets.
- All AZ's in an AWS Region are interconnected with high-bandwidth, low-latency networking, over fully redundant, dedicated metro fibre providing high-throughput & low-latency networking.
- All traffic is encrypted.
- AZs are within 100 km of each other.



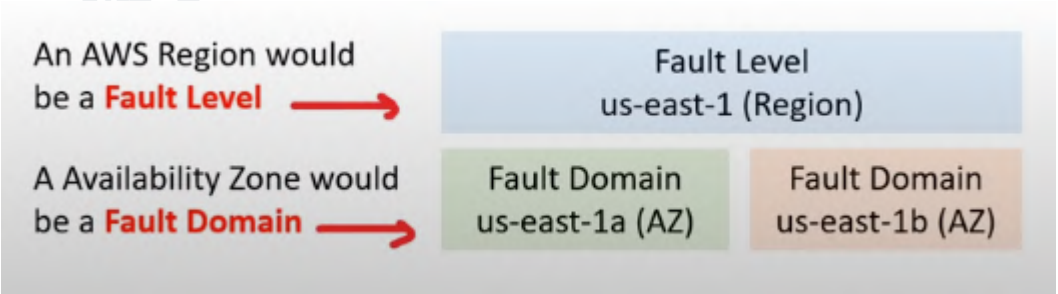
Fault Tolerance

A fault domain is a section of a network that is vulnerable to damage if a critical device or system fails. The purpose of a fault domain is that if a failure occurs it will not cascade outside that domain, limiting the damage possible. You can have fault domains nested inside fault domains.

Fault Level is a collection of fault domains.
The scope of a fault domain could be :

- Specific servers in a rack.
- An entire rack in a data centre.
- An entire room in a data centre.
- The entire data centre building.

It's up to the Cloud Service Provider (CSPs) to define the boundaries of a domain.



Each Amazon Region is designed to be completely isolated from the other Amazon Regions.

- This achieves the greatest possible fault tolerance & stability.
- Each Availability Zone is isolated, but the Availability Zones in a Region are connected through low-latency links.
- Each AZ is designed as an Independent fault domain.

Failure Zone

- AZ's are physically separated within a typical metropolitan region & are located in lower risk flood plains.
- Discrete uninterruptible power supply (UPS) & online backup generation facilities.
- Data centres located in different Availability Zones are designed to be supplied by independent substations to reduce the risk of an event on the power grid impacting more than one AZ.
- AZ are all redundantly connected to multiple tier-1 transit providers.
- Multi-AZ for High Availability : If an application is partitioned across AZs, companies are better isolated & protected from issues such as power outages, lightning strikes, tornadoes, earthquakes & more.

AWS Global Network

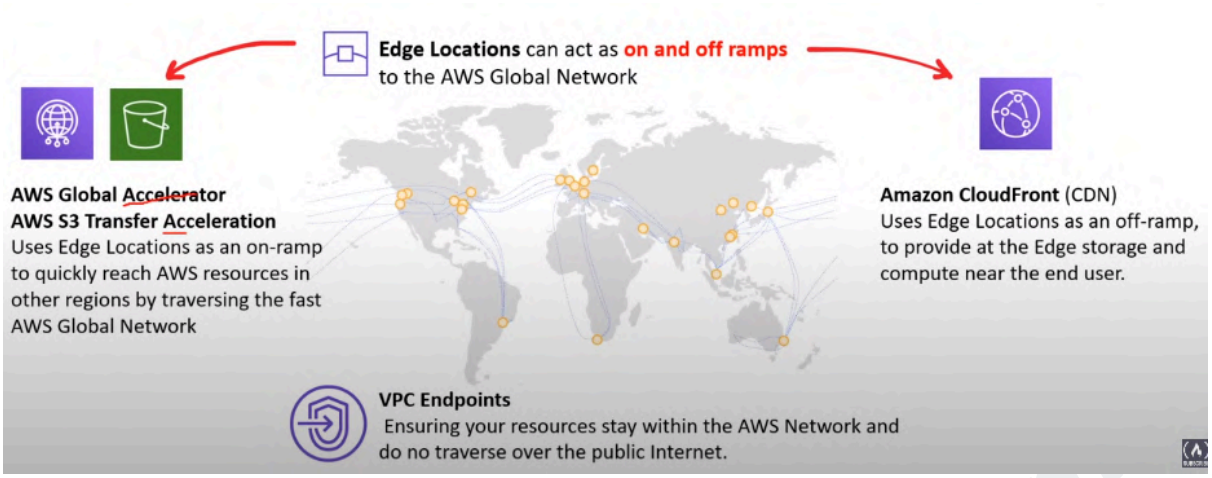
The AWS Global Network represents the interconnections between AWS Global Infrastructure. Commonly referred to as "The Backbone of AWS". A private expressway, where things can move very fast between data centres.

Edge Locations can act as on & off ramps to the AWS Global Network.

AWS Global Accelerator & AWS S3 Transfer Acceleration : uses edge locations as an on-ramp to quickly reach AWS Resources in other regions by traversing the fast AWS Global Network.

Amazon CloudFront (CDN) : Uses Edge Locations as an off-ramp, to provide at the Edge storage & compute near the end user.

VPC Endpoints : Ensuring your resources stay within the AWS Network & do not traverse over the public internet.



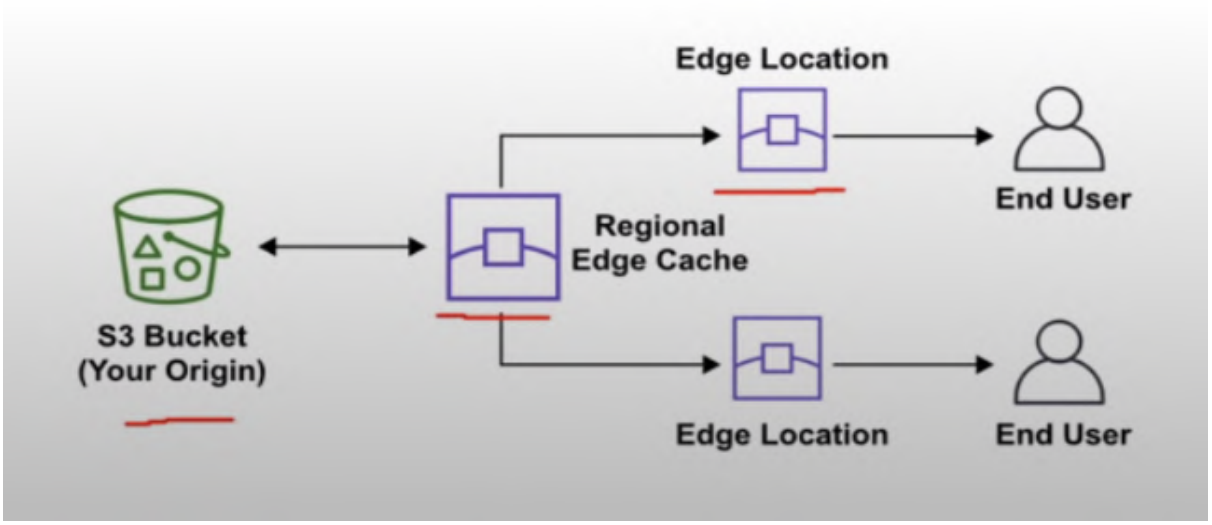
Point of Presence(PoP)

PoP is an intermediate location between an AWS Region & the end user. This location could be a data centre or a collection of hardware.

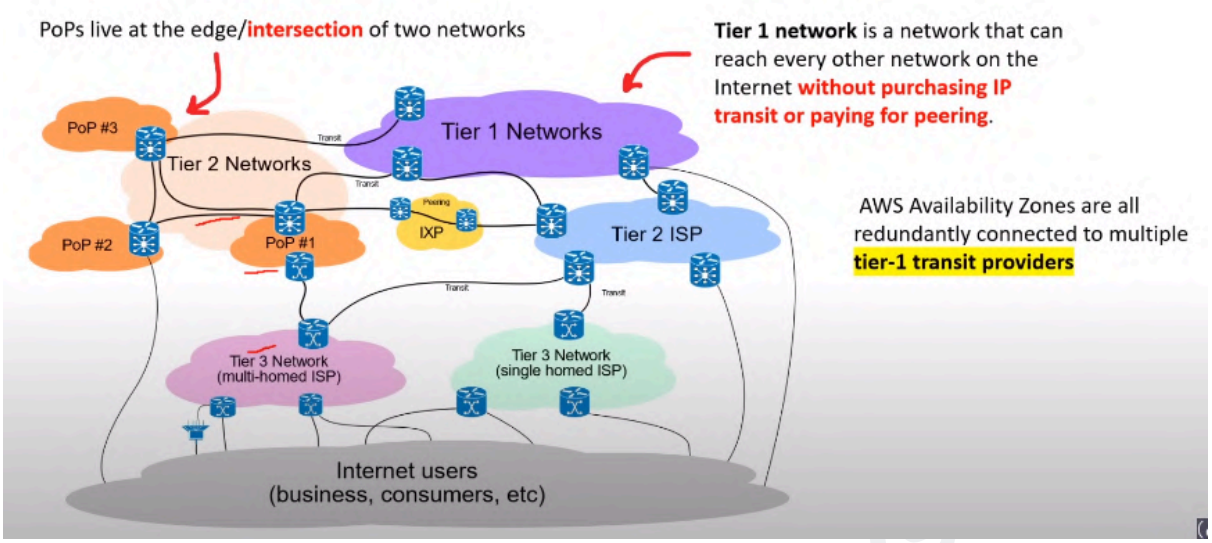
For AWS a PoP is a data centre owned by AWS or a trusted partner that is utilised by AWS Services related for content delivery or expedited upload.
PoP resources are :

1. Edge Locations
2. Regional Edge Caches.

Edge Locations are data centres that hold cached (copy) on the most popular files (webpages, images, videos, etc.) so that the delivery of distance to the users are reduced.
Regional Edge Locations are data centres that hold much larger caches of less-popular files to reduce a full round trip & also the cost of transfer fees.



AWS Cloud Infrastructure - Tier 1



Tier 1 network is a network that can reach every other network on the Internet without purchasing IP transit paying for peering.
AWS Availability Zones are all redundantly connected to multiple tier-1 transit providers.

AWS Services that use PoPs

Amazon CloudFront is a Content Delivery Network (CDN) Service that :

- You point your website to CloudFront so that it will route requests to the nearest Edge Location cache.
- Allows you to choose an origin that will be the source of cache.
- Caches the contents of what origin would return to various Edge Locations around the world.

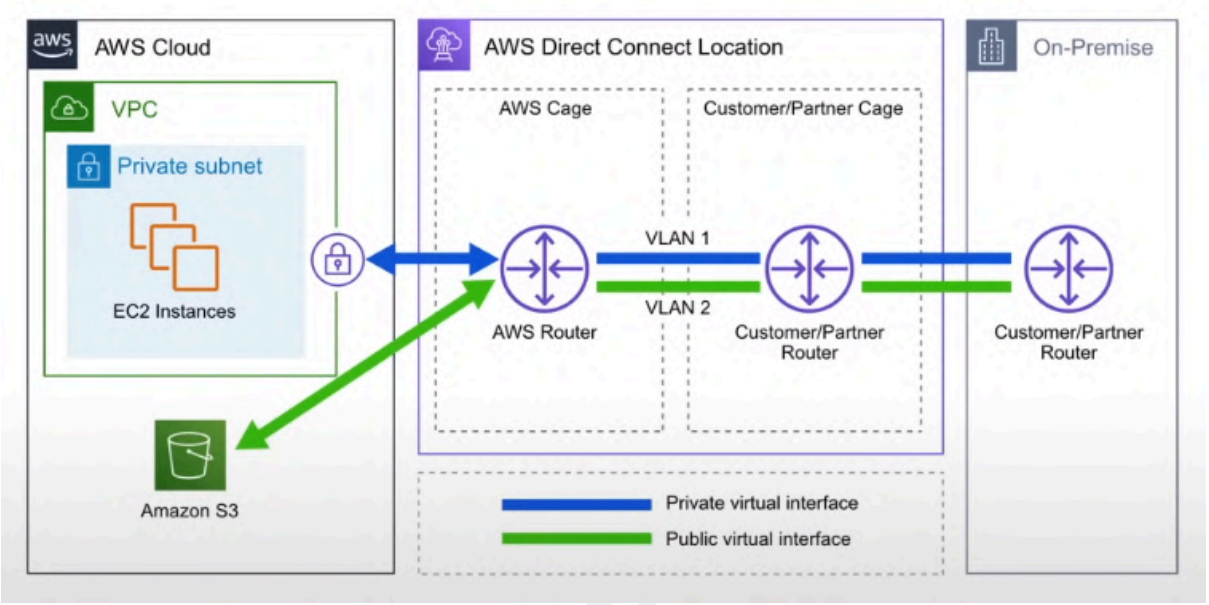
Amazon S3 Transfer Acceleration allows you to generate a special URL that can be used by end users to upload files to a nearby Edge Location. Once a file is uploaded to an Edge Location, it can move much faster within the AWS Network to reach S3.
AWS Global Accelerator can find the optimal path from the end user to your web-servers. Global Accelerators are deployed within Edge Locations so you send user traffic to an Edge Location instead of directly to your web-application.

AWS Direct Connect is a private / dedicated connection between your data centre, office, co-location & AWS.

- Direct Connect has two very-fast network connection options :
    1. **Lower Bandwidth 50MBps - 500MBps**
    2. **Higher Bandwidth 1GBps - 10GBps**

- Helps **reduce network costs** & **increase bandwidth throughput**. (great for high traffic networks)
- Provides a **more consistent network experience** than a typical internet-based connection. (reliable & secure)

A co-location (aka carrier-hotel) is a data centre where equipment, space & bandwidth are available for rental to retail customers.



Direct Connect Locations are trusted partnered data centres that you can establish a dedicated high speed, low-latency connection from your on-premise to AWS.

You would use the AWS Direct Connect Service to order & establish a connection.

Local Zones are data centres located very close to a densely populated area to provide single-digit millisecond low latency performance (eg. 7ms) for that area.

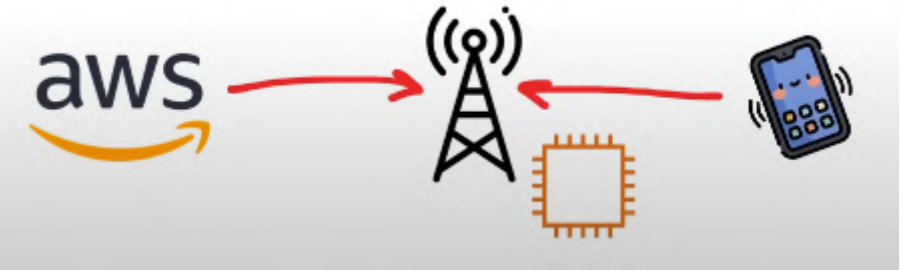
- Los Angeles, California was the first Local Zone to be deployed :
  - It is a logical extension of the US-West Region.
  - The Identifier looks like the following : **us-west-2-lax-1a**
- Only specific AWS Services have been made available :
  - EC2 Instance Types (T3, C5, R5, R5d, I3en, G4)
  - EBS (io1 & gp2)
  - Amazon FSx.
  - Application Load Balancer.
  - Amazon VPC.

The purpose of Local Zone is the support highly-demanding applications sensitive to latencies :

- Media & Entertainment.
- Electronic Design Automation.
- Ad-Tech.
- Machine Learning.

Wavelength Zones

AWS Wavelength Zones allows for edge-computing on 5G Networks. Applications will have ultra-low latency being as close as possible to the users. AWS has partnered with various Telecom companies such as Verizon, KDDI, Vodafone & SK telecom to utilise their 5G networks.



You create a Subnet tied to a Wavelength Zone & then you can launch Virtual Machines (VMs) to the edge of the targeted 5G Networks.

Data Residency is the physical or geographic location of where an organisation or cloud resources reside.

Compliance Boundaries (legal requirement) by a government or organisation that describes where data & cloud resources are allowed to reside.

Data Sovereignty is the jurisdictional control or legal authority that can be asserted over data because its physical location is within jurisdictional boundaries.

For workloads that need to meet compliance boundaries strictly defining the data residency of data & cloud resources in AWS you can use :

AWS Config is a Policy as Code Service. You can create rules to continuously check AWS resources configuration. If they deviate from your expectation you are alerted or AWS Config can in some cases auto-remediate.

IAM Policies can be written explicitly to deny access to specific AWS Regions. A Service Control Policy (SCP) are permissions applied organisation wide.

AWS Outposts is a physical rack of servers that you can put in your data centre. Your data will reside whenever the Outpost Physically resides.

AWS for Government

Federal Risk & Authorisation Management Program (FedRAMP) is a US government-wide program that provides a standardised approach to security assessment, authorisation & continuous monitoring for cloud products & services.

AWS has special regions for US Regulation called **GovCloud.**

Public Sectors include public goods & governmental services such as :

- Military.
- Law Enforcement.
- Infrastructure.
- Public Transit.
- Public Education.
- Health Care.
- Government itself.

AWS can be utilised by the public sector or organisations developing cloud workloads for the public sector.
AWS achieves this by meeting regulatory compliance programs along with specific governance & security controls.

GovCloud is a Cloud Service Provider (CSP) generally offering an isolated region to run FedRAMP workloads.

AWS GovCloud Regions allow customers to host sensitive Controlled. Unclassified Information & other types of regulated workloads.

- GovCloud Regions are only operated by employees who are U.S. citizens, on U.S. soil.
- They are **only accessible to U.S. entities** & root account holders who pass a screening process.

Customers can architect secure cloud solutions that comply with:

- FedRAMP High baseline.
- DOJ's Criminal Justice Information Systems (CJIS) Security Policy.

- U.S. International Traffic in Arms Regulations. (ITAR)
- Export Administration Regulations. (EAR)
- Department of Defense (DoD) Cloud Computing Security Requirements Guide.

AWS in China

AWS China is completely isolated intentionally from AWS Global to meet regulatory compliance for Mainland China. AWS China is on its own domain at : amazonaws.cn

In order to operate in a AWS China Region you need to have a Chinese Business Licence (ICP License)
Not all services are available in China. Eg. Route53
Running in Mainland China means you would not need to traverse The Great Firewall.

Sustainability

Amazon co-founded the Climate Pledge to achieve Net-Zero Carbon Emissions by 2040 across all of Amazon's business including AWS.

1. Renewable Energy : AWS is working towards having their AWS Global Infrastructure powered by 100% renewable energy by 2025.
2. Cloud Efficiency : AWS's infrastructure is 3.6 times more energy efficient than the median of U.S. enterprise data centres surveyed.
3. Water Stewardship :
    - Direct evaporative technology to cool our data centre.
    - Use of non-profitable water for cooling purposes.
    - On-site water treatment allows removal of scale-forming minerals & refuses good scolding for more water cycles.
    - Water Efficiency metrics to determine & monitor optimal water use for a reach.CN, etc.
4. AWS purchases & retires environmental attributes to cover the non-renewable energy for AWS Global Infrastructure :
    - Renewable Energy Credits. (RECs)
    - Guarantees of Origin. (GOs)

AWS Ground Station

It is a fully managed service that lets you control satellite communications, process data & scale your operations without having to worry about building or managing your own ground station infrastructure.

Use cases for Ground Station :

- Weather Forecasting.
- Surface Imaging.
- Communications.
- Video Broadcasts.
- Eg : A company reaches an agreement with a Satellite Imagery Provider to take satellite photos of a specific region. They'll use AWS Ground Station to communicate with that company's satellite & download the S3 Image data.

To use Ground Station :

- You schedule a Contact (select satellite, start & end time, & the ground station)
- Use the AWS Ground Station EC2 AMI to launch EC2 instances that will uplink & downlink data during the contact or receive downlinked data in an Amazon S3 bucket.

AWS Outposts

AWS Outposts is a rack of servers running AWS Infrastructure on your physical location.

It is a fully managed service that offers the same AWS infrastructure, AWS services, APIs, & tools to virtually any data centre, co-location space or on-premises facility for a truly consistent hybrid experience.
Server Rack : A frame design to hold & organise IT equipment.
Rack Heights : U stands for "rack units'' or "U spaces'' which is equal to 1.75 inches. The industry standard rack size is 48U (7 Foot Rack)
Full-size rack cage is 42U high : equipment is typically 1U, 2U, 3U or 4U high.

AWS Outposts comes in 3 form factors : 42U, 1U & 2U
Full Rack of servers : 42U
Servers for Existing Racks : 1U / 2U

42U : AWS delivers it to your preferred physical site fully assembled & ready to be rolled into the final position. It is installed by AWS & the rack needs to be simply plugged into power & network.

1U : Suitable for 19-inch wide,
24-inch deep cabinets.

AWS Graviton2 (up to 64 vCPUs).
128 GiB Memory.
4 TB of local NVMe Storage.

2U : Suitable for 19-inch wide,
36-inch deep cabinets.
Intel Processor (up to 128 vCPUs)
256 GiB Memory.
8 TB of local NVMe Storage.

Cloud Architecture

Solutions Architect is a role in a technical organisation that architects a technical solution using multiple systems via researching, documentation, experimentation.

Cloud Architect is a solutions architect that is focused solely on architecting technical solutions using cloud services.

A cloud architect needs to understand the following terms & factor them into their designed architecture based on the business requirements.

1. **Availability** - Your ability to ensure a service remains available eg. *Highly Available* (HA)
2. **Scalability** - Your ability to grow rapidly or unimpeded.
3. **Elasticity** - Your ability to shrink & grow to meet the demand.
4. **Fault Tolerance** - Your ability to prevent a failure.
5. **Disaster Recovery** - Your ability to recover from a failure eg. *Highly Durable* (DR)

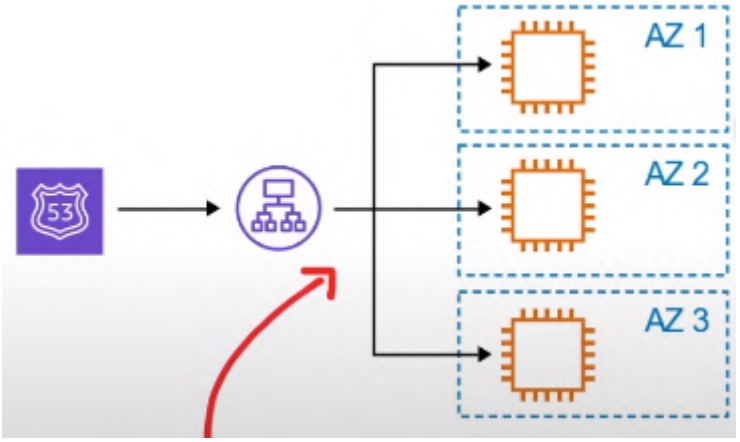
A Solutions Architect needs to always consider the following business factors :

- **Security** : How Secure is the solution?
- **Cost** : How much is this going to cost?

1. High Availability

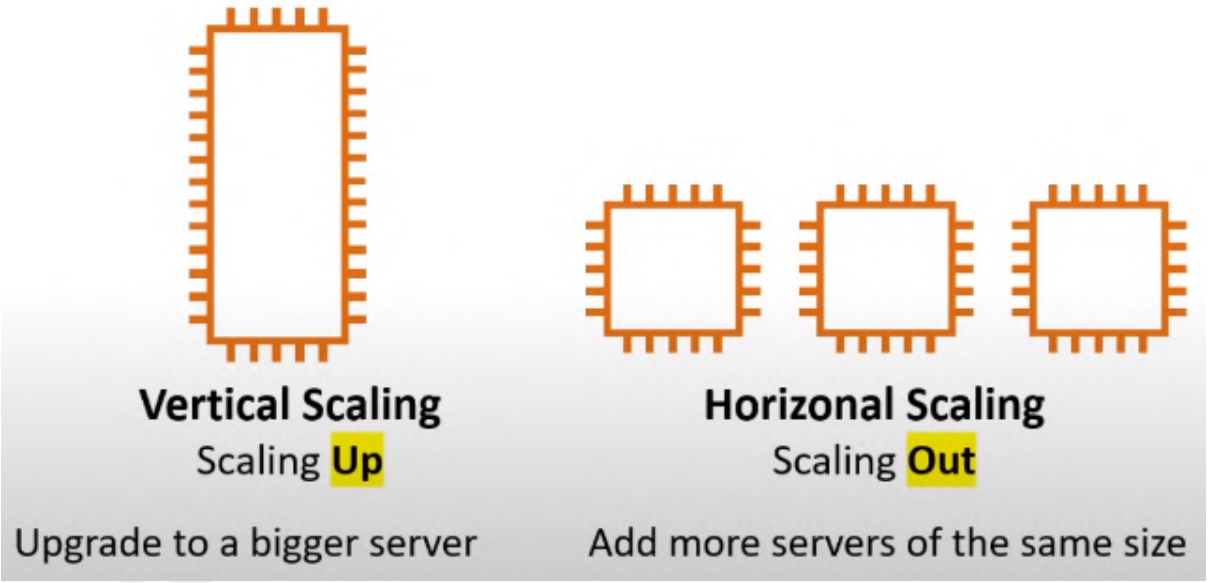
Your ability for your service to remain available by ensuring there is *no single point of failure &/or ensure a certain level of performance.

> **Elastic Load Balancer** allows you to evenly distribute traffic to multiple servers in one or more data centres. If a data centre or server becomes unavailable the load balancer will route the traffic to only available data centres with servers.
> Running your workload across multiple AZ ensures that if 1 or 2 AZs become unavailable your service / applications remain available.



2. High Scalability

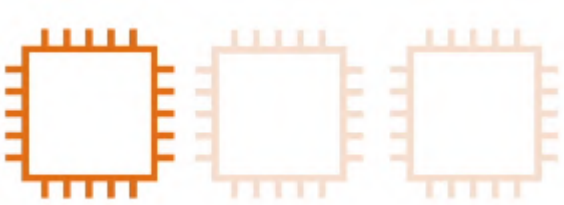
Your ability to increase your capacity based on the increasing demand of traffic, memory & computing power.



3. High Elasticity

Your ability to automatically increase / decrease your capacity based on the current demand of traffic, memory & computing power.

> **Auto Scaling Groups (ASG)** is an AWS Feature that will automatically add / remove servers based on scaling rules you define based on metrics.



Horizontal Scaling
Scaling Out - Add more servers of the same size
Scaling In - Removing underutilised servers of the same size.

Vertical Scaling is generally hard for traditional architecture so you'll usually only see horizontal scaling described with Elasticity.

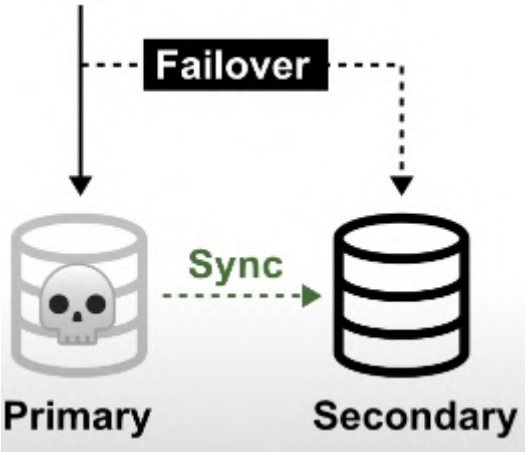
4. Highly Fault Tolerant

Your ability for your service to ensure there's no single point of failure.
Preventing the chance of failure.

Fail-overs is when you have a plan to *shift traffic* to a redundant system in case the primary system fails.

RDS Multi-AZ is when you run a duplicate standby database in another AZ in case your primary database fails.

A common example is having a copy (secondary) of your database where all ongoing changes are synced. The secondary system is not in use until a failover occurs & it becomes the primary database.



5. High Durability

Your ability to recover from a disaster & to prevent the loss of data Solutions that recover from a disaster is known as Disaster Recovery (DR)

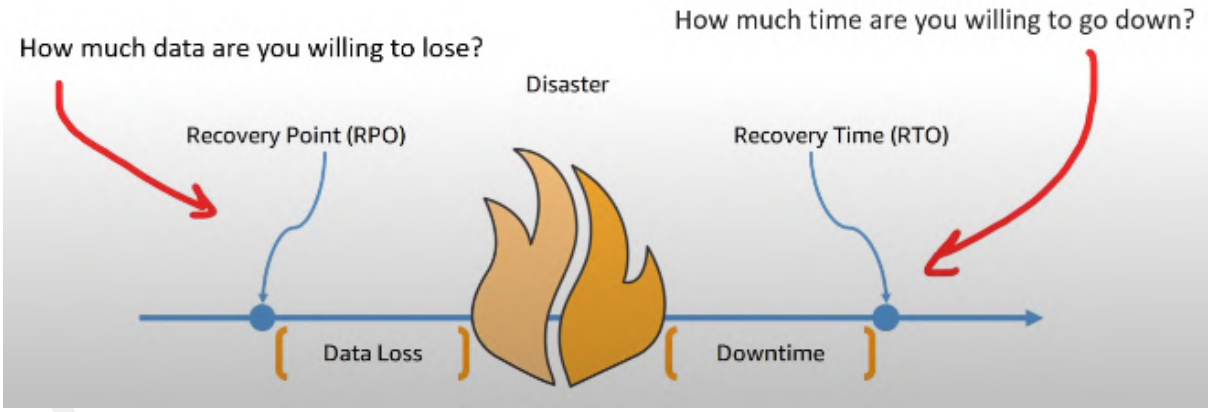
- Do you have a backup?
- How fast can you restore that backup?
- Does your backup still work?
- How do you ensure current live data is not corrupt?

CloudEndure Disaster Recovery continuously replicates your machines into a low-cost staging area in your target AWS Account & preferred Region enabling fast & reliable recovery in case of IT data centre failures.

Business Continuity Plan (BCP)

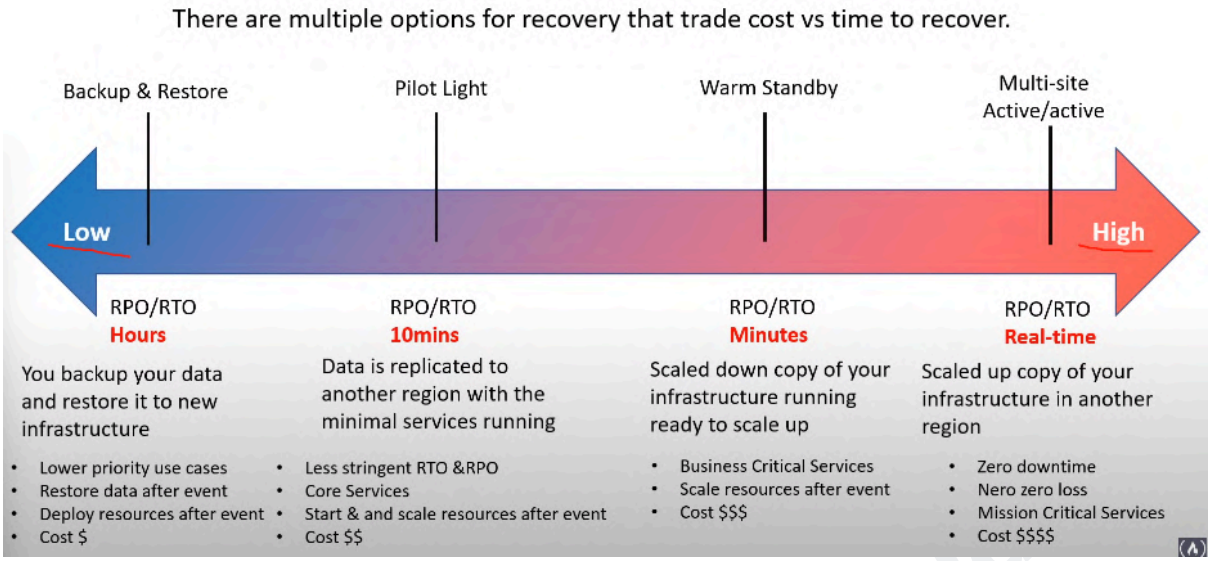
A **BCP** is a document that outlines how a business will continue operating *during an unplanned disruption in services.*

Recovery Point Objective (RPO) the maximum acceptable amount of data loss after an unplanned data-loss incident, expressed as an amount of time.
Recovery Time Objective (RTO) the maximum amount of downtime your business can tolerate without incurring a significant financial loss.

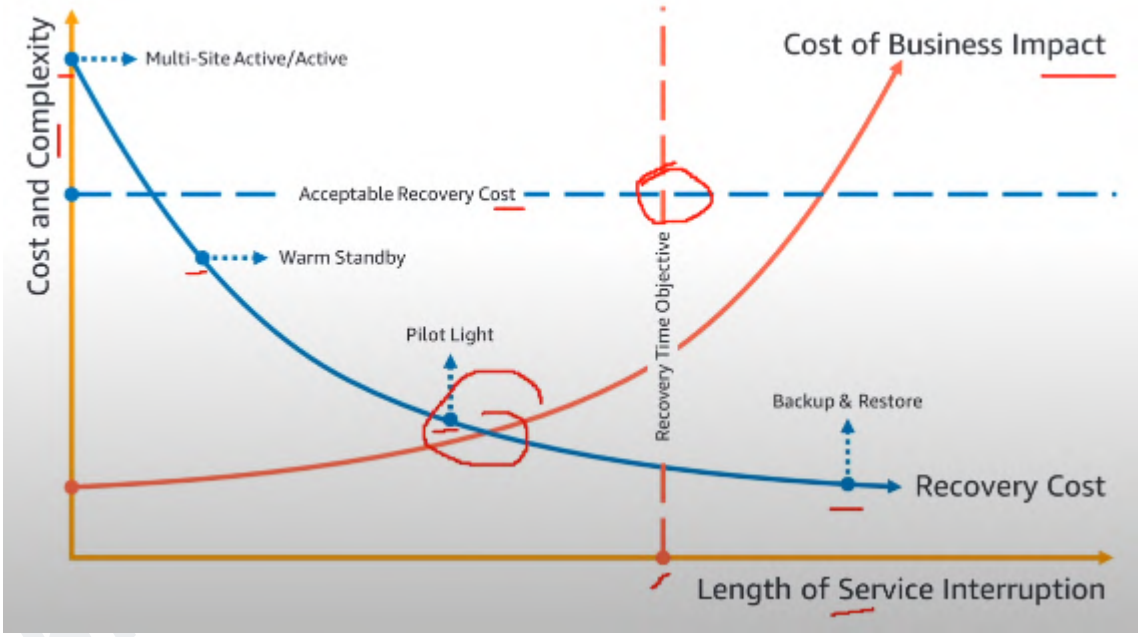


Disaster Recovery Options

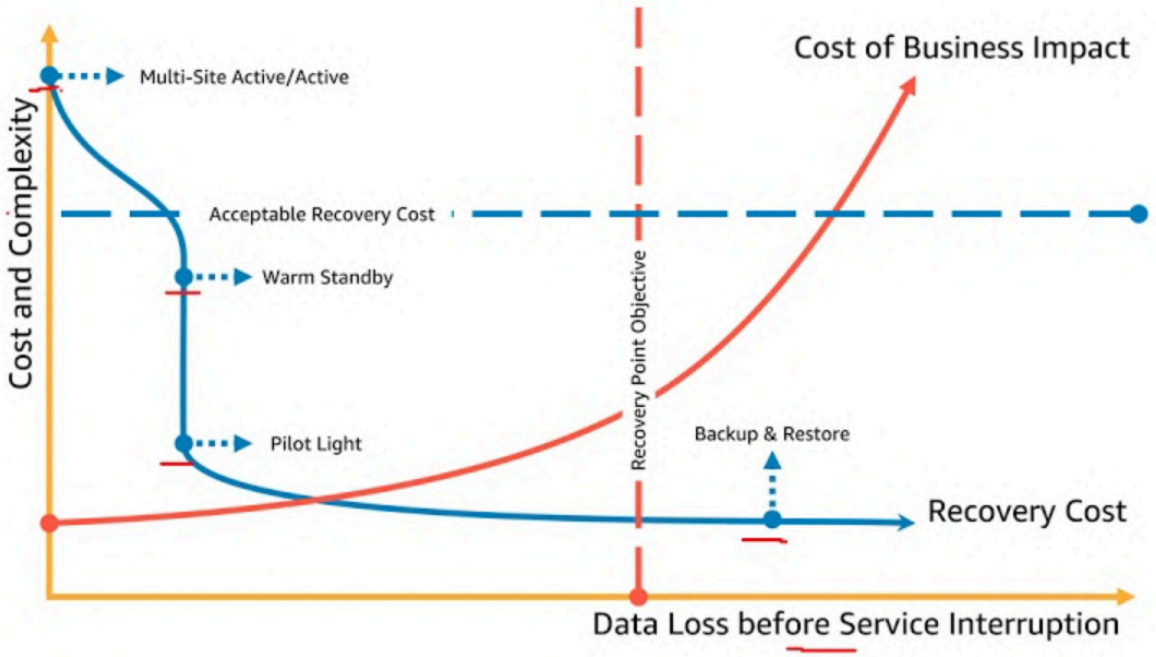
There are multiple options for recovery that trade cost vs time to recover.



Recovery Time Objective (RTO) is the maximum acceptable delay between the interruption of service & restoration of service. This objective determines what is considered an acceptable time window when service is unavailable & is defined by the organisation.



Recovery Point Objective (RPO) is the maximum acceptable amount of time since the last data recovery point. This objective determines what is considered as an acceptable loss of data between the last recovery point & the interruption of service & is defined by the organisation.

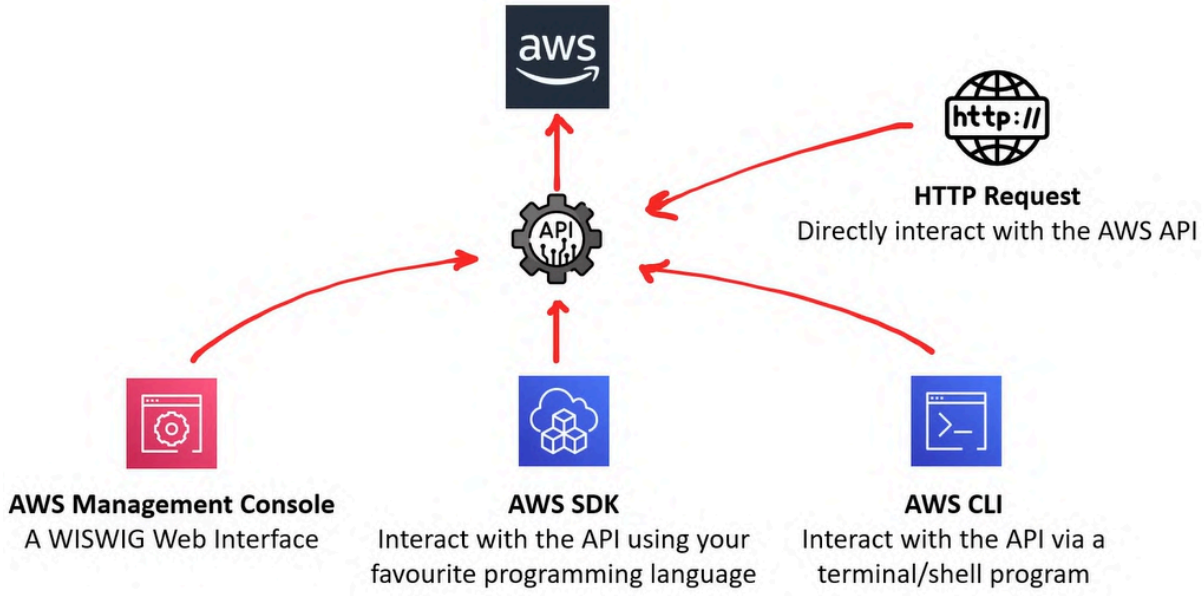


Management & Development Tools

AWS Application Programming Interface (API)

Application Programming Interface (API) is a software that allows two applications / services to talk to each other. Most common type of API is via HTTP/S requests.

AWS API is an HTTP API & you can interact by sending HTTPS requests using an application interacting with APIs like *Postman*.



AWS Management Console is a web-based unified console Build, manage & monitor everything from simple web apps to complex cloud deployments.

ClickOps : The process of pointing & clicking to manually launch & configure AWS Resources with limited programming knowledge.

Service Console

AWS Services each have their own customised console. You can access these consoles by *searching* the service name.

Some AWS Services Consoles will act as an umbrella containing many AWS Services.

Examples

- VPC Console.
- EC2 Console.
- Systems Manager Console.
- SageMaker Console.
- CloudWatch Console.

AWS Account ID

Every AWS Account has a unique Account ID. The Account ID can be easily found by dropping the current user in the Global Navigation.

The AWS Account ID is composed of 12 digits eg : 123456789012

AWS Account ID can be used for

- When logging in with a non-root user account.
- Cross-account roles.
- Support cases.

AWS Tools for PowerShell

Powershell is a task automation & configuration management framework. A command-line shell & a scripting language.
Powershell is *built on top of the .NET Common Language Runtime (CLR)* & accepts & returns .NET objects.

AWS Tools for Powershell lets you interact with the AWS API via Powershell Cmdlets.

Amazon Resource Names (ARNs)

Amazon Resource Names (ARNs) uniquely identify AWS resources. They are required to specify a resource unambiguously across all of AWS.

Format Variations :

- arn:partition:service:region:account-id:resource-id
- arn:partition:service:region:account-id:resource-type/resource-id
- arn:partition:service:region:account-id:resource-type:resource-id

Partition :

- aws - AWS Regions
- aws-cn : China Regions
- aws-us-gov : AWS GovCloud (US) Regions

Service - identifies the service :

- ec2
- s3
- iam

Region - which AWS resource :

- us-east-1
- ca-central -1

Account ID :

- 121212121212

Resource ID - Could be a number name or path :

- user/Bob
- instance/i-1234567890abcdef0

In the AWS Management Console it's common to be able to copy the ARN to your clipboard.

Paths in ARNs

Resource ARNs can include a path.
Paths can include a wildcard character, namely an asterisk (*)

IAM Policy ARN Path

arn:aws:iam::123456789012:user/Development/product_1234/*

S3 ARN Path

arn:aws:s3:::my_corporate_bucket/Development/*

AWS Command Line Interface (CLI)

Command Line Interface (CLI) processes commands to a computer program in the form of lines of text.
Operating systems implement a command-line interface in a shell.

Terminal is a text only interface (input / output environment).

Console is a physical computer to physically input information into a terminal.

Shell is the command line program that users interact with to input commands.
Popular shell programs :

- Bash
- Zsh
- PowerShell

AWS Command Line Interface (CLI) allows users to programmatically interact with the AWS API via entering single or multi-line commands into a shell or terminal.

The AWS CLI is a Python executable program

- Python is required to install AWS CLI.

The AWS CLI can be installed on Windows, Mac or Linux/Unix
The name of the CLI program is **aws**.

AWS Software Development Kit (SDK)

A **Software Development Kit (SDK)** is a collection of software development tools in one installable package.

AWS SDK can be used to programmatically create, modify, delete or interact with AWS Resources.

AWS SDK is offered in various programming languages :

- Java
- Python
- Node.js
- Ruby
- Go
- .NET
- PHP
- JavaScript
- C++

AWS CloudShell

AWS CloudShell is a browser-based shell built into the AWS Management Console.

AWS CloudShell is scoped per region, Same credentials as logged in user. Free Service!
AWS CloudShell is only available in selected regions.

Pre-installed Tools : AWS CLI, Python, Node.js git, make, pip, sudo, tar, tmux, vim, wget, zip & more

Storage Included : 1 GB of storage free per AWS Region.

Saved Files & Settings : Files saved in your home directory are available in future sessions for the same AWS Region.

Shell Environments :
Seamlessly switch between

- Bash
- PowerShell
- Zsh

Infrastructure as Code (IaC)

You write a configuration script to *automate creating, updating or destroying* cloud infrastructure.
Infrastructure as Code (IaC) is a blueprint of an infrastructure.
IaC allows you to easily share,version or inventory your cloud infrastructure.

AWS has two offerings for writing Infrastructure as Code.

1. **AWS CloudFormation (CFN) :** A Decorative IaC tool.
   Declarative :
   - What you see is what you get. ***Explicit***
   - More verbose, but zero chance of mis-configuration.
   - Uses scripting languages eg. JSON, YAML, XML.

   -Allows you to write Infrastructure as Code (IaC) as either a JSON or YAML file.
   -CloudFormation is simple but it can lead to large files or is limited in some regard to creating dynamic or repeatable infrastructure compared to CDK.
   -CloudFormation can be easier for DevOps Engineers who do not have a background in web programming languages.
   -Since CDK generates CloudFormation it's still important to be able to read & understand CloudFormation in order to debug IaC stacks.

2. **AWS Cloud Development Kit (CDK) :** An Imperative IaC tool.
   Imperative :
   - You say what you want & the rest is filled in. ***Implicit***
   - Less verbose, you could end up with misconfiguration
   - Does more than Declarative.
   - Uses programming languages eg. Python, Ruby, JavaScript.

Cloud Development Kit

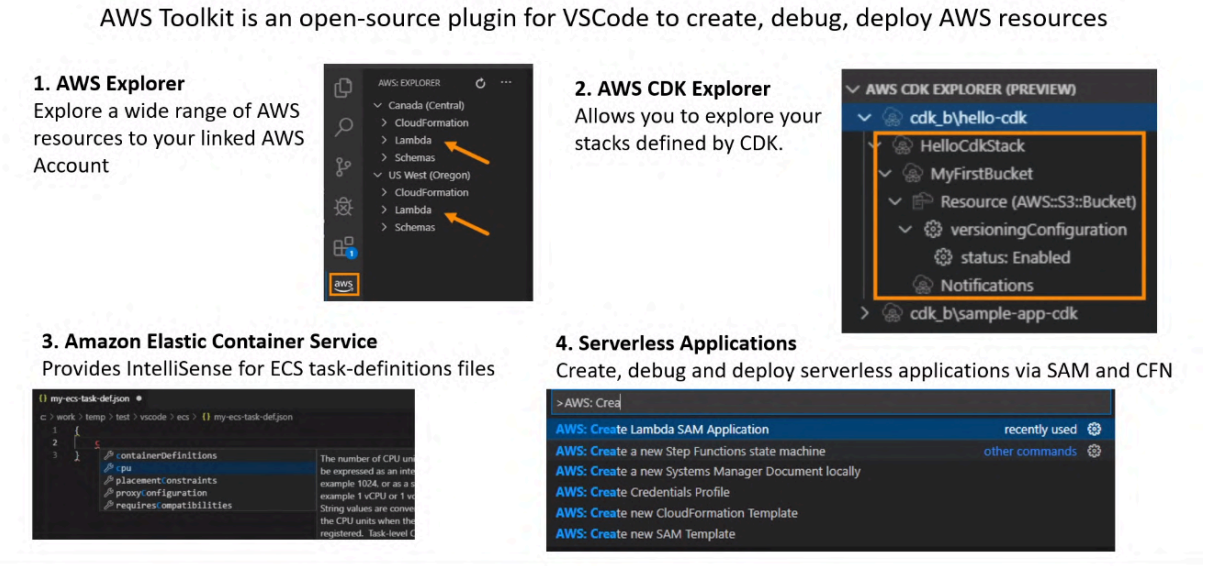
AWS CDK allows you to use your favourite programming language to write Infrastructure as Code.

- CDK is powered by CloudFormation. (it generates out CloudFormation templates)
- CDK has a large library of reusable cloud components called CDK Construct .
  - https://constructs.dev
- CDK comes with its own CLI.
- CDK Pipelines to quickly setup CI/CD pipelines for CDK projects.
- CDK has a testing framework for Unit & Integration Testing.

SDK looks similar but the key difference is CDK ensures Idempotent of your Infrastructure.

AWS Toolkit for VSCode

AWS Toolkit is an open-source plugin for VSCode to create, debug & deploy AWS resources.



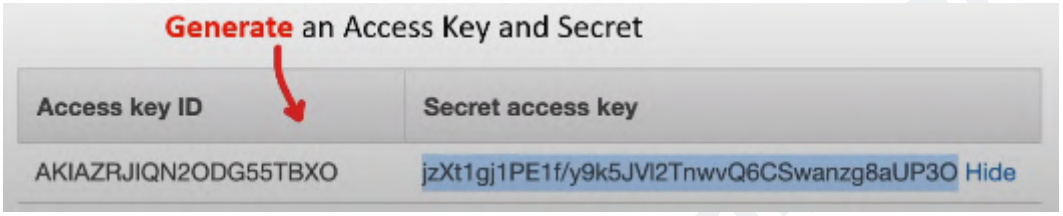
Access Keys

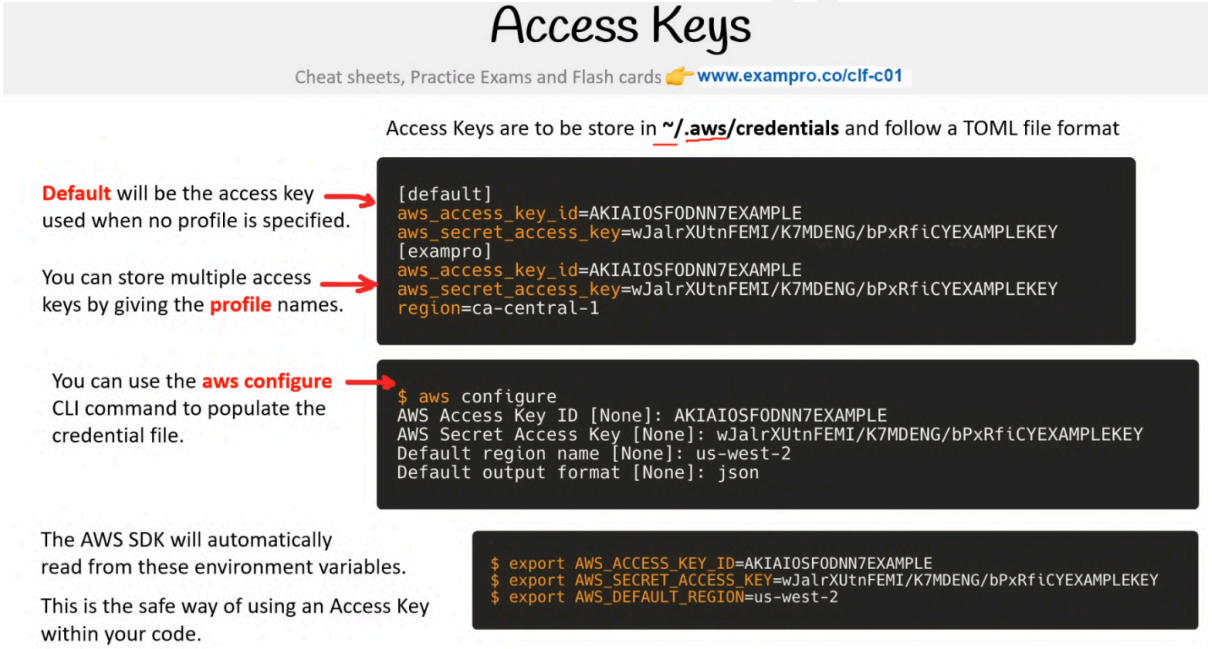
Access Keys is a **key & secret** required to have programmatic access to AWS resources when interacting with the AWS API outside of the AWS Management Console.

An Access Key is commonly referred to as AWS Credentials

A user must be granted access to use Access Keys

- Never share your access keys.
- Never commit access keys to a codebase.
- You can have two active Access Keys.
- You can deactivate Access Keys.
- Access Keys have whatever access a user has to AWS resources.





AWS Documentation

AWS Documentation is a large collection of technical documentation on how to use AWS Services.

- Amazon Documentation

AWS is very good about providing detailed information about every AWS service.

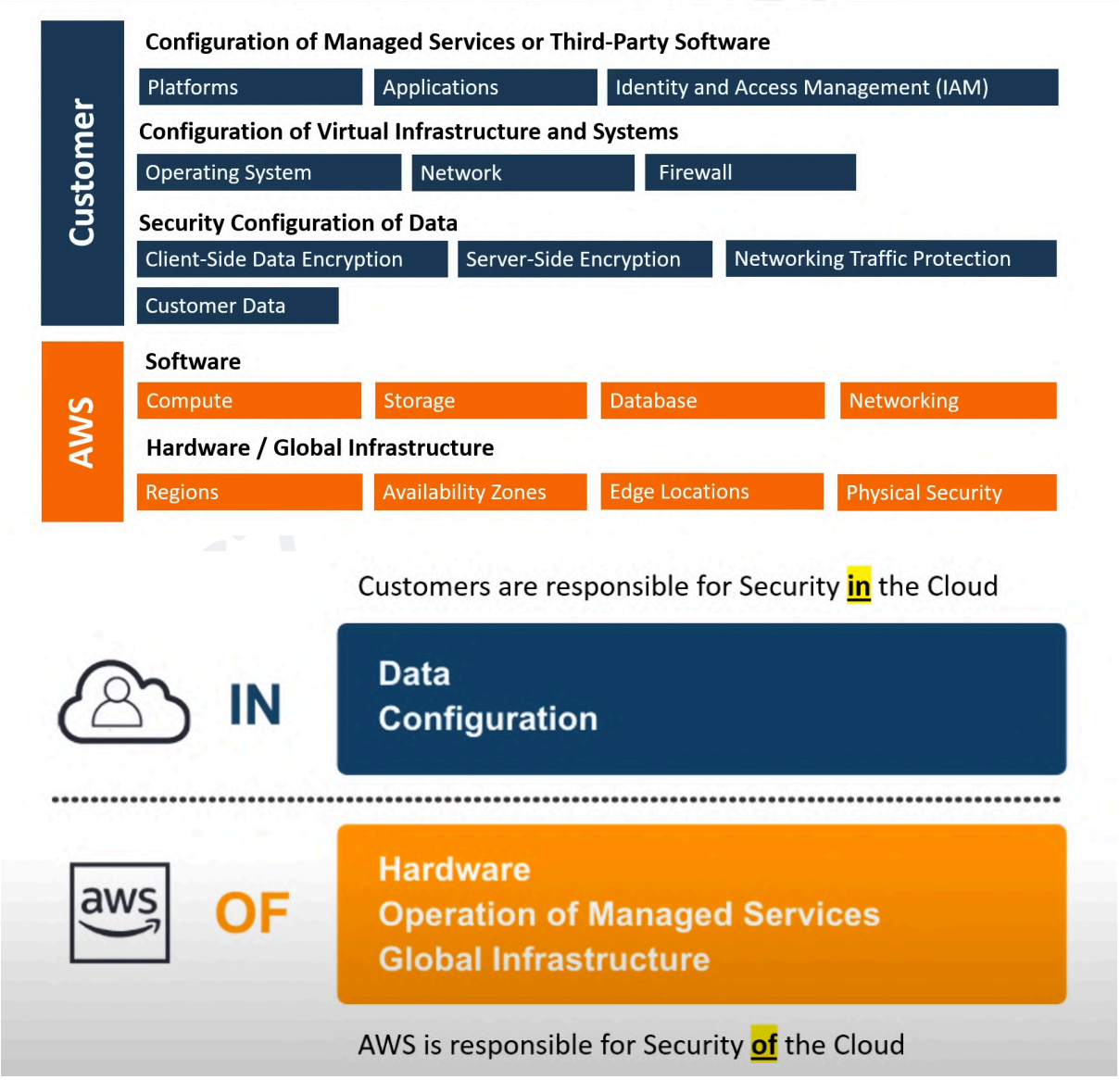
The basis of this course & for any AWS Certification will derive mostly from the AWS Documentation.

Shared Responsibility Model

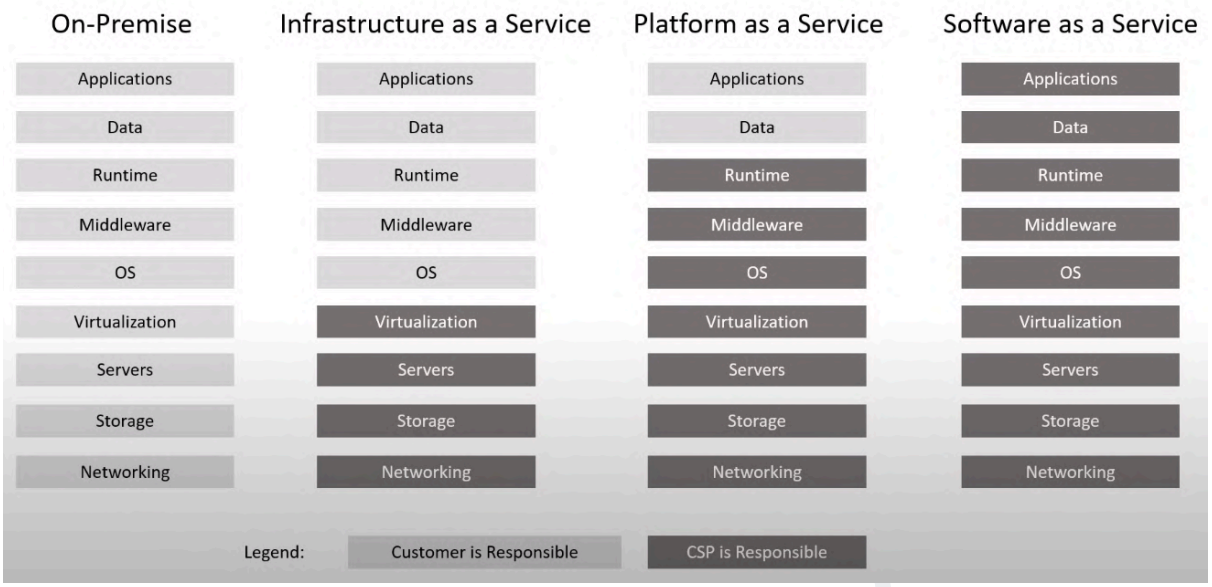
Introduction to Shared Responsibility Model

Shared Responsibility Model is a cloud security framework that defines the security obligations of the customer versus the Cloud Service Provider (CSP).
Example - AWS
Each CSP has their own variant of the Shared Responsibility Model but they are all generally the same.
The type of cloud deployment model &/or the scope of cloud service category can result in specialised Shared Responsibility Models.

AWS Shared Responsibility Model



SRM According to Service



SRM According to Compute

Let us take a look at **compute** as a comparison example of the Shared Responsibility Model

Infrastructure as a Service (IaaS)

Bare Metal
EC2 Bare Metal Instance

Customer:
- The Host OS Configuration
- Hypervisor

AWS
- Physical machine

Virtual Machine
Elastic Cloud Compute (EC2)

Customer:
- The Guest OS Configuration
- Container Runtime

AWS
- Hypervisor, Physical machine

Containers
AWS Elastic Container Service(ECS)

Customer:
- Configuration of containers
- Deployment of Containers
- Storage of containers

AWS
- The OS, The Hypervisor, Container Runtime

Platform as a Service (PaaS)

Managed Platform
AWS Elastic Beanstalk

Customer:
- Uploading your code
- Some configuration of environment
- Deployment strategies
- Configuration of associated services

AWS
- Servers, OS, Networking, Storage, Security

Software as a Service (SaaS)

Content Collaboration
Amazon WorkDocs

Customer:
- Contents of documents
- Management of files
- Configuration of sharing access controls

AWS
- Servers, OS, Networking, Storage, Security

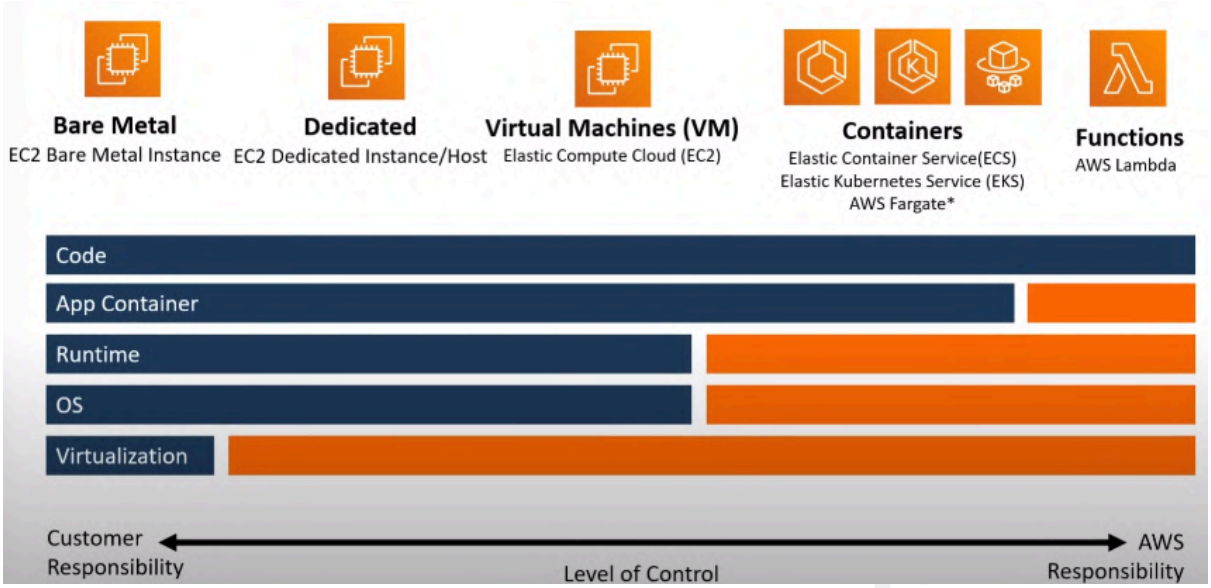
Function as a Service (FaaS)

Functions
AWS Lambda

Customer:
- Upload your code

AWS
- Deployment, Container Runtime, Networking, Storage, Security, Physical Machine, (basically everything)

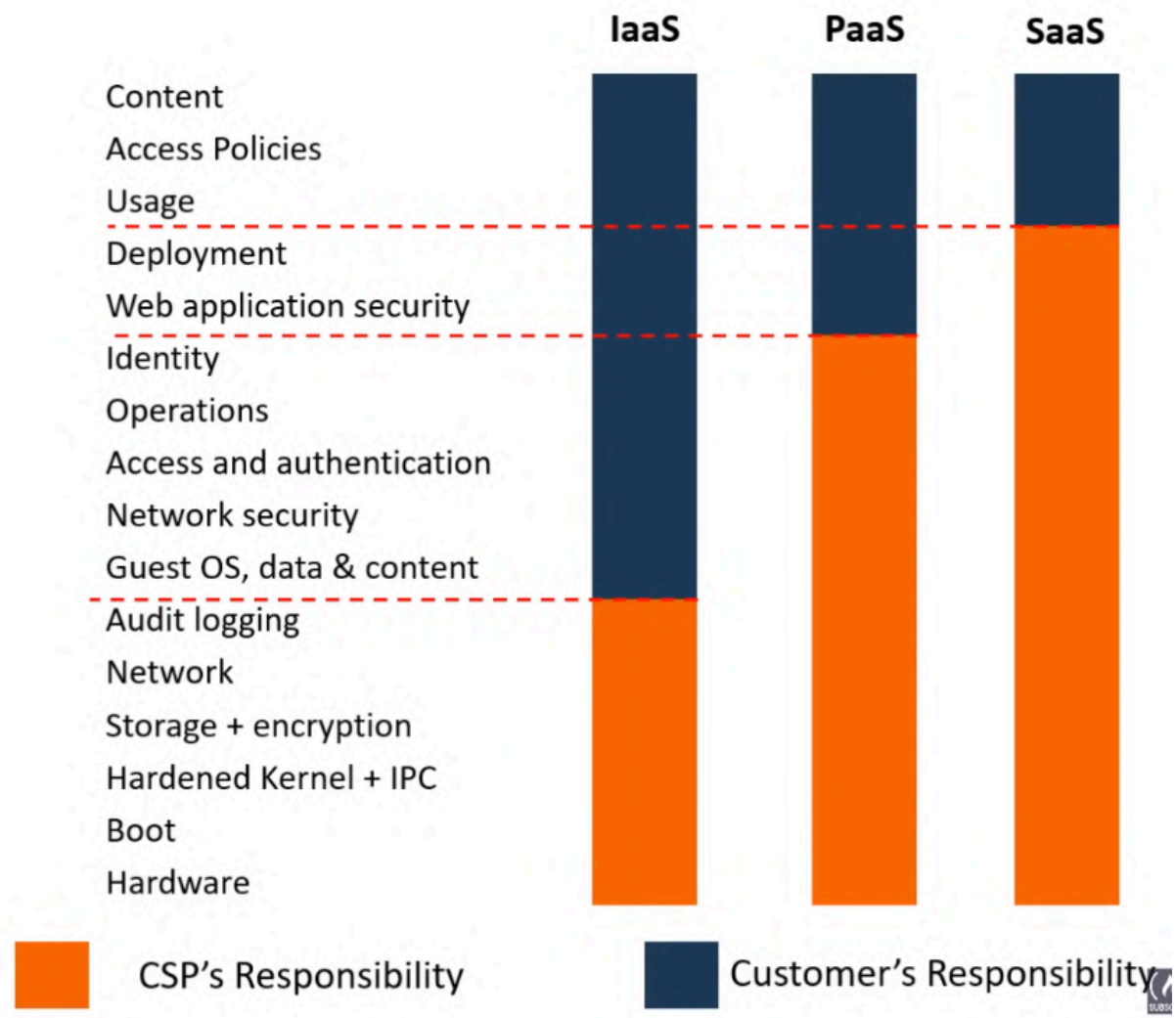
SRM According to Level of Control



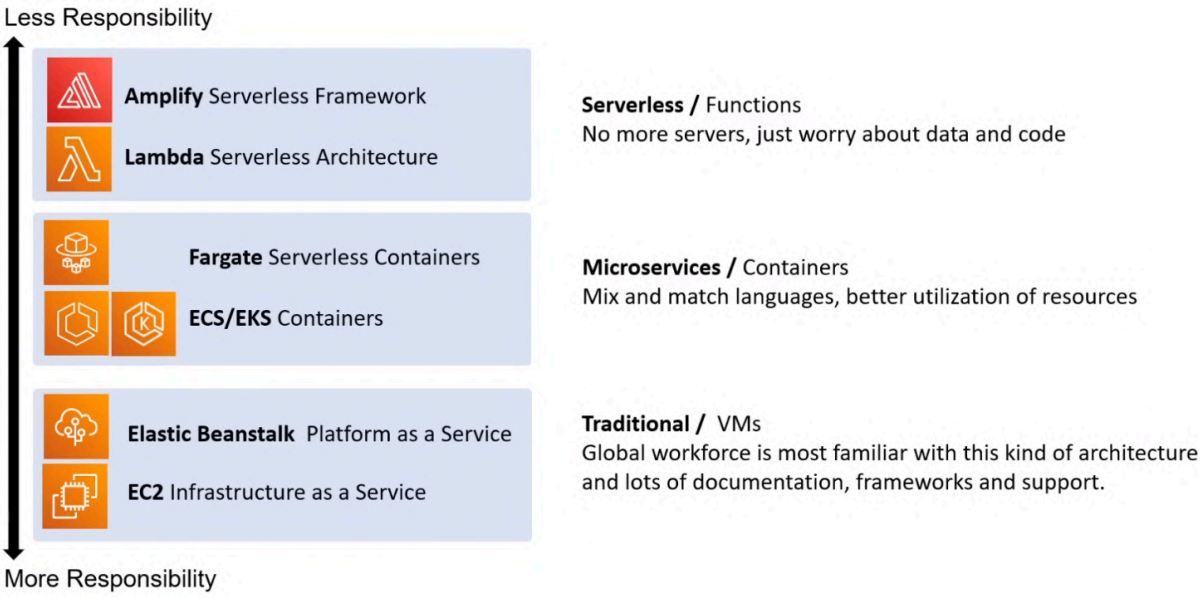
Shared Responsibility Model is a simple visualisation that helps determine what the customer is responsible for & what the CSP is responsible for related to AWS.
The customer is responsible for the data & the configuration of access controls that resides in AWS.
The customer is responsible for the configuration of cloud services & granting access to users via permissions.

CSP is generally responsible for the underlying Infrastructure.

Responsibility of assets in the Cloud :
If you can configure or store it then you are responsible for it.
Responsibility of assets of the Cloud :
If you can not configure it then CSP is responsible for it.



SRM According to Architecture



Serverless / Functions
No more servers, just worry about data and code

Microservices / Containers
Mix and match languages, better utilization of resources

Traditional / VMs
Global workforce is most familiar with this kind of architecture and lots of documentation, frameworks and support.

Computing Services

Elastic Compute Cloud (EC2)

EC2 allows you to launch Virtual Machines (VM)

What is a Virtual Machine?

A VM is an emulation of a physical computer using software.
Server Virtualization allows you to easily create, copy, resize or migrate your server.
Multiple VMs can run on the same physical server so you can share costs with other customers.
Imagine if your server or computer was an executable file on your computer.
When we launch a VM we call it an 'Instance'

EC2 is a highly configurable server where you can choose AMI that affects options such as

- The amount of CPUs
- The amount of Memory (RAM)
- The amount of Network Bandwidth
- The Operation System (OS) eg. Windows 10, Ubuntu, Amazon Linux 2
- Attach multiple virtual hard-drives for storage. Example - Elastic Block Store (EBS)

An **Amazon Machine Image (AMI)** is a predefined configuration for a Virtual Machine.

EC2 is also considered the backbone of AWS because the majority of AWS services are using EC2 as their underlying servers. Examples - S3, RDS, DynamoDB, Lambdas.

Virtual Machines - an emulation of a physical computer using software

1. **Amazon LightSail** is a managed virtual server service. It is the "friendly" version of EC2 Virtual Machine.
   When you need to launch a Linux or Windows server but don't have much AWS knowledge. Example - Launch a Workpress

Containers - virtualizing an Operating System (OS) to run multiple workloads on a single OS instance. Containers are generally used in micro-service

architecture (when you divide your application into smaller applications that talk to each other

1. **Elastic Container Service (ECS) is a container orchestration service that supports Docker containers.** Launches a cluster of server(s) on EC2 instances with Docker installed. When you need Docker as a Service, or you need to run containers.
2. **Elastic Container Registry (ECR) is a repository for container images.** In order to launch a container you need an image. An image just means a saved copy. A repository just means a storage that has version control.
3. **ECS Fargate is a serverless orchestration container service.** It is the same as ECS except you have a pay-on-demand port running container (With ECS you have to keep an EC2 server running even if you have no containers running) AWS manages the underlying server, so you don't have to scale or upgrade the EC2 Server.
4. **Elastic Kubernetes Service (EKS) is a fully managed Kubernetes service.** Kubernetes (K8) is an open-source orchestration software that was created by Google & is generally the standard for managing microservices. When you need to run Kubernetes as a Service.

Serverless - when the underlying servers are managed by AWS. You don't worry or configure servers.

1. **AWS Lambda is a serverless functions service.** You can run code without provisioning or managing servers. You upload a small piece of code, choose how much memory & how long the function is allowed to run before timing out. You are charged based on the runtime of the serverless function rounded to nearest 100ms.

Higher Performance Computing Services

A cluster of hundreds of thousands or servers with fast connections between each of them with the purpose of boosting computing capacity. When you need a supercomputer to perform computational problems too large to run on a standard computer or would take too long.

The **Nitro System is a combination of dedicated hardware & lightweight hypervisor** enabling faster innovation & enhanced security. All new EC2 instance types use the Nitro System.

- Nitro Cards - specialised cards for VPC, EBS, Instance Storage & Controller Card.

- Nitro Security Chips - integrated into the motherboard. Protects hardware resources.
- Nitro Hypervisor - lightweight hypervisor Memory & CPU allocation Bare Metal-like performance.

Bare Metal Instance is a EC2 instance that can be launched without having a hypervisor. This means you can run workloads directly on the hardware for maximum performance & control. The **M5** & **R5** EC2 instances run are bare metal.

Bottlerocket is a Linux-based open-source operating system that is purpose-built by AWS for running containers on VMs or Bare Metal Hosts.

AWS ParallelCluster is an AWS-supported open source cluster management tool that makes it easy for you to deploy & manage High Performance Computing (HPC) clusters on AWS.

Edge & Hybrid Computing Services

Edge Computing is when you push your computing workloads outside of your networks to run close to the destination location. Example - Pushing computing to run on phones, IoT devices or external servers not within your cloud network.

Hybrid Computing is when you're able to run workloads on both your on-premise datacenter & AWS Virtual Private Cloud (VPC)

AWS Outposts is a physical rack of servers that you can put in your datacenter. AWS Outposts allows you to use AWS API & Services such as EC2 right in your datacenter.

AWS Wavelength allows you to build & launch your application in a telecom datacenter. By doing this your applications will have ultra-low latency since they'll be pushed over a 5G network & be closest to the end user.

VMWare Cloud on AWS allows you to manage on-premise virtual machines using VMWare as EC2 instances.

AWS Local Zones are edge data centres located outside of an AWS region so you can use AWS closer to the end destination. When you need faster computing, storage & databases in populated areas that are outside of an AWS Region.

Cost & Capacity Management

Cost Management - How do we save money?
Capacity Management - How do we meet the demand of traffic & usages through adding or upgrading servers?

EC2 Spot Instances, Reserved Instanced & Savings Plan

Ways to save on computing, by paying up in full or partially, by committing to yearly contracts or by being flexible about availability & interruption to computing service.

AWS Batch

Plans, Schedules & executes your batch computing workloads across the full range of AWS compute services & can utilise Spot Instance to save money.

AWS Compute Optimizer

Suggests how to reduce costs & improve performance by using machine learning to analyse previous usage history.

EC2 Autoscaling Groups (ASGs)

Automatically adds or removes EC2 Servers to meet the current traffic demand of traffic. Will save you money & meet capacity since you only run the amount of servers you need.

Elastic Load Balancer (ELB)

Distributes traffic to multiple instances & can re-route traffic from unhealthy instances to healthy instances. Can route traffic to EC2 instances running in different AZs.

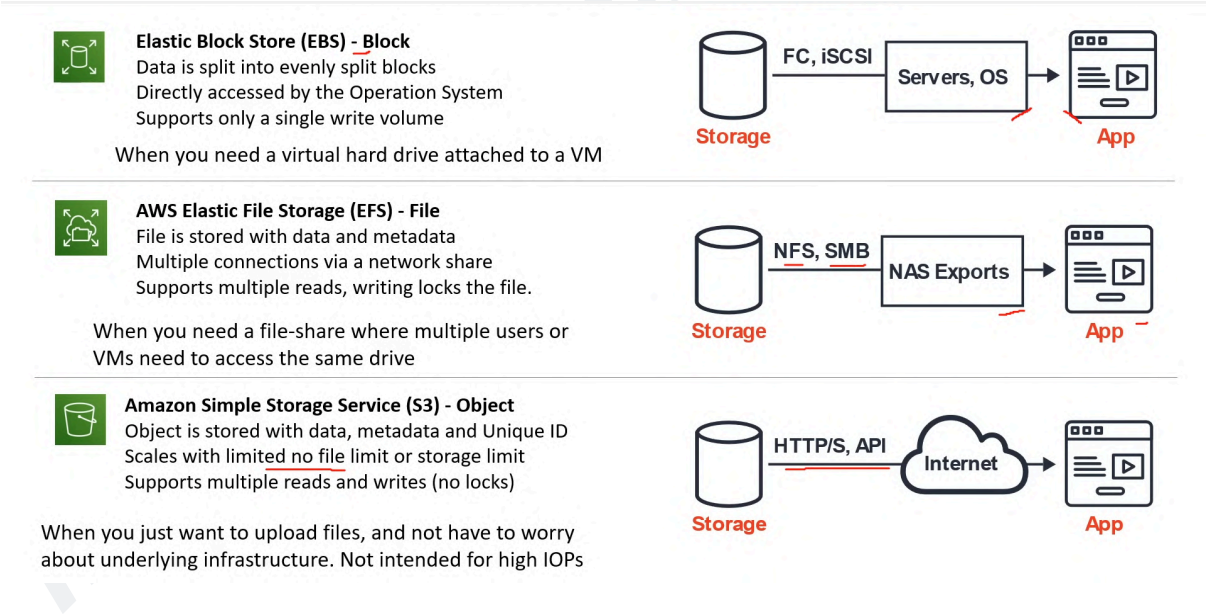
AWS Elastic Beanstalk (EB)

It is for easily deploying web-applications without developers having to worry about setting up & understanding the underlying AWS Services. Similar to Heroku.

Storage Services

Types of Storage Services

1. **Elastic Block Store (EBS) - Block : When you need a virtual hard drive attached to a VM.**
    - Data is split evenly into blocks.
    - Directly accessed by the operating system.
    - Supports only a single write volume
2. **AWS Elastic File Storage - File : When you need a file-share where multiple users or VMs need to access the same drive.**
    - File is stored with data & metadata.
    - Multiple connections via a network drive.
    - Supports multiple reads, writing locks the file
3. **Amazon Simple Storage Service (S3) - Object : When you just want to upload files, & not have to worry about the underlying infrastructure. Not intended for high IOPs.**
    - Object is stored with data, metadata & Unique ID.
    - Scales with no file limit or storage limit.
    - Supports multiple reads & writes (No locks).



Introduction to S3 - Simple Storage Service

What is Object Storage (Object - Based Storage) ? : Data storage architecture that manages data as objects, as opposed to other storage architectures.

- File systems which manage data as files & file hierarchy.
- Block storage which manages data as blocks within sectors & tracks.

S3 : Simple Storage Service

- S3 provides unlimited storage.
- You don't need to think about the underlying infrastructure.
- The S3 Console provides an interface for you to upload & access your data.

S3 Object : Objects contain your data. They are like files.

- **Key** the name of the object.
- **Value** the data itself made up of a sequence of bytes.
- **Version ID** when versioning enabled, the version of object.
- **MetaData** additional information attached to the object.

S3 Bucket : Buckets hold objects. Buckets can also have folders which in turn hold objects.

- S3 is a universal namespace so bucket names must be unique. (like domain names)

You can store an individual object from 0 Bytes to 5 Terabytes in size

S3 Storage Classes

AWS offers a range of S3 storage classes that trade **Retrieval Time, Accessibility & Durability** for Cheaper Storage.**Getting Cheaper down the line.**

1. **S3 Standard (default) :** Fast! 99.99% Availability, 11 9's Durability, Replicated across at least three AZs
2. **S3 Intelligent Tiering :** Uses ML to analyse objects & determine the appropriate storage class.
3. **S3 Standard-IA (Infrequent Access) :** Still Fast! Cheaper if you access files less than once a month. Additional retrieval fee is applied. 50% less than standard (Reduced Availability)
4. **S3 One-Zone-IA :** Still Fast! Objects only exist in one AZ. Availability (is 99.5%) but cheaper than standard IA by 20% less (Reduced durability) Data could get destroyed. A retrieval fee is applied.
5. **S3 Glacier :** For long term cold storage.Retrieval of data can take minutes to hours but the off is very cheap storage.
6. **S3 Glacier Deep Archive :** The lowest cost storage class. Data retrieval time is 12 hours.
7. **S3 Outposts has its own storage class.**

AWS Snow Family

AWS Snow Family are **storage & compute devices used to physically move data in or out the cloud** when moving data over the internet or private connection it to slow, difficult or costly.

All data is delivered to Amazon S3

1. **Snowcone :**
    a. Two sizes:
        i. 8 TB of Storage (HHD)
        ii. 14 TB of Storage (SSD)
2. **Snowball Edge :**
    a. Comes generally in two types:
        i. Storage Optimised.
            1. 80 TB
        ii. Compute Optimised.
            1. 39.5 TB
3. **Snowmobile :**
    a. 100 PB of storage.

Storage Services

1. **Simple Storage Service (S3)** is a serverless object storage service. You can upload very large files & an unlimited amount of files. You pay for what you store. You don't worry about the underlying file-system or upgrading the disk size.
2. **S3 Glacier** is a cold storage service. It is designed as a low cost storage solution for archiving & long-term backup. It uses previous generation HDD drives to get that low cost. It's highly secure & durable.
3. **Elastic Block Store (EBS)** is a persistent block storage service. It is a virtual hard drive in the cloud you could attach to EC2 instances. You can choose different kinds of hard drives : SSD, IOPS SSD, Throughput HDD, Cold HDD.
4. **Elastic File Storage (EFS)** is a cloud-native NFS file system service. File storage you can mount to multiple EC2 instances at the same time. When you need to share files between multiple servers.
5. **Storage Gateway** is a hybrid cloud storage service that extends your on-premise storage to cloud.
    a. **File Gateway** extends your local storage to AWS S3
    b. **Volume Gateway** caches your local drives to S3 so you have a continuous backup on local files in the cloud.
    c. **Tape Gateway** stores files onto virtual tapes for backing up your files on very cost effective long-term storage.

6. **AWS Snow Family** are storage devices used to physically migrate large amounts of data to the cloud.
    a. **Snowball Edge** is a briefcase size data storage device. **50-80 TB**
    b. **Snowmobile** is a cargo container filled with racks of storage & compute that is transported via sem-trailer tractor truck to transfer up to **100PB** of data per trailer.
    c. **Snowcone** is a very small version of Snowball that can transfer **8TB** of data.
7. **AWS Backup** is a fully managed backup service that makes is easy to centralise & automate the backup of data across multiple AWS services eg. EC2, EBS, DynamoDB, EFS, Storage Gateway according to the user created backup plans.
8. **CloudEndure Disaster Recovery** continuously replicates your machines into a low-cost staging area in your target AWS .
9. **Amazon FSx** is a feature rich & highly-performant file system. That can be used for Windows (SWB) or Linux (Lustre).
    a. **Amazon FSx for Window File Server** uses the SMB protocol & allows you to mount FSx to Windows Servers.
    b. **Amazon FSx for Lustre uses Linux's Lustre file** system & allows you to mount FSx to Linux servers.

Databases

What is a Database?

- A database is a data-store that stores semi-structured & structured data.
- A database is more complex data stores because it requires using formal design & modelling techniques.

Databases can be generally categorised as either:

- Relational Databases
    - Structured data that strongly represents tabular data (tables, rows & columns)
        - Row-oriented or Columnar-oriented.
- Non-relational Databases
    - Semi-structured that may or may not distantly resemble tabular data.

Databases have a rich set of functionality :

- Specialised language to query (retrieve data)
- Specialised modelling strategies to optimise retrieval for different use cases
- More fine tune control over the transformation of the data into useful data structures or reports.

Normally a database infers someone is using a relational row oriented data store.

What is a Data Warehouse?

A relational datastore designed for analytics workloads, which is generally column-oriented data-store.
Companies will have TBs & millions of rows of data & they need a fast way to be able to produce analytics reports.

Data warehouses generally perform Aggregation

- Aggregation is grouping data eg. find a total or average.
- Data warehouses are optimised around columns since they need to quickly aggregate column data.

Data warehouses generally be HOT

- Hot means they can return queries very fast even though have vast amounts of data

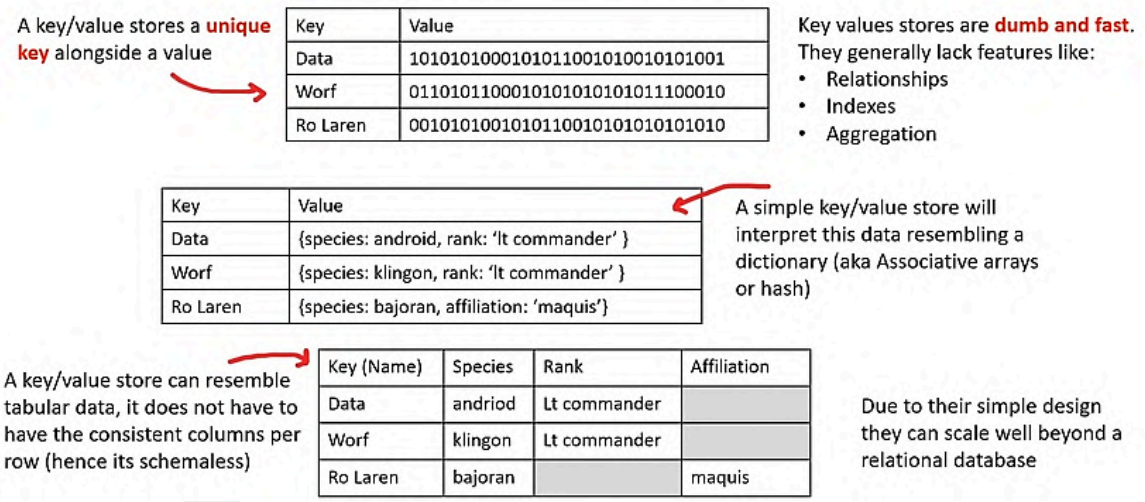
- Data warehouses are infrequently accessed meaning they aren't intended for real-time reporting but maybe once or twice a day or once a week to generate business & user reports.

A data warehouse needs to consume data from relational databases on a regular basis.

What is a Key / Value store?

A key-value database is a type of non-relational database (NoSQL) that uses a simple key-value method to store data.

- Stores a unique key alongside a value.
- Dumb & fast
- Lack features like :
    - Relationships
    - Indexes
    - Aggregation
- Due to their simple design they can scale well beyond a relational database.



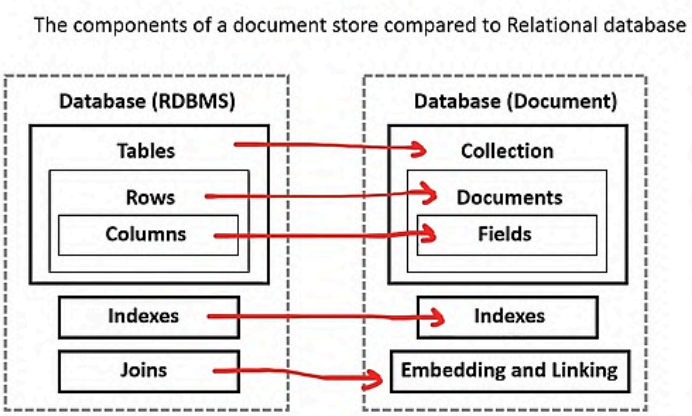
Key	Value
Data	10101010001010110010100101010 01
Worf	01101011000101010101011100010
Ro Laren	00101010010101100101010101010

Key	Value
Data	{species: android, rank: 'lt commander' }
Worf	{species: klingon, rank: 'lt commander' }
Ro Laren	{species: bajoran, affiliation: 'maquis'}

Key (Name)	Species	Rank	Affiliation
Data	andriod	Lt commander	
Worf	klingon	Lt commander	
Ro Laren	bajoran		maquis

What is a Document Store?

- A document store is a NoSQL database that stores documents as its primary data structure.
- A document is an XML but more commonly is JSON or JSON-Like.
- Document stores are sub-class of Key/Value stores.



Components of document store compared to a Relational Database

NoSQL Database Service

DynamoDB is a serverless NoSQL key/value & document database. It is designed to scale to billions of records with guaranteed consistent data return in at least a second. You don't have to worry about managing shards!

- **DynamoDB is AWS's flagship database service** meaning whenever we think of a database service that just scales, is cost effective & very fast we should think of DynamoDB.

When we want a massively scalable database - DynamoDB

In **2019, Amazon** online shopping retail shutdown their last Oracle database & completed their migration to DynamoDB. They had 7500 Oracle Database & 75 Petabytes of data. With DynamoDB they reduce costs by 60% & reduce latency by 40%.

DocumentDB is a NoSQL document database that is "MongoDB compatible"

MongoDB is a very popular NoSQL among developers. There were open-source licensing issues around using open-sources MongoDB, so AWS got around it by just building their own MongoDB database.

When you want a MongoDB Database. - DocumentDB

Amazon Keyspaces is a fully managed Apache Cassandra database. Cassandra is an open-source NoSQL key/value database similar to DynamoDB in that it is a columnar store database but has some additional functionality.

When you want to use Apache Cassandra. - Amazon Keyspaces

Relational Database Services

Relational Database Service (RDS) is a **relational database service** that supports SQL engines. Relational is synonymous with SQL & Online Transactional Processing (OLTP). Relational databases are the most commonly used type of database among tech companies & start-ups.

RDS Supports the following SQL Engines :

- **MySQL -** The most popular open-source SQL Database that was purchased & now owned by Oracle.
- **MariaDB** - When Oracle bought MySQL, MariaDB made a fork of MySQL under a different open-source licence.
- **Postgres (PSQL)** - Most popular open-source SQL database among developers. Has rich features over MySQL but at added complexity.
- **Oracle** - Oracle's proprietary SQL database. Well used by Enterprise companies. You have to buy a licence to use it.
- **Microsoft SQL Server** - Microsoft's proprietary SQL database. You have to buy a licence to use it.
- **Aurora -** Fully managed database

Aurora is a **fully managed** database of either MySQL(5x faster) & PSQL(3x faster) databases.

When you want a highly available, durable, scalable & secure relational database for Postgres or MySQL.

Aurora Serverless is the **serverless on-demand version of Aurora**.

When you want "most" of the benefits of Aurora but can trade to have cold-starts or you don't have lots of traffic data.

RDS on VMWare allows you to deploy RDS Supported engines to an on-premise data-centre. The data centre must be using VMWare for server virtualisation.

When you want databases managed by RDS on your own Database.

Other Database Services

Redshift is a **petabyte-size** data-warehouse. Data-warehouses are for Online Analytical Processing(OLAP)

- Data warehouses can be expensive because they're keeping data 'hot'.

When you need to quickly generate analytics or reports from a large amount of data.

ElastiCache is a managed database of the in-memory & caching open-source databases Redis & Memcached.

When you need to improve the performance of an application by adding a caching layer in-front of a web-server or database.

Neptune is a managed graph database. Data is represented as interconnected nodes.

When you need to understand the connections between data eg. Mapping fraud rings or social media relationships.

Amazon Timestreams is a fully managed time series database. Think of devices that send lots of data that are time-sensitive such as IoT devices.

When you need to measure how things change over time.

Amazon Quantum Ledger Database is a fully managed ledger database that provides transparent, immutable & cryptographically variable transaction logs.

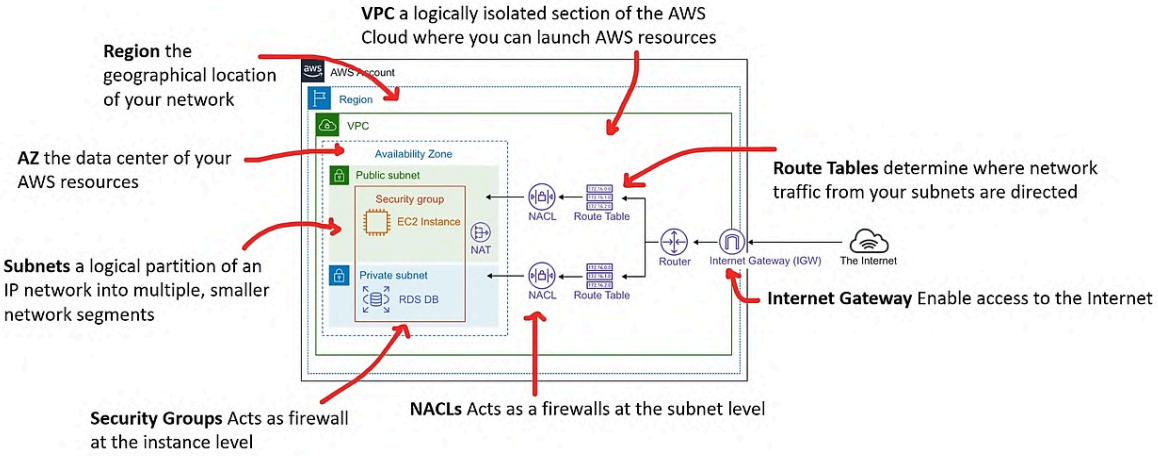
When you need to record the history of financial activities that can be trusted.

Database Migration Service (DMS) is a database migration service.

- You can migrate from :
    - On-premise database to AWS.
    - From two databases in different or same AWS Accounts using different SQL engines.
    - From an SQL to NoSQL database.

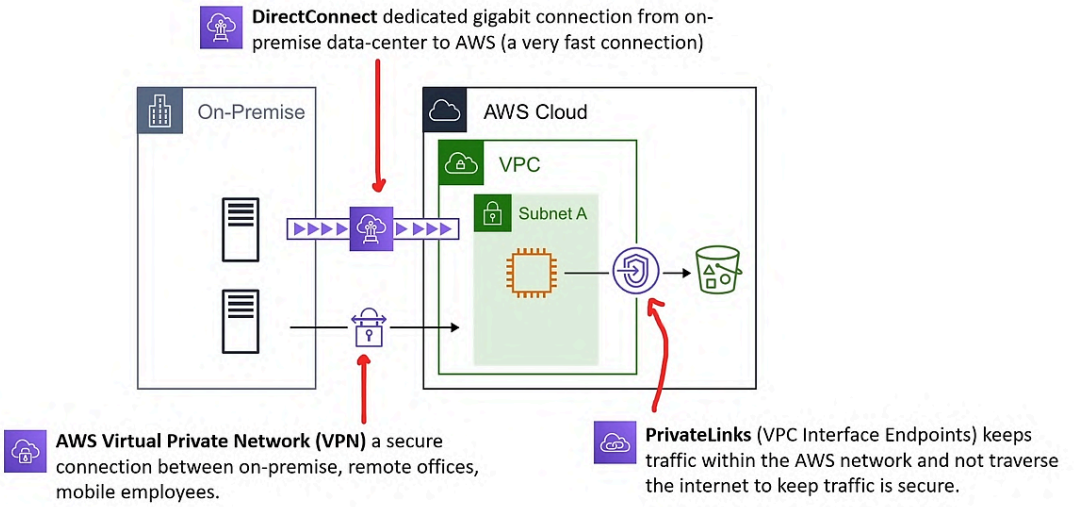
Networking

Cloud-Native Networking Services



Enterprise / Hybrid Networking Services

- **AWS Virtual Private Network (VPN)** a secure connection between on-premise, remote offices, mobile employees.
- **DirectConnect** dedicated gigabit connection from on-premise data-centre to AWS (a very fast connection)
- **PrivateLinks (VPC Interface Endpoints)** keeps traffic within the AWS network & not traverses the internet to keep traffic secure.



Virtual Private Cloud (VPC) & Subnets

Virtual Private Cloud (VPC) is a logically isolated section of the AWS Network where you launch your AWS resources. You choose a range of IPs using CIDR Range.

CIDR - Classless Inter-Domain Routing : Classless Inter-Domain Routing (CIDR) allows network routers to route data packets to the respective device based on the indicated subnet. Instead of classifying the IP address based on classes, routers retrieve the network & host address as specified by the CIDR suffix.

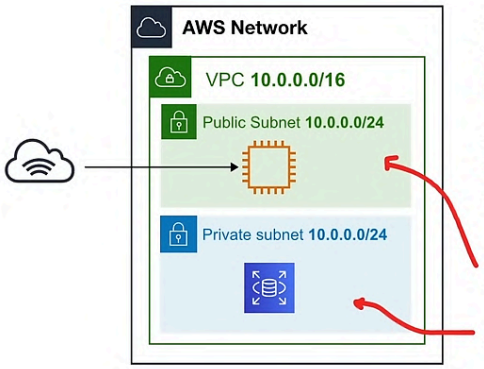
CIDR Range of 10.0.0.0/16 = 65,536 IP Addresses

Subnets is a logical partition of an IP network into multiple smaller network segments. **Breaking up your IP range** for VPC into smaller networks.

Subnets need to have a smaller CIDR Range than the VPC to represent their portion.

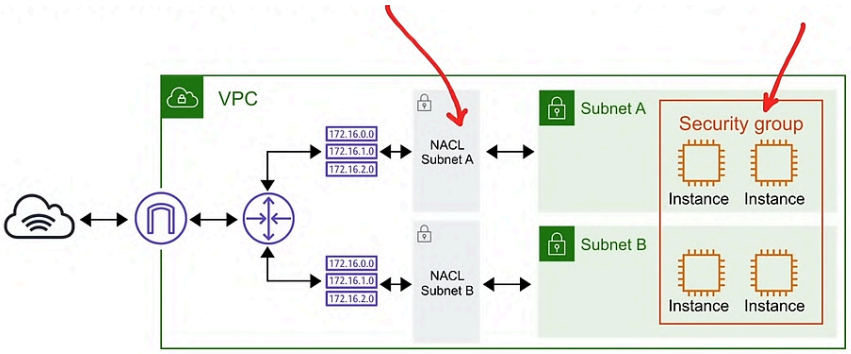
Eg : Subnet CIDR Range 10.0.0.0/24 = 256 IP Addresses

Public Subnet is one that can reach the internet.
Private Subnet is one that cannot reach the internet.



Security Groups VS NACLs

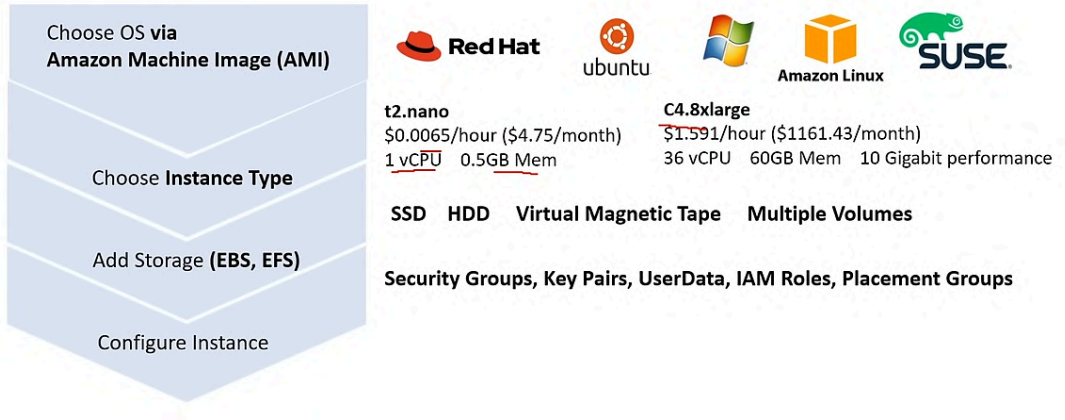
- **Network Access Control Lists (NACLs) :** It acts as a **virtual firewall at the subnet level. You create Allow & Deny rules.**
  - Eg : Block a specific IP address known for abuse.
- **Security Groups** : It acts as a **virtual firewall at the instance level.** Implicitly denies all traffic. **You create only Allow rules.**
  - Eg : Allow an EC2 instance access on port 22 for SSH.
  - Cannot block single IP Address



Elastic Compute Cloud (EC2)

Introduction to EC2

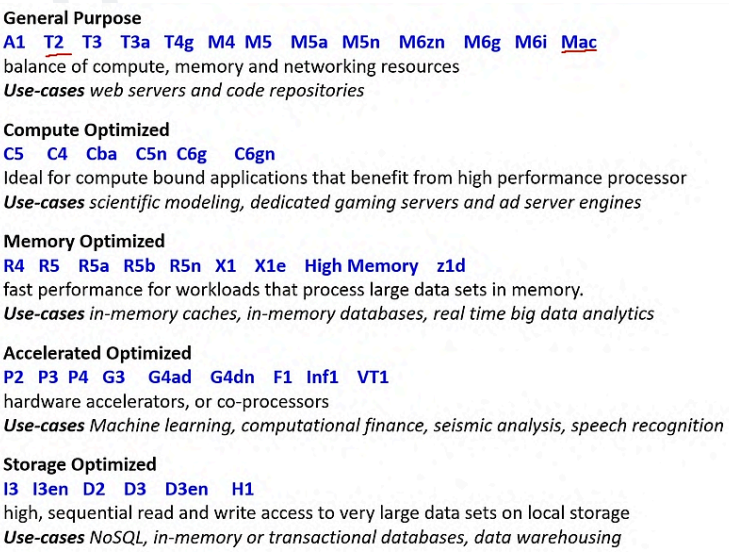
Elastic Compute Cloud (EC2) is a **highly configurable virtual server.** EC2 has a resizable compute capacity. It takes **minutes** to launch new instances. Anything & everything on AWS uses EC2 Instances underneath



EC2 Instance Families

What are Instance Families?

- Instance families are different combinations of CPU, Memory, Storage & Networking capacity.
- Instance Families allow you to choose the appropriate combination of capacity to meet your application's unique requirements.
- Different instance families are different because of the varying hardware used to give them their unique properties.

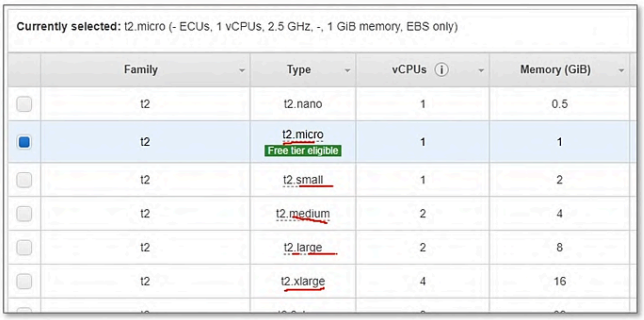


Various Instance Families

EC2 Instance Types

An instance type is a particular instance size & instance family.
Common Pattern for Instance Sizes :

- Nano
- Micro
- Small
- Medium
- Large
- xLarge
- 2xLarge
- 4xLarge
- 8xLarge
-
- Exceptions to this pattern :
    - c6g.metal - is a bare metal machine.
    - C5.9xLarge is not a power of 2 or even number size.



EC2 instance sizes **generally double** in price & key attributes.

Name	vCPU	RAM (GIB)	On-Demand **per hour**	On-Demand **per month**
t2.small	1	12	$0.023	$16.79
t2.medium	2	24	$0.0464	$33.87
t2.large	2	36	$0.0928	$67.74
t2.xlarge	4	54	$0.1856	$135.48

EC2 - Dedicated Instances VS Dedicated Hosts

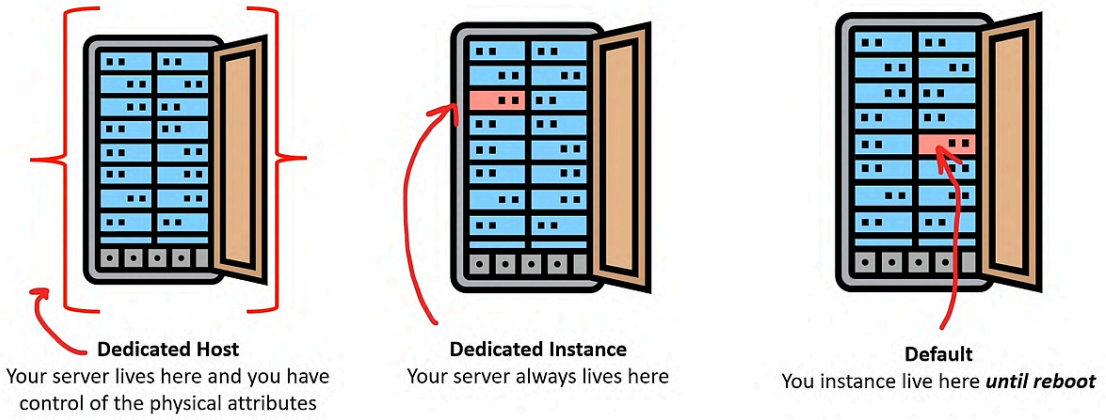
Dedicated Hosts are **single-tenant EC2 instances** designed to let you **Bring-Your-Own-License (BYOL)** based on machine characteristics.

Aspect	Dedicated Instance	Dedicated Hosts
Isolation	Instance isolation	Physical server isolation
Billing	Per instance billing (+2$ per region fee)	Per host billing
Visibility of Physical Characteristics	No visibilities	Sockets, cores, Host ID
Affinity between a host & instance	No affinity	Consistency deploy to the same instances to the same physical server
Targeted Instance placement	No control	Additional Control
Automatic Instance placement	Yes	Yes
Add capacity using an allocation request	No	Yes

EC2 Tenancy

EC2 has 3 levels of Tenancy

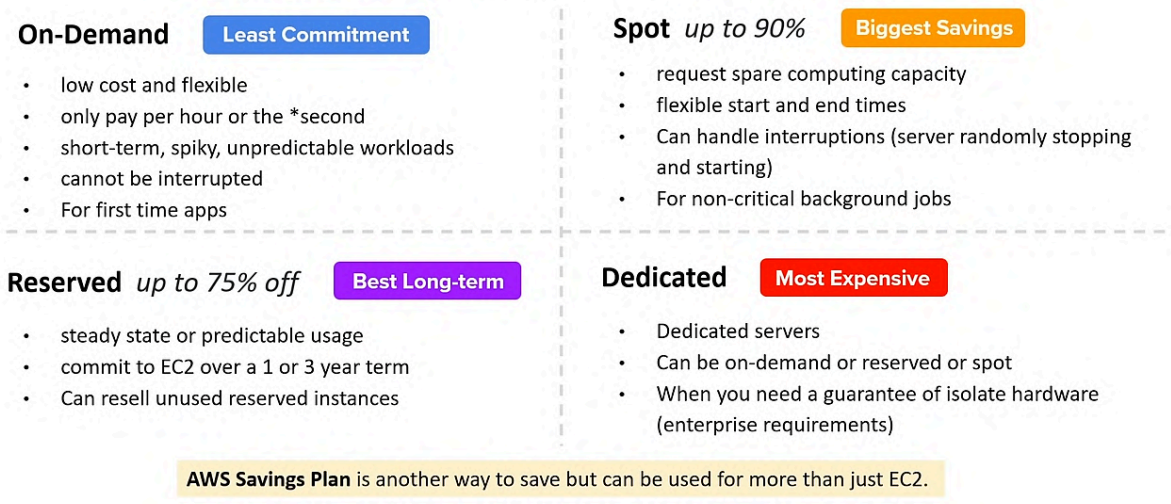
1. **Dedicated Host :** Your server lives here & you have control of the physical attributes.
2. **Dedicated Instance** : Your server always lives here
3. **Default :** Your instance lives here **until reboot**.



EC2 Pricing Models

There are 5 different ways to pay for EC2 (Virtual Machines)

1. **On-Demand** : Least Commitment
    - Low cost & flexible
    - Only pay per hour or per second
    - Short-term, spiky, unpredictable workloads
    - Cannot be interrupted
    - For first time apps
2. **Reserved** : Best Long-term
    - Steady state or predictable usage
    - Commit to EC2 over a 1 or 3 year term
    - Can resell unused reserved instances
3. **Spot** : Biggest Savings
    - Request spare computing capacity
    - Flexible start & end times
    - Can handle interruptions (Server randomly stopping & starting)
    - For non-critical background jobs
4. **Dedicated** : Most Expensive
    - Dedicated Servers
    - Can be on-demand or reserved or spot
    - When you need a guarantee of isolated hardware (entreprise requirements)
5. **AWS Savings Plan** is another way to save but can be used for more than just EC2.



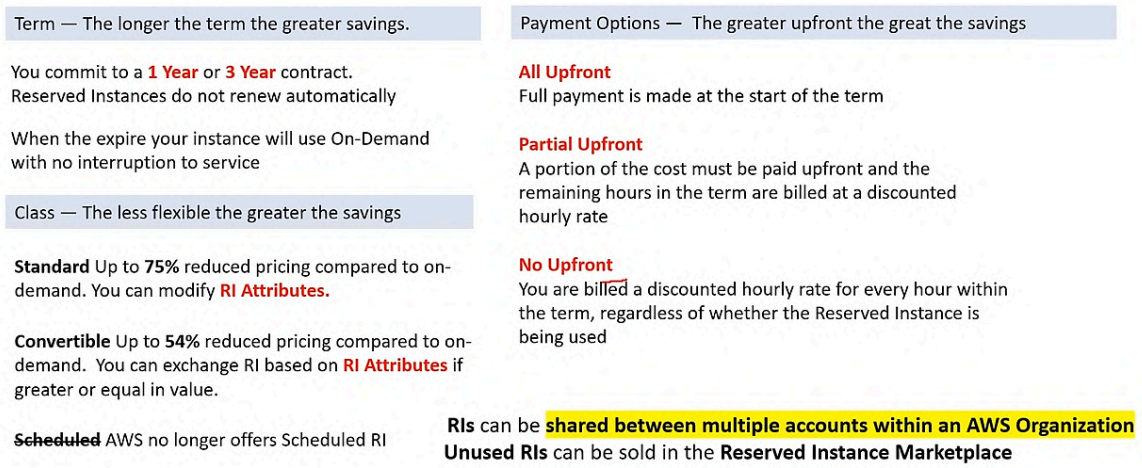
On-Demand Instances

On-Demand is a **Pay-As-You-Go (PAYG) model**, where you consume the compute & then you pay.

- When you launch an EC2 instance it is **by default** using **On-Demand** pricing.
- On-Demand has no up-front payment & no long-term commitment.
- You are charged by the **second (min. of 60 seconds)** or the **hour**.
    - /Second for : Linux, Windows with SQL Enterprise/Standard/Web Instances.(Does not have a separate hourly charge)
    - /Hour for : For all other instance types.
- When looking up pricing, it'll always show EC2 Pricing on an hourly basis.
- **On-Demand** is for applications where the workload is short-term, spikey or unpredictable.
    - Example : When you have a **new app** for development or you want to run an experiment.

Reserved Instances (RI)

- Designed for applications that have a **steady-state**, **predictable usage**, or **require reserved capacity**.
- Reduced Pricing is based on **Term** x **Class Offering** x **RI Attributes** x **Payment Option**
    - **Term :** The longer the term, the greater the savings
    - **Class :** the less flexible the class, the greater the savings
    - **Payment Options :** The greater the upfront payment, the greater the savings.
- **RIs** can be shared between multiple accounts within an AWS Organisation.
- **Unused RIs** can be sold in the **Reserved Instances Marketplace**.



Reserved Instances (RI) Attributes

Reserved Instances Attributes / Instance Attributes are limited based on Class Offering & can affect the final price of an RI instance.

There are 4 Instance Attributes.

1. **Instance Type** : Includes Instance family & instance size.
2. **Region** : The region in which Reserved Instance is purchased
3. **Tenancy** : Whatever your instance runs on shared (default) or single-tenant (dedicated) hardware.
4. **Platform** : The operating system eg. Windows or Linux/Unix.

Regional & Zonal Reserved Instances

When you purchase a Reserved Instance, you determine **the scope** of the Reserved Instance. It doesn't affect the price.

Regional Reserved Instance	Zonal Reserved Instance
Does not reserve capacity	Reserves capacity in the specific Availability Zone
Reserved Instance discount applied to instance usage in any Availability Zone in the region	Reserved Instance discount applies to instance in the selected Availability Zone (No AZ Flexibility)
Reserved Instance discount applied to instance usage within the instance family, regardless of size. Only supported on Amazon Linux/Unix Reserved Instances with default tenancy.	No instance size flexibility. Reserved Instance discounts discount applies to instance usage for the specified instance type & size only.
You can queue purchases	You cannot queue purchases

Reserved Instances Limits

There is a limit to the number or Reserved Instances that you can purchase per month.

20 Regional Reserved Instances / Region / Month.
20 Zonal Reserved Instances / Availability Zone / Month.

Regional Limits : You cannot exceed your running On-Demand instance limit by purchasing regional Reserved Instances. Default On-Demand instance limit is 20
Before purchasing a Reserved Instance ensure your On-Demand limit is equal to or greater than your Reserved Instance you intend to purchase.

Zonal Limits : You can exceed your running On-Demand instance limit by purchasing Zonal Reserved Instances.
If you already have 20 running On-Demand instances & you purchase 20 Zonal Reserved Instances, you can launch further 20 On-Demand Instances that match the specifications of your Zonal Reserved Instances.

Capacity Reservations

EC2 Instances are backed by different kinds of hardware & so there is a **finite amount of servers** available within an Availability Zone per instance type or family.

Capacity Reservation is a service of EC2 that allows you to **request a reserve of EC2 instance type** for a specific region & AZ.
The reserved capacity is charged at the selected instance type's On-Demand rate whether an instance is running in it or not.
You can also use your regional reserved instances with your Capacity Reservations to benefit from billing discounts.

Standard VS Convertible Reserved Instances

Aspect	Standard Reserved Instance	Convertible Reserved Instance
RI attributes that can be modified :	RI attributes that can be modified. Such as :	RI attributes can't be modified (you perform an exchange)
	1. Change the AZ within the same region	
	2. Change the Scope of the Zonal RI to Regional RI or visa versa	
	3. Change the Instance size (Linux/Unix only, default tenancy)	
	4. Change network from EC2-Classic to VPC & vice-versa	
Exchange?	Can't be exchanged	Can be exchanged during the term for another Convertible RI with new RI Attributes.
		1. Instance Family
		2. Instance Type
		3. Platform
		4. Scope
		5. Tenancy
Marketplace	Can be bought or sold in the RI Marketplace	Can't be bought or sold in the RI Marketplace

Reserved Instances Marketplace

EC2 Reserved Instance Marketplace allows you to **sell your unused Standard RI** to recoup your RI & spend for your RI that you do not intend or cannot use.

- Reserved Instances can be sold after they have been active for at least 30 days & once AWS has received the upfront payment (if applicable).
- You must have a US bank account to sell Reserved Instances on the Reserved Instance Marketplace.
- There must be at least one month remaining in the term of the Reserved Instance you are listing.

- You will retain the pricing & capacity benefit of your reservation until it's sold & the transaction is complete. Your company name (& address upon request) will be shared with the buyer for tax purposes.
- A seller can set only the upfront price for a Reserved Instance. The usage price & other configuration (e.g., instance type, Availability Zone, platform) will remain the same as when the Reserved Instance was initially purchased.
- The term length will be rounded down to the nearest month. For example, a reservation with 9 months & 15 days remaining will appear as 9 months on the Reserved Instance Marketplace.
- You can sell up to $20,000 in Reserved Instances per year. If you need to sell more Reserved Instances.
- Reserved Instances in the GovCloud region cannot be sold on the Reserved Instance Marketplace.

Spot Instances

AWS has **unused compute capacity** that they want to maximise the utility of their idle servers.

Spot instances provide a discount of **90%** compared to On-Demand Pricing.
Spot instances can be terminated if the computing capacity is needed by other On-Demand Customers.

Designed for applications that have flexible start & end times or applications that are only feasible at **very low** compute costs.

AWS Batch is an easy & convenient way to use Spot Pricing

Termination Conditions

- Instances can be terminated by AWS at any time.
- If your instance is **terminated by AWS**, you don't get charged for a partial hour of usage.
- If **you terminate** an instance, you will still be charged for any hour that it runs.

Dedicated Instances

Dedicated Instances is designed to meet regulatory requirements.
*When you have **strict server-bound licensing** that won't support multi-tenancy or cloud deployments you use Dedicated Hosts.*

Multi-Tenant : When multiple customers are running workloads on the same hardware. Virtual Isolation is what separates the customers.

Single Tenant : When a single customer has dedicated hardware Physical Isolation is what separates the customers.

Dedicated Server can be offered for

- **On-Demand**
- **Reserved**
- **Spot**

Enterprises & **Large Organisations** may have security concerns or obligations against sharing the same hardware with other AWS Customers.

AWS Savings Plan

Savings Plans offer you the similar discounts as Reserved Instances (RIs) but simplifies the purchasing process.

There are **3 types of Savings Plans**

- **Compute Savings Plans :** Compute Savings Plans provide the most flexibility & help to reduce your costs by up to 66%. These plans automatically apply to EC2 instance usage, AWS Fargate, & AWS Lambda service usage regardless of instance family, size, AZ, region, OS, or tenancy.

- **EC2 Savings Plan :** Provide the lowest prices, offering savings up to 72% in exchange for commitment to usage of individual instance families in a region. Automatically reduces your cost on the selected instance family in that region regardless of AZ, size, OS or tenancy. give you the flexibility to change your usage between instances within a family in that region.

- **SageMaker Savings Plan :** Helps you reduce SageMaker costs by up to 64%. Automatically apply to SageMaker usage regardless of instance family, size, component or AWS region.

You can choose from **two Terms**

- **1 Year**
- **3 Years**

You can choose from **three Payment Options**

- **All Upfront**
- **Partial Upfront**
- **No Upfront**

Identity

Zero-Trust Model

Operates on the principle of **"Trust no one, verify everything."**

Malicious actors being able to by-pass conventional access controls demonstrates traditional security measures are no longer sufficient.
In the Zero Trust Model, **Identity** becomes the primary security perimeter.

What is the Primary Security Perimeter?
The primary or new security perimeter defines the first line of defence & its security controls that protect a company's cloud resources & assets

Network-Centric (Old Way) : Traditional security focused on firewalls & VPNs since there were few employees or workstations outside the office or they were in the specific remote offices.
Identity-Centric (New Way) : Bring-your-own-device, remote workstations is much more common, we can't trust if the employee is in a secure location, we have identity based security controls like MFA or providing provisional access based on the level of risk from where, when & what a user wants to access.

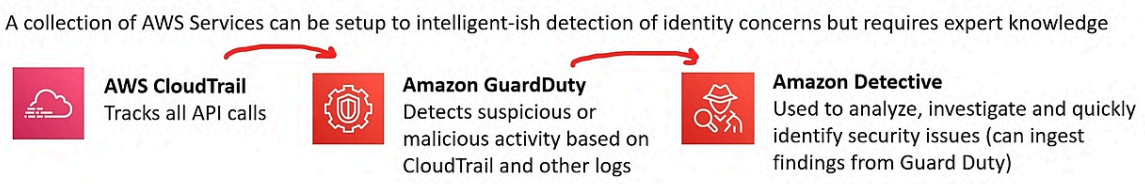
Identity-Centric does not replace but **augments** Network-Centric Security.

Zero Trust on AWS

Identity Security Controls you can implement on AWS to meet the Zero Trust Model.
AWS Identity & Access Management (IAM)

- IAM Policies
- Permission Boundaries
- Service Control Policies (Organisation wide policies)
- IAM Policy Conditions
    - aws:SourceIp - Restrict on IP Address
    - aws:RequestedRegion - Restrict on Region
    - aws:MultiFactorAuthPresent - Restrict if MFA is turned off
    - aws:CurrentTime - Restrict access based on time of day

AWS does not have ready-to-use identity controls that are intelligent, which is why AWS is considered to not have a true Zero Trust offering for customers & third-party services need to be used.



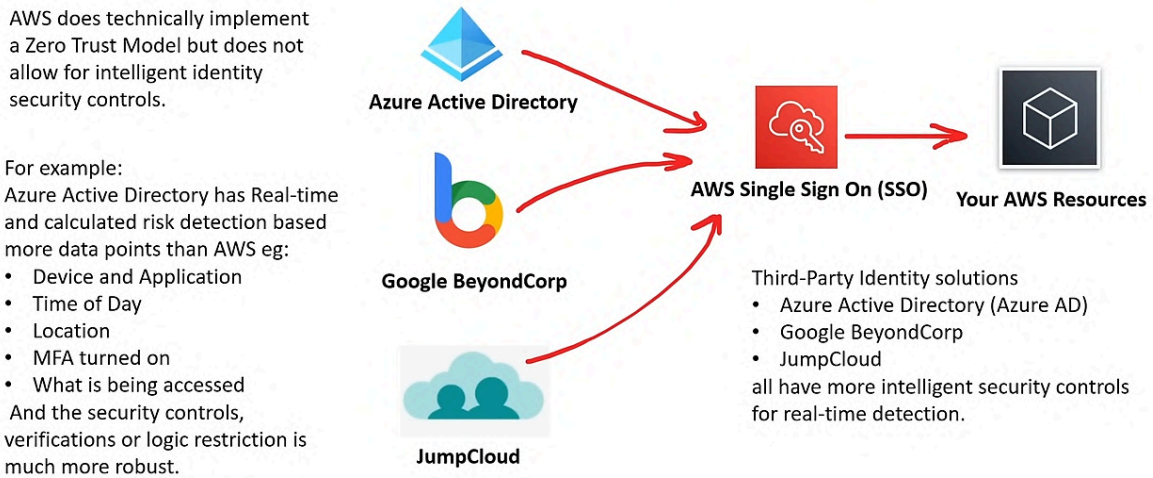
Zero Trust on AWS with Third Parties

AWS does technically implement a Zero Trust Model but does not allow for intelligent identity security controls.

Third Party Identity Solutions :

- Azure Active Directory (Azure AD)
- Google BeyondCorp
- JumpCloud

All of these have more intelligent security controls for real-time detection.



Directory Service

What is a directory service?

A directory service maps the **names of network resources to their network addresses.**

A directory service is shared information infrastructure for locating, managing, administering & organising resources such as

- Volumes
- Folders
- Files

- Printers
- Users
- Groups
- Devices
- Telephone numbers
- Other objects

A directory service is a critical component of a network operating system
A directory server (name server) is a server which provides a directory service.

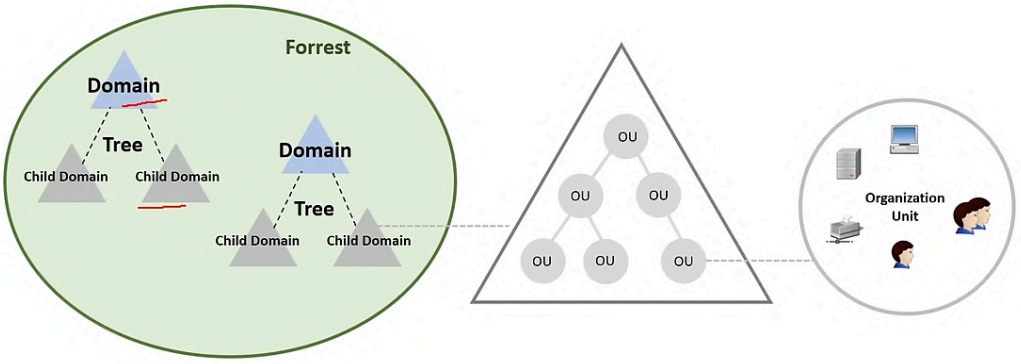
Each resource on the network is considered an object by the directory server. Information about a particular resource is stored as a collection of attributes associated with that resource or object.

Well known directory services

- Domain Name Service (DNS)
    - Directory service for the internet
- **Microsoft Active Directory**
    - Azure Active Directory
- Apache Directory Server
- Oracle Internet Directory (OID)
- OpenLDAP
- Cloud Identity
- JumpCloud

Active Directory

Microsoft introduced **Active Directory** Domain Services in Windows 2000 to give organisations the ability to manage multiple on-premises infrastructure components & systems using a single identity per user.



Identity Providers (IdP)

Identity Provider (IdP) a system entity that creates, maintains & manages identity information for principals & also provides authentication services to applications within a **federation** or distributed network.
A trusted provider of your user identity that lets you authenticate to access other services.
Identity Providers could be : Facebook, Amazon, Google, Twitter, Github, LinkedIn...

Federated identity is a method of linking a user's identity across multiple separate identity management systems.

OpenID : Open Standard & Decentralised Authentication Protocol.

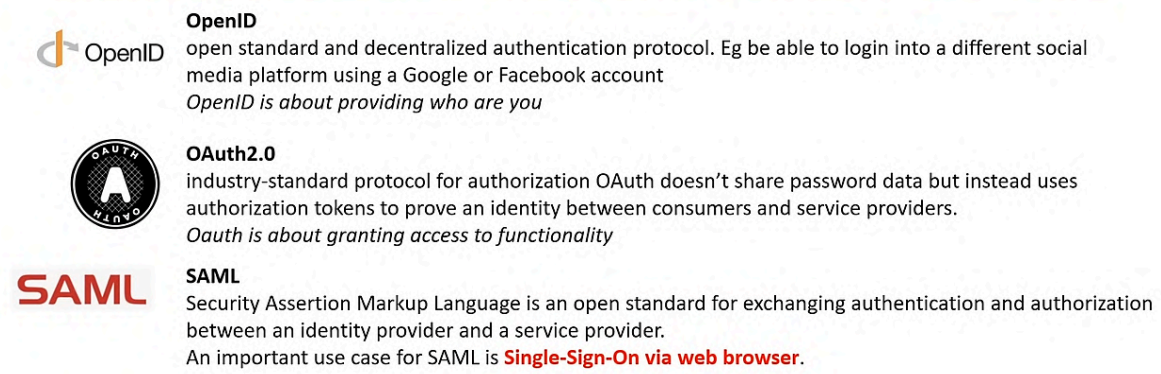
OpenID is about providing who you are.

OAuth2.0 : Industry-standard Protocol for authentication.

Oauth is about granting access to functionality.

SAML : Security Assertion Markup Language

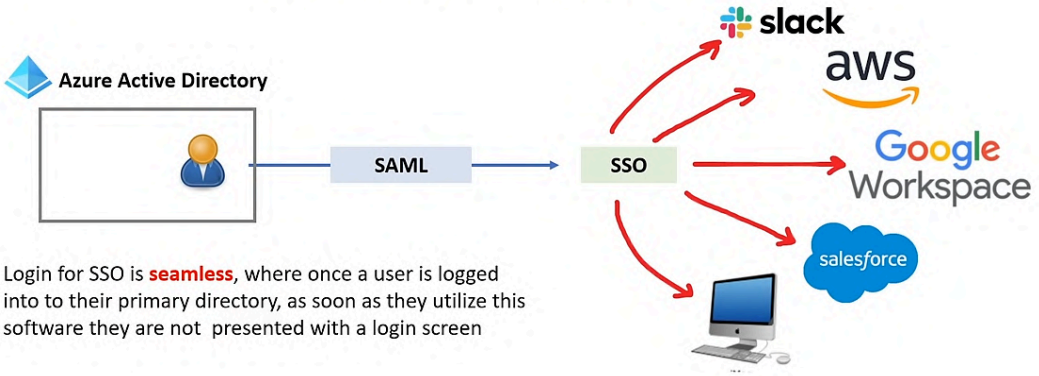
SAML is for Single-Sign-On via Web browser.



Single-Sign-On (SSO)

Single-Sign-On (SSO) is an authentication scheme that **allows a user to log in with a single ID & password to different systems & softwares.**
SSO Allows IT departments to administer a single identity that can access many machines & cloud services.

Login for SSO is seamless, where once a user is logged into their primary directory, as soon as they utilise their software they are not presented with a login screen.j



LDAP - Lightweight Directory Access Protocol

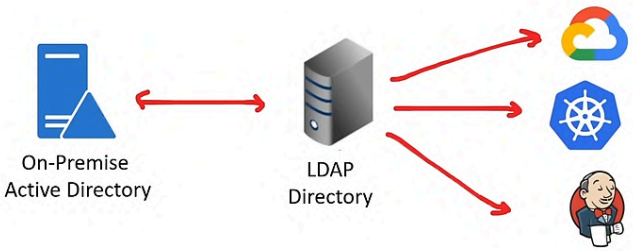
LDAP - Lightweight Directory Access Protocol is an open, vendor-neutral, industry standard application protocol for accessing & maintaining distributed directory information services over an Internet Protocol (IP) Network.

A common use of LDAP is to provide a central place to store usernames & passwords

LDAP enables for same-sign which allocates users to a single ID & password but they have to enter it in every time they want to login.

Why use LDAP when SSO is more convenient?

- Most SSO Systems use LDAP
- LDAP was not designed natively to work with web-applications
- Some systems only support integration with LDAP & not SSO



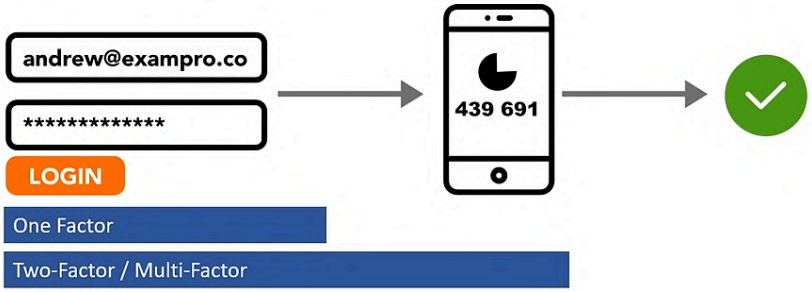
Multi-Factor Authentication (MFA)

What is Multi-Factor Authentication (MFA)?

A security control where after you fill in your username/email & password you have to use a second device such as a phone to confirm that it's you logging in.

MFA protects against people who have stolen your password

MFA is an option in most cloud providers & even social media websites such as Facebook.

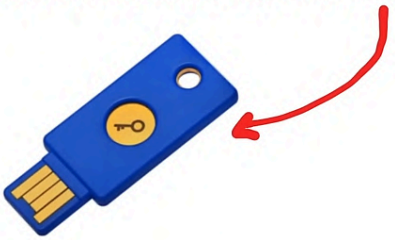


Security Keys

What is a Security Key?
A security device used as the second step in the authentication process to gain access to a device workstation or application.

A security key can resemble a memory stick. When your finger makes contact with a button of exposed metal on the device, it'll generate & autofill a security token.

A popular brand of security key is an Yubikey



- Works out of box with Gmail, Facebook & hundreds more
- Supports FIDO2/WebAuthn, U2F
- Waterproof & crush resistant
- USB-A & NFC dual connectors on a single key

AWS Identity & Access Management (IAM)

AWS Identity & Access Management (IAM) : Create & manage AWS users & groups & use permissions to allow & deny their access to AWS resources.

IAM Policies : JSON documents which grant permissions for a specific user, group or role to access services. Policies are attached to IAM Identities.

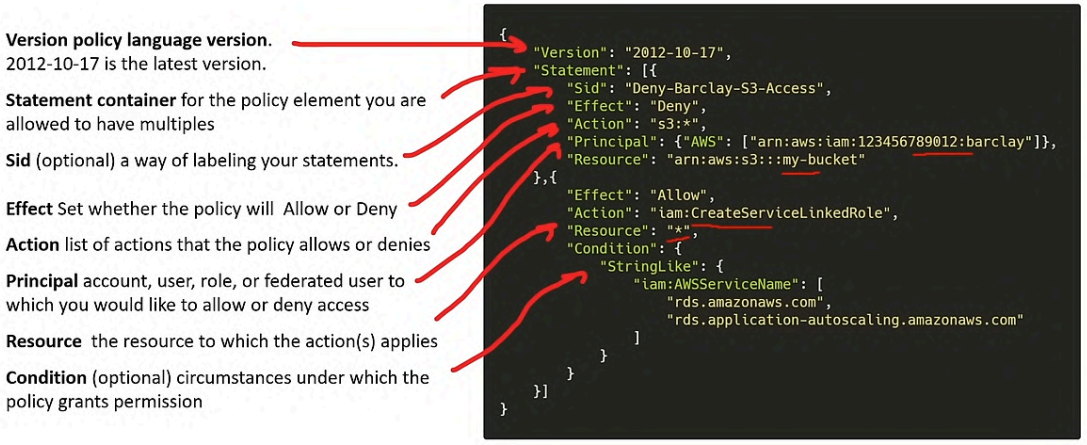
IAM Permission : The API actions that can or cannot be performed are represented in the IAM Policy document.

IAM Identities

- **IAM Users :** End users who log into console or interact with AWS resources programmatically or via clicking UI interfaces
- **IAM Groups** : Group up your Users so they all share permission levels of the group. Eg : Administrators, Developers, Auditors.
- **IAM Roles** : Roles grant AWS resources permissions to specific AWS API actions. Associate policies to a role & then assign it to an AWS resource.

Anatomy of an IAM Policy

IAM Policies are written in JSON & contain permissions which determine what API actions are allowed or denied.



Principle of Least Privilege (PoLP)

Principle of Least Privilege (PoLP) is the computer security concept of providing a user, role or application the least amount of permissions to perform an operation or action.

Just-Enough-Access (JEA) : Permitting only the exact actions for the identity to perform a task.

Just-In-Time (JIT) : Permitting the smallest length of duration an identity can use permissions.

ConsoleMe is an open-source Netflix project to self-serve short-lived IAM Policies so an end user can access AWS Resources while enforcing JEA & JIT.

Risk-based Adaptive Policies :

Each attempt to access a resource generates a risk score of how likely the request is to be from a compromised source.
The risk score could be based on many factors such as : device, user location, IP address, which service is being accessed & when?

AWS at this time does not have Risk-based Adaptive Policies built into IAM

AWS Account Root User

AWS Account - The account which holds all your AWS Resources
AWS Account : Root User - A special account with full access that can't be deleted.
AWS Account : User - A user for common tasks based on assigned permissions.

AWS Account Root User is a special user who is created at the time of AWS account creation :

- The Root User account uses an Email & Password to login.
    - A regular user has to provide the Account ID / Alias, Username & Password.
- The Root user account cannot be deleted.
- The Root user account has full permissions to the account & its permissions ***cannot be limited.**
    - You cannot use IAM policies to explicitly deny the root user access to resources.
    - You can only use an AWS Organisations Service Control Policy (SCP) to limit the permissions of the root user.
- There can only be one Root user per AWS Account.
- The Root user is instead for very specific & specialised tasks that are infrequently or rarely performed.
    - An AWS Root account should not be used for daily or common tasks.
- It's strongly recommended to never use Root User Access Keys.
- It's strongly recommended to turn on MFA for the Root User.
- **Administrative Tasks that only the Root User can perform :**
    - Change your account settings.
        - Includes : Account Name, Email Address, Root User Password & Root User Access Keys.
        - Other account settings such as Contact information, Payment currency options & Regions don't require Root User credentials.
    - Restore IAM User Permissions.

        - If the only IAM administrator accidentally revokes their own permissions, you can sign in as the root user to edit policies & restore those permissions.
    - Activate IAM Access to the Billing & Cost Management Console.
    - View certain tax invoices.
    - Close your AWS Account.
    - Change or Cancel AWS Support plan.
    - Register as a seller in the Reserved Instance Marketplace.
    - Enable MFA Delete on an S3 bucket.
    - Edit or delete an Amazon S3 bucket policy that includes an invalid VPC ID or VPC endpoint ID.
    - Sign up for GovCloud.

AWS Single-Sign-On

AWS Single-Sign-On (AWS SSO) is where you create or connect your workforce identities in AWS once & manage access centrally across your AWS Organisation.

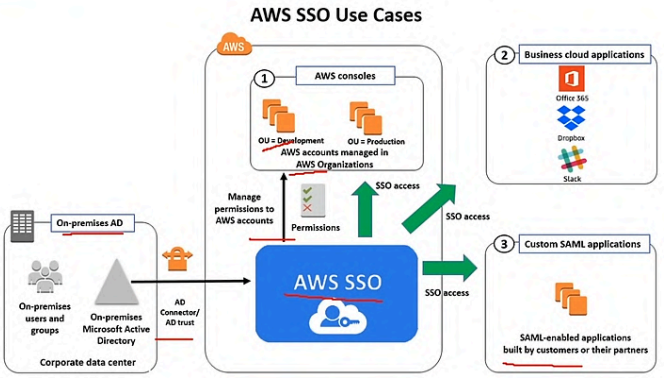
Choose your Identity Source

- AWS SSO
- Active Directory
- SAML 2.0 IpD

Managed User Permissions Centrally

- AWS Account
- AWS Applications
- SAML Applications

Uses get Single Click Access



Application Integration

Introduction to Application Integration

What is Application Integration?
Application Integration is the process of letting two independent applications communicate & work with each other, commonly facilitated by an intermediate system.

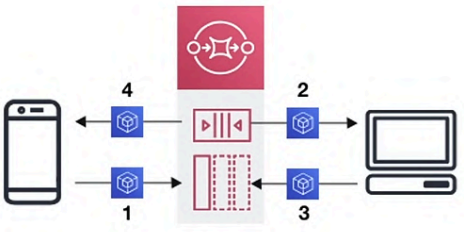
Cloud workloads encourage systems & services to be loosely coupled & so AWS has many services for the specific purpose of application integration.
The common systems or design patterns utilised for Application Integration generally are :

- Queueing
- Streaming
- Pub/Sub
- API Gateways
- State Machine
- Event Bus

Queueing & SQS

What is a Messaging System?
Used to provide asynchronous communication & decouple processing via messages / events.
From a sender & receiver (Producer & Consumer)

What is a Queueing System?
A Queueing system is a messaging system that generally will delete messages once they are consumed.
Simple communication.
Not Real-time.
Have to pull.
Not reactive.

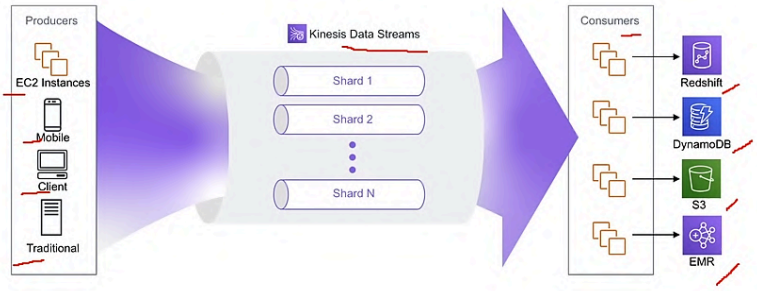


Simple Queueing Service (SQS)
Fully managed queueing service that enables you to decouple & scale microservices, distributed systems & serverless applications.
Use case : You need to queue up transactional emails to be sent.
Example : Signup, Reset Password.

Streaming

What is streaming?
Multiple consumers can react to events (messages)
Events live in the stream for long periods of time, so complex operations can be applied. Real-time

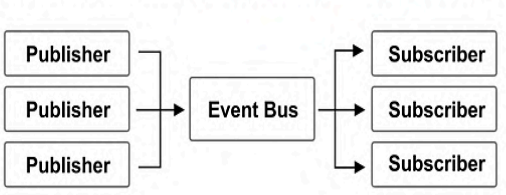
Amazon Kinesis
Amazon Kinesis is the AWS fully managed solution for collecting, processing & analysing streaming data in the cloud.



Publish - Subscribe

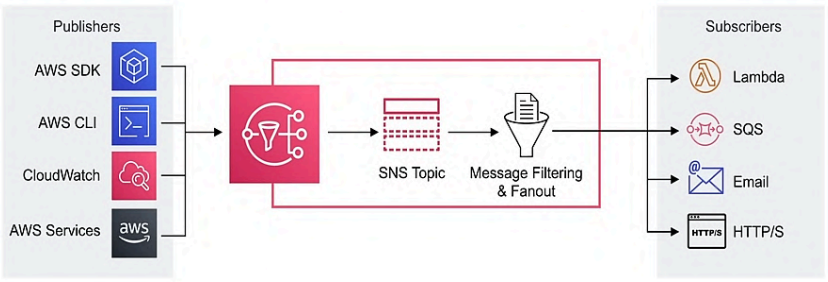
What is Pub/Sub?
Publish - Subscribe pattern commonly implemented in **messaging systems**.
In a pub/sub system the sender of messages (publishers) do not send their messages directly to receivers.
They instead send their messages to an event bus. The event bus categorises their messages into groups.
Then receivers of messages (subscribers) subscribe to these groups.
Whenever new messages appear within their subscription the messages are immediately delivered to them.

- Publishers have no knowledge of who their subscribers are.
- Subscribers do **not pull** for messages.
- Messages are instead automatically & immediately pushed to subscribers.
- Messages & events are interchangeable terms in pub/sub.



Use Case : A real-time chat system. A web-hook system.

Simple Notification Service (SNS) is a highly available, durable, secure, fully managed pub/sub messaging service that enables you to decouple microservices, distributed systems & serverless applications.

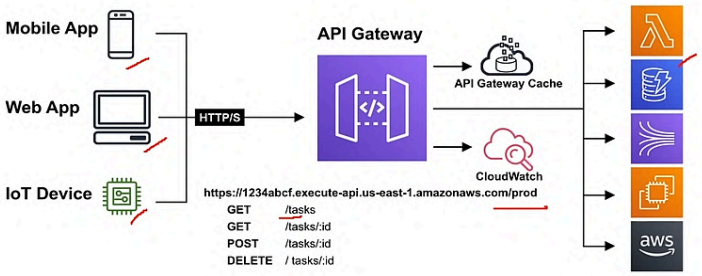


API Gateway

What is an API Gateway?

An API Gateway is a program that sits between a single-entry point & multiple backends.
API Gateway allows for throttling, logging, routing logic or formatting of the request & response.

Amazon API Gateway is a solution for creating secure APIs in your cloud environment at any scale. Create APIs that act as a front door for applications to access data, business logic or functionality from back-end services.



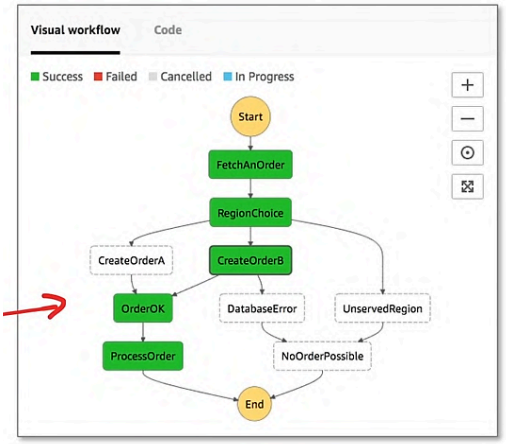
State Machines

What is a state machine?

A state machine is an abstract model which decides how one state moves to another based on a series of conditions. Think of a state machine like a flow chart.

What is AWS Step Function?

- Coordinate multiple AWS Services into a serverless workflow
- A graphical console to visualise the components of your application as a series of steps.



- Automatically triggers & tracks each step & retries when there are errors, so your application executes in order & as expected, every time.
- Logs the state of each step, so when things go wrong, you can diagnose & debug problems quickly.

Event Bus & Amazon Events Bridge

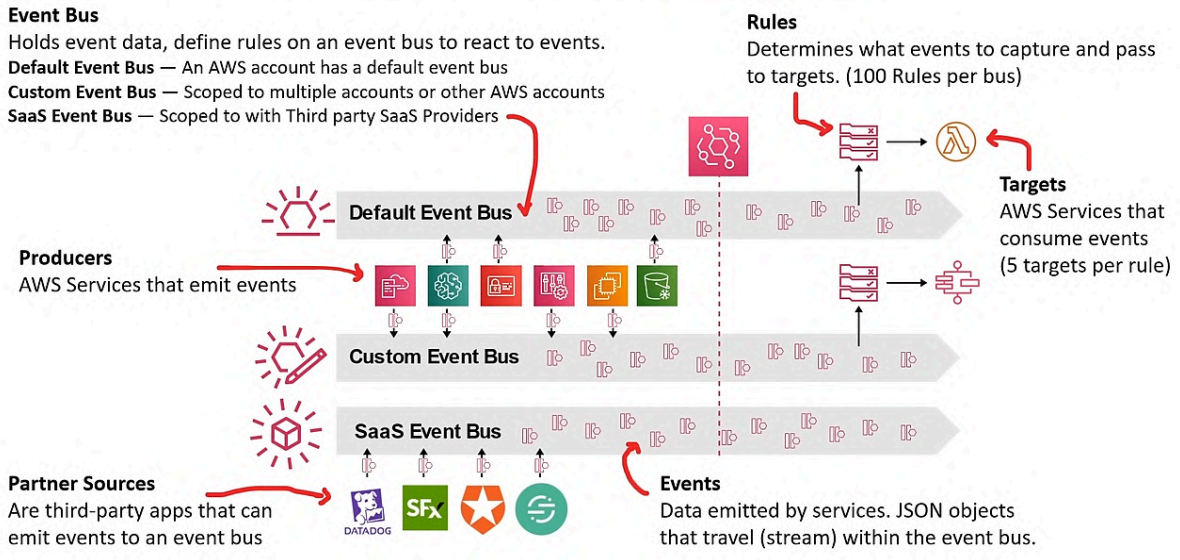
Event Bus

What is an Event Bus?
An event bus **receives events** from a **source** & **routes events** to a **target** based on rules

EventBridge is a serverless event bus service that is used for application integration by **streaming real-time** data to your applications.

EventBridge was formerly called **Amazon CloudWatch** Events.

- Hold event data, define rules on an event bus to react to events.
- **Default Event Bus** - An AWS Account has a default event bus.
- **Custom Event Bus** - Scoped to multiple accounts or other AWS Accounts.
- **SaaS Event Bus** - Scoped with Third party SaaS Providers.
- **Producers** - AWS Services that emit events.
- **Events** - Data emitted by services. JSON Objects that travel (stream) with the event bus.
- **Partner Sources** - Third party apps that can emit events to an event bus.
- **Rules** - Determine what events to capture & pass to targets (100 rules per bus)
- **Targets** - AWS Services that consume events (5 targets per rule)



Application Integration Services

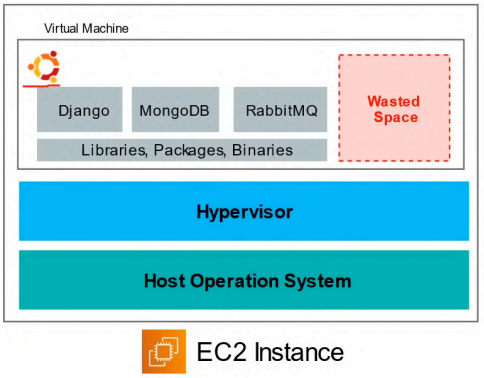
- **Simple Notification Service (SNS) -** A **Publisher-Subscriber messaging system**. Sends notifications via various formats such as Plain-text, Email, HTTP/s (webhooks) SMS (text messages), SQS & Lambda. Push messages which then are sent to subscribers
- **Simple Queue Service (SQS) -** A **queueing messaging service**. Send events to a queue. Other applications pull the queue for messages. Commonly used for background jobs.
- **Step Functions -** A **state machine service**. It coordinates multiple AWS services into serverless workflows. Easily share data among Lambdas. Have a group of lambdas wait for each other. Create logical steps. Also works with Fargate Tasks.
- **EventBridge (CloudWatch Events) -** A **serverless event bus** that makes it easy to connect applications together from your own application, third-party services & AWS services.
- **Kinesis -** A **real-time streaming data service**. Create Producers which send data to a stream. Multiple Consumers can consume data within a stream. Use for real-time analytics, click streams, ingesting data from a fleet of IOT Devices
- **Amazon MQ -** A **managed message broker service that uses Apache ActiveMQ.**
- **Managed Kafka Service (MSK) -** A **fully managed Apache Kafka service**. Kafka is an open-source platform for building real-time streaming data pipelines & applications. Similar to Kinesis but more robust
- **API Gateway -** A fully-managed service for developers to create, publish, maintain, monitor, & secure APIs. You can create API endpoints & route them to AWS services.
- **AppSync -** A **fully managed GraphQL service**. GraphQL is an open-source agnostic query adaptor that allows you to query data from many different data sources.

Containers

VMs vs Containers

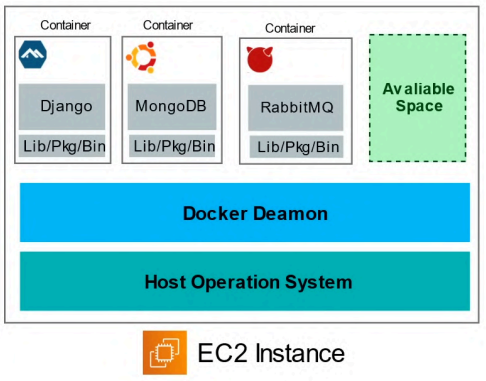
VMs

VMs **do not** make best use of space.
Apps are not isolated which could cause **config conflicts**, **security problems** or **resource hogging**.



Containers

Containers allow you to run multiple apps which are virtually isolated from each other.
Launch new containers & configure OS Dependencies per container.

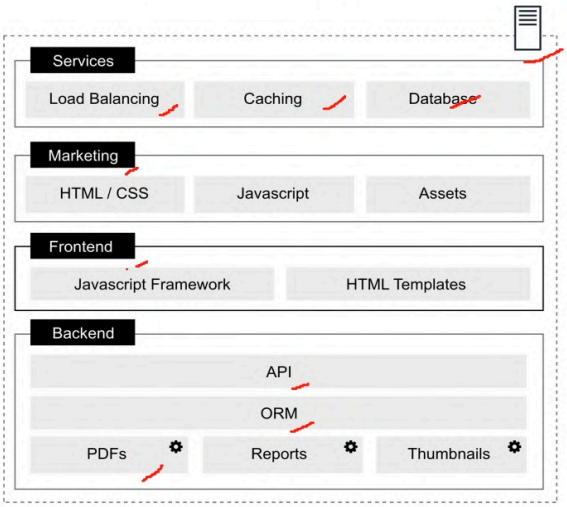


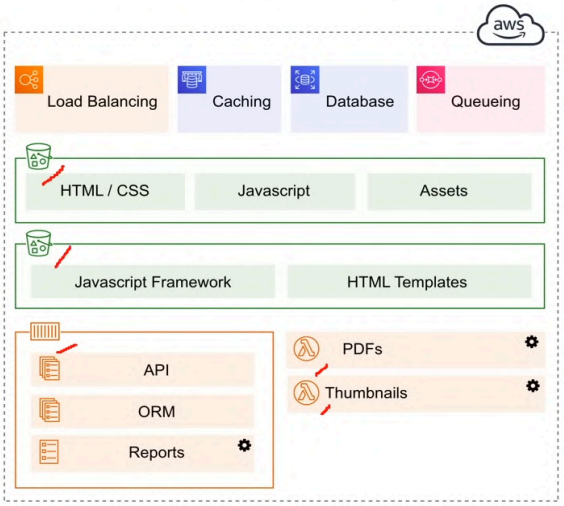
Microservices

Monolithic vs Microservices Architecture

Monolithic Architecture	Microservices Architecture
One App which is responsible for everything.	Multiple Apps which are each responsible for one thing.

Functionality is tightly coupled.	Functionality is isolated & stateless.





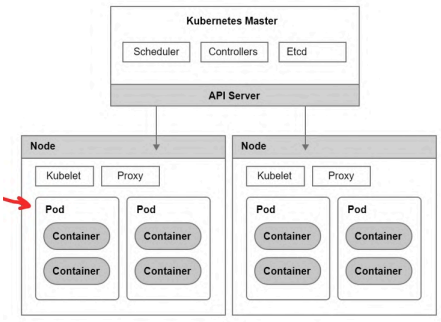
Kubernetes

Kubernetes is an **open-source container orchestration system** for automating deployment, scaling & management of containers.

Originally created by Google & now maintained by the **Cloud Native Computing Foundation (CNCF)**

Kubernetes is commonly called **K8**

- The 8 represent the remaining letters “ubernete”



The advantage of Kubernetes over docker is the ability to run containers distributed across multiple VMs

A unique component of Kubernetes is Pod.
Pod is a group of one or more containers with shared storage, network resources & other shared settings.

Kubernetes is ideal for microservice architectures where a company has **tens to hundreds of services** they need to manage.

Docker

Docker is a set of Platform as a Service (SaaS) Products that use OS-level virtualisation to deliver software in packages called containers.
Docker was the earliest popularised open-source container platform.
When people think of containers, they think of Docker.

- **Docker CLI** - CLI Commands to download, upload, build, run & debug containers.
- **DockerFile** - a configuration file on how to provision a container.
- **Docker Compose** - A Tool & configuration file when working with multiple containers.
- **Docker Swarm** - An orchestration tool for managing deployed multi-containers architectures.
- **DockerHub** - A public online repository for containers published by the community for download.

The Open Container Initiative (OCI) is an open governance structure for creating open industry standards around container formats & runtime.
Docker established the OCI & it is now maintained by the Linux Foundation.

Docker has been losing favour with developers due to their handling of introducing a paid open-source model & alternatives like Podman are growing.

Podman, Buildah & Skopeo

Podman is a container engine that is OCI-complaint & is a drop-in replacement for Docker.

- Podman is daemon-less where Docker uses containerized Daemon.
- Podman allows to create Pods like K8, Docker doesn't have pods.
- Podman only replaces one part of Docker. Podman is to be used alongside Buildah & Skopeo.

Buildah is a tool used to build OCI Images.
Skopeo is a tool for moving container images between different types of container storages.

Container Services

Primary Services

- **Elastic Container Service (ECS)**
  - No Cold Starts
  - Self-managed EC2
- **AWS Fargate**
  - More robust than Lambda
  - Scale to Zero Cost
  - AWS-managed EC2
- **Elastic Kubernetes Services (EKS)**
  - Open Source
  - Avoid Vendor Lock-In
- **AWS Lambda**
  - Only think about code
  - Short running tasks
  - Can deploy custom containers

Provisioning & Deployment

- **Elastic Beanstalk (EB)**
  - ECS on training wheels
  - Platform as a Service
- **App Runner**
  - Platform as a service
  - Specifically for Containers
- **AWS Copilot CLI**
  - Build, release & operate production ready containerized applications on AWS App Runner, Amazon ECS & AWS Fargate.

Supporting Services

- **Elastic Container Registry (ECR)**
  - Repos for your Docker Images.
- **X-Ray**
  - Analyse & debug between microservices.
- **Step Functions**
  - Stitch together Lambdas & ECS Tasks.

Governance

Organisations & Accounts

AWS Organisations allow the creation of new AWS Accounts. Centrally manage billing, control access, compliance, security & share resources across your AWS Accounts.

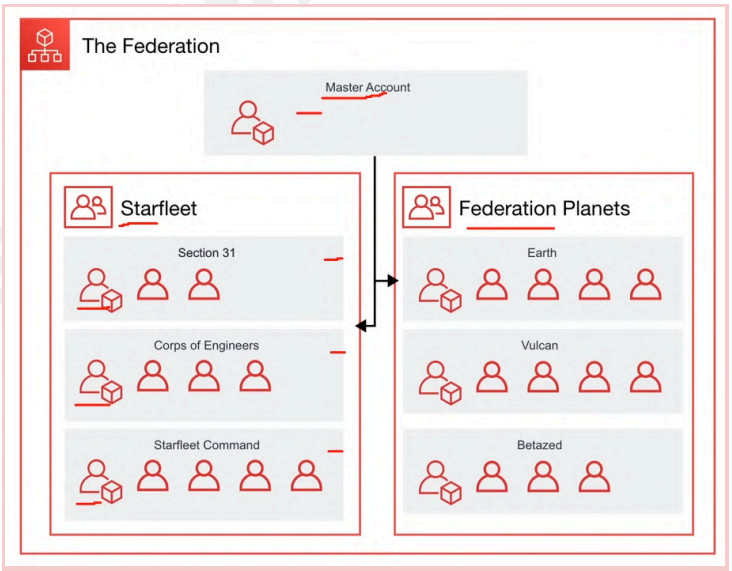
Root Account User is a single sign-in identity that has complete access to all AWS Services & resources in an account. Each account has a Root Account User.

Organisation Unit is a group of AWS Accounts within an organisation which can also contain other organisational units - creating a hierarchy.

Service Control Policies give central control over the allowed permissions for all accounts in your organisation, helping to ensure your accounts stay within your organisation's guidelines.

AWS Organisations must be turned on, once turned on it cannot be turned off. You can create as many AWS Accounts as you like, one account will be the Master/Root Account.

AWS Account is not the same as **User Account**



AWS Control Tower

AWS Control Tower helps Enterprises quickly set-up a secure, **AWS multi-account.**

Provides you with a **baseline environment** to get started with a **multi-account architecture.**

- **Landing Zone**
  - A Landing zone is a baseline environment following well-architectured & best practices to start launching production ready workloads.
    - AWS SSO enabled, Centralised logging for AWS CloudTrail & Cross-account security auditing.
- **Account Factory**
  - Automated provisioning of new accounts in your organisation.
  - Standardise the provisioning of new accounts with pre-approved account configurations.
  - Configure your account factory with pre-approved network configuration & region selections.
  - Enable self-service for your builders to configure & provision new accounts using AWS Service Catalog.
- **Guardrails**
  - Pre-packaged governance rules for security, operations & compliance that customers can select & apply enterprise-wide or to specific groups of accounts.

AWS Control Tower is the *replacement* for retired **AWS Landing Zones.**

AWS Config

What is Change Management?
Change management in the context of Cloud Infrastructure is when we have a **formal process** to

1. Monitor Changes.
2. Enforce Changes.
3. Remediate Changes.

What is Compliance-as-Code (CaC)?
Compliance as code is when we utilise programming to automate the monitoring, enforcing & remediating changes to stay compliant with a compliance program or expected configuration.

What is AWS Config?
AWS Config is a **Compliance-as-Code framework** that allows us to **manage change** in AWS accounts on a per region basis.

When should you use AWS Config?

- I want this **resource** to **stay configured** in a **specific way** for compliance.
- I want to **keep track** of configuration **changes** to resources.
- I want **a list of all resources** within a region.
- I want to analyse potential security weaknesses, you need detailed historical information.

AWS Quick Starts

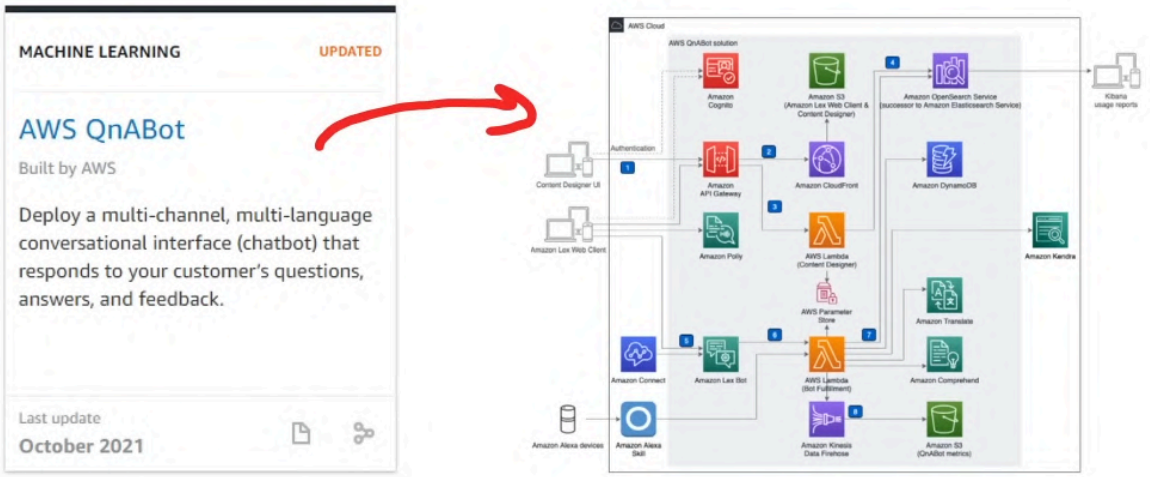
AWS Quick Starts are **prebuilt templates** by AWS & AWS Partners to **help deploy a wide range of stacks.**

Reduce hundreds of manual procedures into just a few steps.

A Quick Start is composed of **3 Parts :**

1. A reference architecture for the deployment.
2. AWS CloudFormation templates that automate & configure the deployment.
3. A deployment guide explaining the architecture & implementation in detail.

Most Quick start reference deployments enable you to spin up a fully functional architecture in less than an hour!



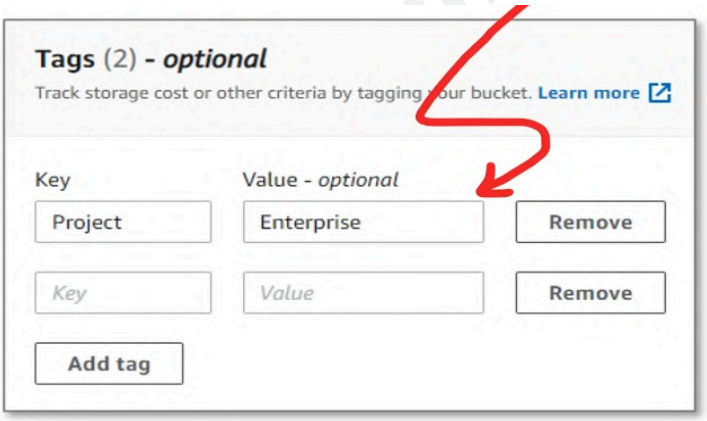
Tagging

A tag is **a key & a value pair** that you can assign to AWS Resources.

Tags allow you to organise your resources in the following ways :

- **Resource Management**
  - Specific workloads.
  - Environments.
  - Eg. Developer Environments.
- **Cost Management & Optimization**

    - Cost Tracking.
    - Budgets.
    - Alerts.
- **Operations Management**
    - Business commitments.
    - SLA Operations.
    - Eg : Mission-Critical Services.
- **Security**
    - Classification of Data.
    - Security Impact.
- **Governance & Regulatory Compliance**
- **Automation**
- **Workload Optimization.**



Tag Examples :

- Dept : Finance
- Status : Approved
- Team : Compliance
- Environment : Production
- Project : Enterprise
- Location : Canada

Resource Groups

Resource Groups are a collection of resources that share one or more **tags**

Helps you organise & consolidate information based on your project & the resources that you use.

Resource Groups can display details about a group of resource based on

- Metrics
- Alarms
- Configuration Settings

At any time you can modify the settings of your resource groups to change what resources appear.

Resource Groups appear in the **Global Console header** & under **Systems Manager**

Business Centric Services

- **Amazon Connect** is a **virtual call centre service.** You can create workflows to route callers. You can record phone calls. Manage a queue of callers. Based on the same proven system used by the Amazon customer service teams.
- **WorkSpaces** is a **virtual remote desktop service.** Secure managed service for provisioning either Windows or Linux desktops in just a few minutes which quickly scales up to thousands of desktops
- **WorkDocs** is a **shared collaboration service.** A centralised storage to share content & files. It is similar to Microsoft SharePoint. Think of it as a shared folder where the company has ownership
- **Chime** is a **video-conference service.** It is similar to Zoom or Skype. You can screen share, have multiple people on the call. It is secure by default & it can show you a calendar of your upcoming calls.
- **WorkMail** is a **managed business email, contacts, & calendar service** with support for existing desktop & mobile email client applications. (IMAP). Similar to Gmail or Exchange.
- **Pinpoint** is a **marketing campaign management service.** Pinpoint is for sending targeted email via SMS, push notifications, & voice messages. You can perform A/B testing or create Journeys (complex email response workflows)
- **Simple Email Service (SES)** is a **transactional email service.** You can integrate SES into your application to send emails. You can create a common template, track open-rates & keep track of your reputation.
- **QuickSight** is a **Business Intelligence (BI) service.** Connect multiple data sources & quickly visualise data in the form of graphs with little to no programming knowledge.

Provisioning

Provisioning Services

What is provisioning?
The allocation or creation of resources & services to a customer.
AWS Provisioning Services are responsible for setting up & then managing those AWS Services.

- **Elastic Beanstalk (EB)** is a **Platform as a Service (PaaS) to easily deploy web-applications.** EB will provision various AWS services, including EC2, S3, Simple Notification Service (SNS), CloudWatch, EC2 Auto Scaling Groups, & Elastic Load Balancers. If you have ever used **Heroku** it the AWS equivalent.
- **AWS OpsWorks** is a **configuration management service** that also provides managed instances of the open-source configuration managed software **Chef** & **Puppet**.
- **CloudFormation** is an **infrastructure modelling & provisioning service.** Automate the provisioning of AWS Services by writing CloudFormation templates in either **JSON** or **YAML files.** This is known as **Infrastructure as Code (IaC).**
- **AWS QuickStarts** are pre-made packages that can launch & configure your AWS computer, network, storage, & other services required to deploy a workload on AWS.
- **AWS Marketplace** is a **digital catalogue** of **thousands** of software listings from independent software vendors you can use to find, buy, test, & deploy software.
- **AWS Amplify** is **a mobile & web-application framework** that will provision multiple AWS services as your backend.
- **AWS App Runner** is a fully managed service that makes it easy for developers to quickly deploy containerized web applications & APIs, at scale & with no prior infrastructure experience required.
- **AWS Copilot** is a command line interface (CLI) that enables customers to quickly launch & easily manage containerized applications on AWS.
- **AWS CodeStar** provides a unified user interface, enabling you to easily manage your software development activities in one place. Easily launch common types of stacks eg. LAMP.
- **AWS Cloud Development Kit (CDK)** is an Infrastructure as Code (IaC) tool. Allows you to use your favourite programming language. Generates CloudFormation templates as the means for IaC.

AWS Elastic Beanstalk

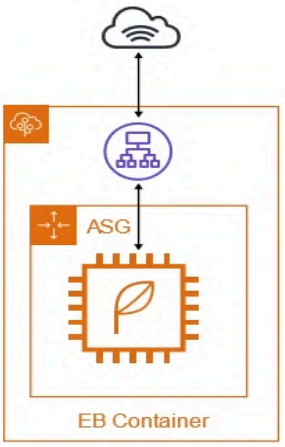
What is Platform as a Service? (PaaS)
A PasS allows customers to develop, run, & manage applications without the complexity of building & maintaining the infrastructure typically associated with developing & launching an app.

Elastic Beanstalk is a PaaS for deploying web-applications with little-to-no knowledge of the underlying infrastructure so you can focus on writing application code instead of setting up an automated deployment pipeline & DevOps tasks. Choose a platform, upload your code & it runs with little knowledge of the infrastructure.

Not Recommended for "Production" applications. (AWS is talking about enterprise, large companies.)

Elastic Beanstalk is powered by a CloudFormation template for you:

- Elastic Load Balancer.
- Autoscaling Groups.
- RDS Database.
- EC2 Instance preconfigured (or custom) platforms.
- Monitoring (CloudWatch, SNS).
- In-Place & Blue/Green deployment methodologies.
- Security (Rotates passwords).
- Can run Dockerized environments.



Serverless

Serverless Services

What is Serverless?
When the underlying servers, infrastructure & Operating System (OS) is taken care of by the Cloud Service Provider (CSP). Serverless is generally by default highly available, scalable & cost-effective. You pay for what you use.

- **DynamoDB** is a serverless **NoSQL key/value & document database.** It is designed to scale to **billions of records** with guaranteed consistent data return in at least a second. You don't have to worry about managing shards!
- **Simple Storage Service (S3)** is a **serverless object storage service.** You can upload a very large & an unlimited amount of files. You pay for what you store. You don't worry about the underlying file-system, or upgrading the disk size.
- **ECS Fargate** is a **serverless orchestration container service.** It is the same as ECS except you pay-on-demand port running containers (With ECS you have to keep an EC2 server running even if you have no containers running) AWS manages the underlying server, so you don't have to scale or upgrade the EC2 server.
- **AWS Lambda** is a **serverless functions service.** You can run code without provisioning or managing servers. You upload small pieces of code, choose much memory & how long the function is allowed to run before timing out. You are charged based on the runtime of the serverless function rounded to the nearest 100ms.
- **Step Functions** is a **state machine service.** It coordinates multiple AWS services into serverless workflows. Easily share data among Lambdas. Have a group of lambdas wait for each other. Create logical steps. Also works with Fargate Tasks.
- **Aurora Serverless** is the **serverless on-demand version of Aurora.** **When you want "most" of the benefits of Aurora but can trade to have cold-starts or you don't have lots of traffic demand.**

A serverless service could have all or most of the following characteristics:

- Highly elastic & scalable
- highly available
- Highly durable
- Secure by default

- Abstracts away the underlying infrastructure & are billed based on the execution of your business task.
- Serverless can **Scale-to-Zero** meaning when not in use the serverless resources cost nothing.

Pay-for-Value (you don't pay for idle servers).
An analogy of serverless could be similar to an energy rating label which allows consumers to compare the energy efficiency of a product. Some services are more serverless than others.

Serverless Architecture

What is Serverless Architecture?
Serverless architecture generally describes fully managed cloud services. The classification of a cloud service being serverless is not a Boolean answer (yes or no), but an answer on a scale where a cloud service has a degree of serverless.

Windows on AWS

Windows on AWS

AWS has multiple cloud services & tools to make it easy for you to run Windows workloads on AWS.

- **Windows Servers on EC2** : You can select from a number of Windows Server versions including the latest version, Windows Server 2019
- **SQL Server on RDS** : You can select from a number of SQL Server database versions
- **AWS Directory Service** : You can run **Microsoft Active Directory (AD) as a managed service.**
- **AWS License Manager** : You can easily manage your software licences from software vendors such as Microsoft.
- **Amazon FSx for Windows File Server** : You can **fully manage scalable storage** built for Windows.
- **AWS Software Development Kit (SDK)** : You can write code in your favourite language to interact with AWS API. The SDK supports **.NET** a language favourite for Windows Developers
- **Amazon WorkSpaces** : You can run a virtual desktop. You can launch a **Windows 10 desktop** to provide a secure & durable workstation that is accessible from wherever you have an internet connection.
- **AWS Lambdas** supports **PowerShell** as a programming language to write your serverless functions!
- **AWS Migration Acceleration Program (MAP)** for Windows is a migration methodology for moving large enterprises. AWS has Amazon Partners that specialise in providing professional services for MAP.

AWS Licence Manager

What is Bring-Your-Own-License? (BYOL)
The process of reusing an existing software licence to run vendor software on a cloud vendor's computing service. BYOL allows companies to save money since they may have purchased the licence in bulk or at a time that provided a greater discount than if purchased again.

Eg. **Licence Mobility** is Microsoft Volume Licensing customers with eligible server applications covered by active Microsoft Software Assurance (SA)

AWS License Manager is a service that makes it easier for you to manage your software licences from software vendors centrally across AWS & your on-premises environments.
AWS Licence Manager software that is licensed based on **virtual cores (vCPUs), physical cores, sockets, or number of machines**. This includes a variety of software products from Microsoft, IBM, SAP, Oracle, & other vendors.
AWS License Manager works with:

- EC2 - Dedicated Instances, Dedicated Hosts, Spot Instances
- RDS - (Only for Oracle databases)

For **Microsoft Windows Server** & **Microsoft SQL Server licence** you generally need to use a **Dedicated Host**

Logging

Logging Services

CloudTrail - logs all **API calls** (SDK, CLI) between **AWS services** (who can we blame)

- To find out
  - Who created this bucket?
  - Who spun up that expensive EC2 instance?
  - Who launched this SageMaker Notebook?
- Detect developer misconfiguration
- Detect malicious actors
- Automate responses

CloudWatch is a collection of multiple services

- CloudWatch **Logs** A centralised place to store your cloud services log data or application logs.
- CloudWatch **Metrics** Represents a time-ordered set of data points. A variable to monitor
- CloudWatch **Events (EventBridge)** trigger an event based on a condition eg. ever hour take snapshot of server
- CloudWatch **Alarms** triggers notifications based on metrics
- CloudWatch **Dashboard** create visualisations based on metrics

AWS X-Ray is a **distributed tracing system**. You can use it to pinpoint issues with your microservices. See how data moves from one app to another, how long it took to move, & if it failed to move forward.

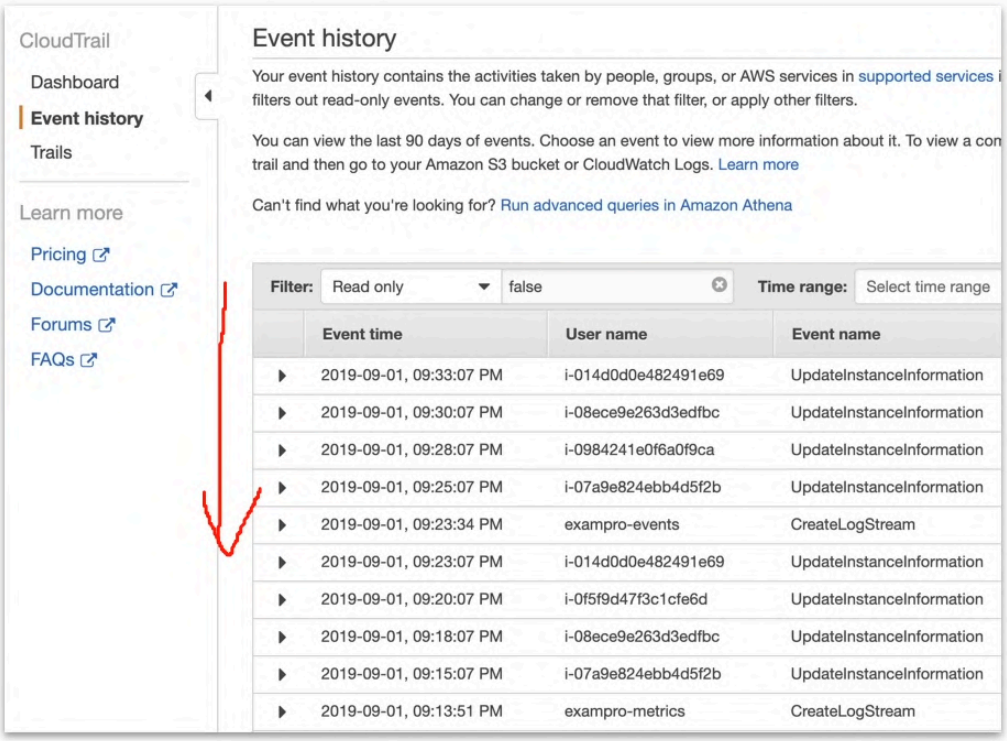
AWS CloudTrail

AWS CloudTrail is a service that enables governance, compliance, operational auditing, & risk auditing of your AWS account.

AWS CloudTrail is used to monitor API calls & Actions made on an AWS account & easily identify which users & accounts made the call to AWS .

- **Where** – Source IP Address.
- **When** – EventTime.
- **Who** – User, UserAgent.
- **What** – Region, Resource, Action.

CloudTrail is already logging by default & will collect logs for the **last 90 days** via **Event History**.
If you need more than 90 days you need to create a **Trail**.
Trails are output to S3 & do not have GUI like Event History. To analyse a Trail you'd have to use **Amazon Athena**.



CloudWatch Alarms

A CloudWatch Alarm monitors a **CloudWatch Metric** based on **a defined threshold**.
When an alarm breaches (goes outside the defined threshold) then it changes **state**.

- **Metric Alarm States**
  - **OK** The metric or expression is **within** the defined threshold
  - **ALARM** The metric or expression is **outside** of the defined threshold
  - **INSUFFICIENT DATA**
    - The alarm has **just started**
    - the metric is **not available**
    - **Not enough data** is available

When it changes state we can define what **action it should trigger**. Such as

- Notification
- Auto Scaling Group
- EC2 Action

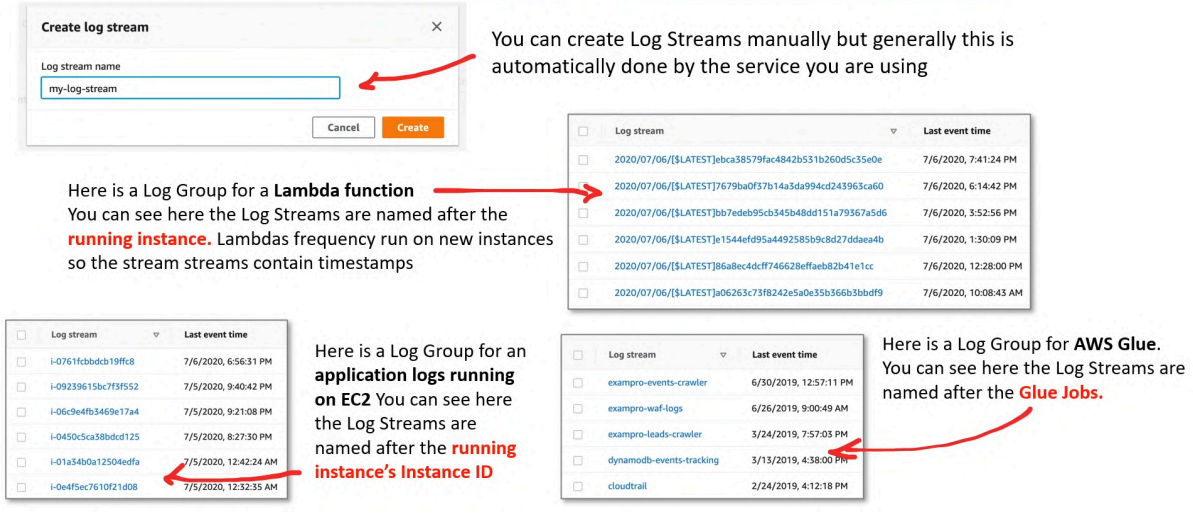
Anatomy of an Alarm

- **Threshold Condition :** Defines when a datapoint is breached
- **Metric :** The actual data we are measuring
- **NetworkIn :** The volume of incoming network traffic. Measured in bytes. When using 5 min monitoring divide by 300 to get Bytes/second.
- **Data Point :** Represents the metric's measurement at a given period.
- **Period :** How often it checks to evaluate the Alarm.
- **Evaluation Periods :** Number of previous periods.
- **Data Points to Alarm :** 1 data point is breached in an evaluation period going back 4 periods. *This is what triggers the alarm.*

CloudWatch Logs

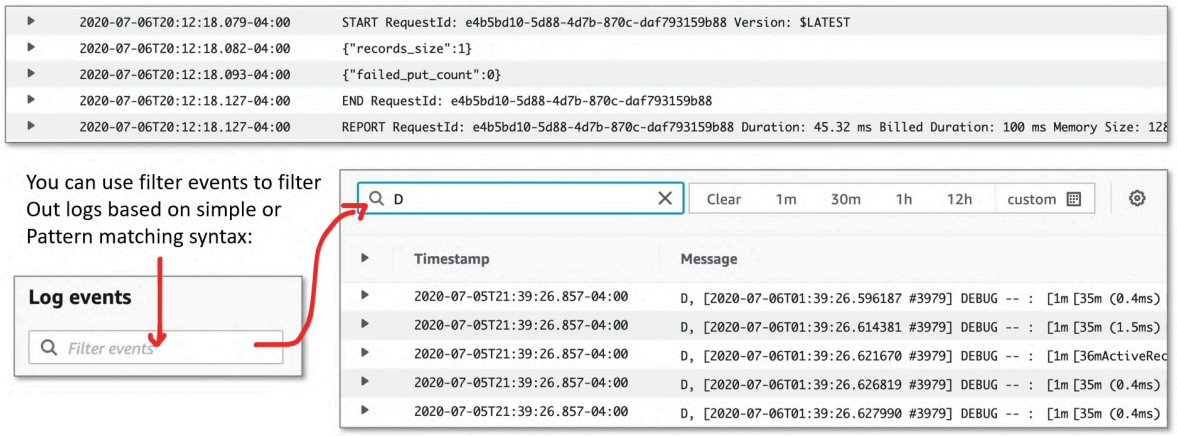
Log Streams

A Log stream represents a **sequence of events** from an **application or instance being monitored.**
You can create Log Streams manually but generally this is automatically done by the service you are using.



Log Events

Represents a single event in a log file. Log events can be seen within a Log Stream.



Log Insights

CloudWatch Logs Insights enables you to **interactively search & analyse your CloudWatch log data** & has the following advantages:

- More robust filtering then using the simple Filter events in a Log Stream
- Less burdensome than having to export logs to S3 & analyse them via Athena.

CloudWatch Logs Insights supports all types of logs.

- CloudWatch Logs Insights is commonly used via the console to do ad-hoc queries against logs groups.

CloudWatch Insights has its own language called:

- CloudWatch Logs Insights **Query Syntax**.
- A single request can query up to **20 log groups**.
- Queries **timeout after 15 minutes**, if they have not completed.
- Query results are **available for 7 days**.

AWS provides sample queries that can get you started for common tasks & to ease learning the Query Syntax. A good example Is filtering VPC Flow Logs.

You can create & save your own queries to make future repetitive tasks easier.

CloudWatch Metrics

A **CloudWatch Metric** represents a **time-ordered set of data points**. It is a **variable** that is **monitored over time**.

CloudWatch comes with many **predefined** metrics that are generally name spaced by AWS Service

EC2 Per-Instance Metrics

- CPUUtilization
- DiskReadOps
- Disk WriteOps
- DiskReadBytes
- DiskWriteBytes
- NetworkIn
- NetworkOut
- NetworkPacketsin
- NetworkPacketsOut

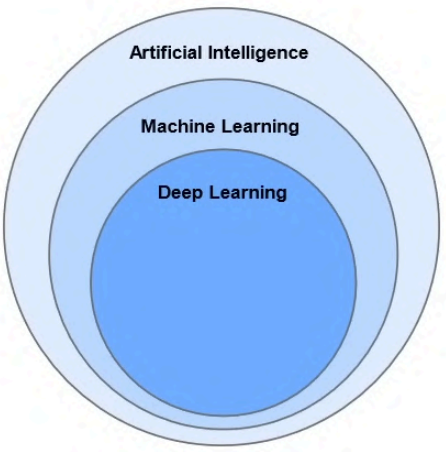
Machine Learning, Artificial Intelligence & Big Data

Introduction to Machine Learning & Artificial Intelligence

What is Artificial Intelligence (AI)?
Machines that perform jobs that mimic human behaviour

What is Machine Learning (ML)?
Machines that get better at a task without explicit programming

What is Deep Learning (DL)?
Machines that have an artificial neural network inspired by the human brain to solve complex problems.



Amazon SageMaker is a fully managed service to **build, train, & deploy machine learning models** at scale

- Apache MXNet on AWS, open-source deep learning framework
- TensorFlow on AWS open-source machine intelligence library
- PyTorch on AWS open-source machine learning framework

Amazon SageMaker Ground Truth is a **data-labelling service**. Have humans label a dataset that will be used to train machine learning models

Amazon Augmented AI human-intervention review service. When SageMaker's use machine Learning to make a prediction it is not confident it has the right answer to queue up the prediction for human review.

Machine Learning & Artificial Intelligence Services

- **Amazon CodeGuru** is a **machine-learning code analysis service**. CodeGuru performs code-reviews & will suggest changes to improve the quality of code. It can show visual code profiles (show the internals of your code) to pinpoint performance.
- **Amazon Lex** is a **conversion interface service**. With Lex you can build **voice & text chatbots**.

- **Amazon Personalise** is a **real-time recommendations** service. Same technology used to make product recommendations to customers shopping on the Amazon platform.
- **Amazon Polly** is a **text-to-speech service**. Upload your text & an audio file spoken by synthesised voice is generated.
- **Amazon Rekognition** is an **image & video recognition service**. Analyse images & videos to detect & label objects, people, celebrities.
- **Amazon Transcribe** is a **speech-to-text service**. Upload your audio file & it is converted.
- **Amazon Textract** & **OCR (extract text from scanned documents) service**. When you have paper forms & you want **to digitally extract the data**.
- **Amazon Translate** **neural machine learning translation service**. Uses deep learning models to deliver more accurate & natural sounding translations.
- **Amazon Comprehend** is a **Natural Language Processing (NLP) service**. Find relationships between text to produce insights. Looks at data such as Customer emails, support tickets, social media & makes predictions.
- **Amazon Forecast** is a **time-series forecasting service**. Forecast business outcomes such as product demand, resource needs or financial performance.
- **AWS Deep Learning AMIs** Amazon EC2 instances **pre-installed with popular deep learning frameworks & interfaces** such as TensorFlow, PyTorch, Apache MXNet, Chainer, Gluon, Horovod, & Keras.
- **AWS Deep Learning Containers** Docker images instances pre-install with popular deep learning frameworks & interfaces such as TensorFlow, PyTorch, & Apache MXNet.
- **AWS DeepComposer** is a **machine-learning enabled musical keyboard**.
- **AWS DeepLens** is a **video-camera that uses deep-learning**.
- **AWS DeepRacer** a **toy race car** that can be powered with machine-learning to perform **autonomous driving**.
- **Amazon Elastic Inference** allows you to attach low-cost GPU-powered acceleration to EC2 instances to reduce the cost of running deep learning inference by up to 75%.
- **Amazon Fraud Detector** is a **fully managed fraud detection service**. identify potentially fraudulent online activities such as online payment fraud & the creation of fake accounts.
- **Amazon Kendra** **enterprise machine learning search engine service**. Uses natural language to suggest answers to questions instead of just simple keyword matching.
- **Amazon SageMaker** is a unified ML platform for building ML solutions end-to-end.

- **Amazon Bedrock** is a **Large Language Model (LLM) cloud service offering** to generate text & image responses. **Think like ChatGPT**.
- **Amazon CodeWhisper** is an AI code generator that will predict code to meet your use case. Think like **Github Copilot**.
- **Amazon DevOps Guru** uses ML to analyse your operational data & application metrics & events to detect operational abnormalities. Is there something wrong with our cloud operations?
- **Amazon Lookout** for Equipment / Metrics / Vision uses ML models for quality control & performs automated inspections.
- **Amazon Monitron** uses ML models to predict unplanned equipment downtime. Monitor has an IOT sensor that captures vibrations & sensor data
- **AWS Neuron** is an SDK used to run deep learning workloads on AWS Inferentia & AWS Trainium based instance

Big Data & Analytics Services

What is BigData?
A term used to describe **massive volumes of structured/unstructured data** that is so large it is difficult to **move & process** using traditional database & software techniques.

- **Amazon Athena** is a **serverless interactive query service**. It can take a bunch of CSV or JSON files in an S3 Bucket & load them into temporary SQL tables so you can run SQL queries. When you want to query CSV or JSON files
- **Amazon CloudSearch** is a fully managed **full-text search service**. When you want to add search to your website
- **Amazon Elasticsearch Service (ES)** is a **managed Elasticsearch cluster**. Elasticsearch is an open-source full-text search engine. It is more robust than CloudSearch but requires more server & operational maintenance.
- **Amazon Elastic MapReduce (EMR)** is for data processing & analysis. It can be used for creating reports just like Redshift but is more suited when you need to transform unstructured data into structured data on the fly.
- **Kinesis Data Streams** is a **real-time streaming data service**. Create **Producers** which send data to a stream. **Multiple Consumers** can consume data within a stream. Use for real-time analytics, clickstreams, ingesting data from a fleet of IoT Devices
- **Kinesis Firehose** is serverless & a simpler version of Data Streams, You pay on-demand based on how much data is consumed through the stream & you don't worry about the underlying servers.

- **Amazon Kinesis Data Analytics** allows you to run queries against data that is flowing through your real-time stream so you can create reports & analysis on emerging data.
- **Amazon Kinesis Video Streams** allows you to analyse or apply processing on real-time streaming video.
- **Managed Kafka Service (MSK)** a **fully managed Apache Kafka service**. Kafka is an open-source platform for building real-time streaming data pipelines & applications. It is similar to Kinesis but with more robust functionalities
- **Redshift** is a **petabyte-size data-warehouse**. Data-warehouses are for Online Analytical Processing) OLAP Data-warehouses can be expensive because they are keeping data "hot". Meaning that we can run a very complex query & a large amount of data & get that data back very fast. When you want to quickly generate analytics or reports from a large amount of data.
- **Amazon QuickSight** is a **business intelligence (BI) dashboard**. You can use it to create business dashboards to power business decisions. It requires little to no programming knowledge, connects & ingests to many different types of databases
- **AWS Data Pipeline** **automates the movement of data**. You can reliably move data between compute & storage services.
- **AWS Glue** is an **Extract, Transform, Load (ETL) service**. Moving data from one location to another & where you need to perform transformations before the final destination. Similar to Database Migration Service (DMS) but more robust
- **AWS Lake Formation** is a **centralised, curated, & secured repository that stores all your data**. A **data lake** is a storage repository that holds a vast amount of raw data in its native format until it is needed.
- **AWS Data Exchange** is a catalogue of third-party datasets. You can download for free, subscribe or purchase datasets. Eg. COVID-19 Foot Traffic Data, IMDB TV & Movie data, Historical Weather Data

Amazon QuickSight

Amazon QuickSight is a **Business Intelligence (BI) Dashboard** that allows you to ingest data from various AWS storage or database services to **quickly visualise business data** with minimal programming or data formula knowledge.

QuickSight uses **SPICE** (super-fast, parallel, in-memory, calculation engine) to achieve blazing fast performance at scale

Amazon QuickSight ML Insights – Detects Anomalies, Perform accurate forecasting, Generate Natural Language Narratives.

Amazon QuickSight Q - Ask questions using natural language, on all your data, & receive answers in seconds.

Generative Artificial Intelligence

What is Generative AI?
Generative AI is a type of Artificial intelligence **capable of generating new content** such as **text**, **images**, **music** or **other forms** of media.
Example : **MidJourney** generative AI for Graphics.

Machine Learning / Deep Learning Frameworks & Tools

Here are some common Machine Learning & Deep Learning Frameworks Most of which can be used within SageMaker, or have direct support.

- **Apache MXNet** – adopted by AWS, supports both imperative & symbolic
- **PyTorch** – optimised tensor library for deep learning using GPUs & CPU (created by Facebook)
- **TensorFlow** – low-level machine learning framework (created by Google)
- **Keras** – high level machine learning framework built on top & ships with TensorFlow
- **Apache Spark** – unified analytics engine for large-scale data processing.
- **SparkML** – uniform set of high-level APIs that help users create & tune practical machine learning pipelines.
- **Chainer** – powerful, flexible & intuitive deep learning framework, supports CUDA.
- **Hugging Face** - An AI Community of ML Models & dataset.

ML frameworks which you might come across but are no longer in active development have been absorbed or abandon or the community just doesn't really use them: Caffe2, Theano, DSSTNE, CNTK

Apache MXNet

Apache MXNet is **a deep learning machine learning framework** which supports many different programming languages :

- Python, Java, Julia, Matlab, R, Javascript, Go, R, Scala, Perl, Wolfram language

They key features of Apache MXNet:

- **Scalable –** designed for distributed cloud infrastructure.
- **Flexible –** supports both imperative & symbolic programming.
- **Portable –** can be used on low-end or edge devices & on serverless compute.
- **Multiple –** Programming languages.

AWS has made Apache **MXNet their ML framework of choice**
There is lots of support to use Apache MXNet within AWS SageMaker & AWS ML containers.

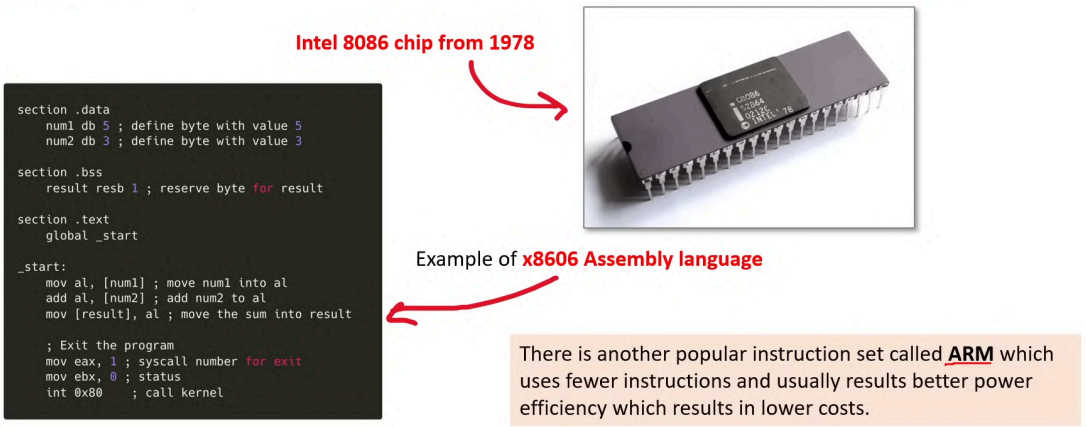
MXNet has two high-level interfaces:

- Gluon API - **imperative** programming.
- Module API - **symbolic** programming.

Intel

What is Intel?

Intel is a multinational corporation & is one of the world's largest semiconductor chip manufacturers. Intel is the inventor of the **x86 instruction set**.



There is another popular instruction set called ARM which uses fewer instructions & usually results in better power efficiency which results in lower costs.

Intel Xeon Scalable & Intel Gaudi

Intel Xeon Scalable Processor – The Intel Xeon Scalable Processor is a high-performance CPU designed for enterprise & server applications, commonly used in AWS instances

Intel Habana Gaudi – AI training processor developed by Habana Labs, a company acquired by Intel. Gaudi is tailored for training deep learning models. Gaudi is often viewed as a competitor to NVIDIA's GPUs. While NVIDIA's GPUs are highly versatile & widely adopted, Gaudi offers a specialised alternative that's optimised specifically for AI training.

Graphics Processing Unit

What is a GPU?
A General Processing Unit (GPU) is a processor that is specialised to quickly render high-resolution images & video **concurrently**. **Great for Video Games**.

GPUs can perform parallel operations on multiple sets of data, & so they are commonly used for non-graphical tasks such as machine learning & scientific computation.

CPUs can have an average of 4 to 16 processor cores... GPUs can have **thousands of processor cores**. 4 to 8 GPUs can provide as many as 40,000 cores.

GPUs are best suited for repetitive & highly-parallel computing tasks:

- Rendering graphics.
- Cryptocurrency mining.
- Deep Learning & ML.

Compute Unified Device Architecture (CUDA)

NVIDIA is a company that manufactures **graphical processing units (GPUs)** for gaming & professional markets.

CUDA stands for Compute Unified Device Architecture & is a **parallel computing platform** & **API** by NVIDIA that allows developers to use **CUDA-enabled GPUs** for general-purpose computing on GPUs. (GPGPU)

All major deep learning frameworks are integrated with **NVIDIA Deep Learning SDK**.

- EC2 P3 Instances have up to 8 NVIDIA **Tesla V100** GPUs
- EC2 G3 Instances have up to 4 NVIDIA **Tesla M60** GPUs
- EC2 G4 Instances have up to 4 NVIDIA **T4** GPUs
- EC2 P4 Instances have up to 8 NVIDIA **Tesla A100** GPUs

The NVIDIA Deep Learning SDK is a collection of NVIDIA libraries for deep learning. One of those libraries is the **CUDA Deep Neural Network library (cuDNN)**
cuDNN provides highly tuned implementations for standard routines such as:

- Forward & backward convolution.
- Pooling.
- Normalisation.
- Activation layers.

AWS Well-Architected Framework

AWS Well-Architected Framework

The AWS Well-Architected Framework is a Whitepaper created by AWS to help customers build using best practices defined by AWS.

AWS Well-Architected

The framework is divided into 6 sections called pillars which address different aspects of “lenses” that can be applied to a cloud workload.

6 Pillars

- Operational Excellence.
- Security.
- Reliability.
- Performance Efficiency.
- Cost Optimization.
- [New] Sustainability.

AWS Well-Architected Framework

- General Definitions.
- General Design Principles.
- The Review Process.

General Definitions

- **Operational Excellent Pillar** – Run & monitor systems.
- **Security Pillar** – Protect data & systems, mitigate risk.
- **Reliability Pillar** – Mitigate & recover from disruptions.
- **Performance Efficiency Pillar** – Use computing resources effectively .
- **Cost Optimization Pillar** – Get the lowest price.
- **[New] Sustainability Pillar** – Use environmental best practices.
- **Component** – Code, Configuration, & AWS Resource against a requirement.
- **Workload** – A set of components that work together to deliver business value.
- **Milestones** – Key changes of your architecture through product life cycle.
- **Architecture** – How components work together in a workload.
- **Technology Portfolio** – A collection of workloads required for the business to operate.

On Architecture

- The AWS Well-Architected Framework is designed around a different kind of team structure.
- Enterprises generally have centralised teams with specific roles whereas AWS has distributed teams with flexible roles.
- Distributed teams can come with new risks, AWS mitigates these with Practices, Mechanisms, & Leadership Principles.

On-Premise Enterprise	Amazon Web Services

Centralised Team
• Technical Architect. (infrastructure)
• Solution Architect. (software)
• Data Architect.
• Networking Architect.
• Security Architect.
	Distributed Teams
	• Practices.
	  ○ Team Experts. (Raise the Bar)
	• Mechanisms.
	  ○ Automated Checks for Standards.
	• *Amazon Leadership Principle.

Managed by either TOGAF or Zachman Framework.	Supported by a virtual community of SMEs, Principle Engineers.

Amazon Leadership Principles

The **Amazon Leadership Principles are a set of principles** used during the company **decision-making**, **problem-solving**, **simple brainstorming** & **hiring**.

1. Customer Obsession.
2. Ownership.
3. Invent & Simplify.
4. Are Right, A Lot.
5. Learn & Be Curious.
6. Hire & Develop the Best.
7. Insist on the Highest Standards.
8. Think Big.
9. Bias for Action.
10. Frugality.
11. Earn Trust.
12. Dive Deep.
13. Have Backbone; Disagree & Commit.
14. Deliver Results.
15. Strive to be Earth's Best Employer.
16. Success & Scale Bring Broad Responsibility.

General Design Principles

Stop guessing your capacity needs
Eg. Cloud computing you use is as little or much-based **on demand**.

Test systems at production scale
Eg. Clone production env to testing, Teardown testing not in use to save money.

Automate to make architectural experimentation easier
Eg. Using CloudFormation with ChangeSets, StackUpdate, & Drift Detection

Allow for evolutionary architectures
Eg. CI/CD, rapid or nightly releases, Lambdas deprecating run-times forcing you to evolve

Drive architectures using data
Eg. CloudWatch, Cloud Trail automatically turned on collecting data

Improve through game days
Eg. simulate traffic on production or purposely kill EC2 instances to see test recovery

Anatomy of a Pillar

A Pillar of the Well-Architected Framework is **structured** as follows:

- Design Principles
    - A list of design principles that need to be considered during implementation.
- Definition
    - An overview of the best practice categories.
- Best Practices
    - Detailed information about each best practice with AWS Services.
- Resources
    - Additional documentation, whitepapers, & videos to implement this pillar.

Design Principles

Operational Excellence

- **Perform operations as code**
    - Apply the same engineering discipline you would to application code to your cloud infrastructure. By treating your operations as code you can limit human error & enable consistent responses to events.
    - Eg. Infrastructure as Code.

- **Make frequent, small, reversible changes**
    - Design workloads to allow components to be updated regularly.
    - Eg. rollbacks, incremental changes, Blue/Green, CI/CD.

- **Refine operations procedures frequently**
    - Look for continuous opportunities to improve your operations
    - Eg. Use game days to simulate traffic or event failure on your production workloads.

- **Anticipate failure**
    - Perform post-mortems on system failures to better improve, write test code, kill production serves to test recovery.

- **Learn from all operational failures**
    - Share lessons learned in a knowledge base for operational events & failures across your entire organisation.

Security

- **Implement a strong identity foundation**
    - Implement the Principle of Least Privilege (PoLP). Use Centralised identity. Avoid Long-lived credentials.

- **Enable traceability**
    - Monitor alert & audit actions & changes to your environment in real-time Integrate log & metric collection & automate investigation & remediation.

- **Apply security at all layers**

    - Take Defence-in-depth approach with multiple security controls for everything eg. Edge Network, VPC, Load Balancing Instances, OS, Application Code.

- **Automate security best practices**

- **Protect data in transit & at rest**

- **Keep people away from data**

- **Prepare for security events**
    - Incident management systems & investigation policy & processes. Tools to detect, investigate & recover from incidences.

Reliability

- **Automatically recover from failure**
    - Monitor Key Performance Indicators (KPIs) & trigger automation when a threshold is breached.

- **Test recovery procedures**
    - Test how your workload fails, & you validate your recovery procedures. You can use automation to simulate different failures or to recreate scenarios that led to failures before.

- **Scale horizontally to increase aggregate system availability**
    - Replace one large resource with multiple small resources to reduce the impact of a single failure on the overall workload. Distribute requests across multiple, smaller resources to ensure that they don't share a common point of failure.

- **Stop guessing capacity**
    - On-premises it takes a lot of guesswork to determine the elasticity of your workload demands. With Cloud, you don't need to guess how much you need because you can request the right size of resources on-demand.

- **Manage change in automation**
    - Making changes via Infrastructure as Code will allow for a formal process to track & review infrastructure.

Performance Efficiency

- **Democratise advanced technologies**
    - Focus on product development rather than procurement, provisioning, & management of services. Take advantage of advanced technology specialised & optimised for your use-case with on-demand cloud services.

- **Go global in minutes**
    - Deploying your workload in multiple AWS Regions around the world allows you to provide lower latency & a better experience for your customers at minimal cost.

- **Use serverless architectures**
    - Serverless architectures remove the need for you to run & maintain physical servers for traditional compute activities. Removes the operational burden of managing physical servers, & can lower transaction costs because managed services operate at cloud scale.

- **Experiment more often**
    - With virtual & automatable resources, you can quickly carry out comparative testing using different types of instances, storage, or configurations.

- **Consider mechanical sympathy**
    - Understand how cloud services are consumed & always use the technology approach that aligns best with your workload goals. For example, consider data access patterns when you select database or storage approaches.

Cost Optimization

- **Implement Cloud Financial Management**
    - Dedicate time & resources to build capability Cloud Financial Management & Cost Optimization tooling.

- **Adopt a consumption model**
    - Pay only for the computing resources that you require & increase or decrease usage depending on business requirements

- **Measure overall efficiency**

    - Measure the business output of the workload & the costs associated with delivering it. Use this measure to know the gains you make from increasing output & reducing costs.

- **Stop spending money on undifferentiated heavy lifting**
    - AWS does the heavy lifting of data centre operations like racking, stacking, & powering servers. It also removes the operational burden of managing operating systems & applications with managed services. This allows you to focus on your customers & business projects rather than on IT infrastructure.

- **Analyse & attribute expenditure**
    - The cloud makes it easier to accurately identify the usage & cost of systems, which then allows transparent attribution of IT costs to individual workload owners. This helps measure return on investment (ROI) & gives workload owners an opportunity to optimise their resources & reduce costs.

Sustainability

- **Measure & Monitor Resource Consumption**
    - Track energy usage, carbon emissions, & resource utilisation with AWS tools like Cost Explorer, CloudWatch, & Sustainability Hub.
    - Establish baselines & set targets for sustainability improvements.

- **Optimise Resource Provisioning**
    - Right-size instances & services to match workloads, avoiding over-provisioning.
    - Utilise Auto Scaling to dynamically adjust capacity based on demand.
    - Employ serverless technologies like AWS Lambda to eliminate idle resources.
    - Consider managed services that handle optimization for you.

- **Leverage Renewable Energy**
    - Choose AWS regions powered by renewable energy sources.
    - Track your renewable energy use through AWS tools & reports.
    - Participate in AWS programs like the Amazon Sustainability Data Initiative.

- **Design for Energy Efficiency**
    - Architect applications & workloads with energy efficiency in mind.

    - Use efficient storage options like S3 Glacier & S3 Intelligent-Tiering.
    - Enable automatic sleep & hibernation for inactive resources.
    - Adopt efficient coding practices & algorithms.

- **Extend the Lifespan of Resources**
    - Prioritise reuse & repurposing of cloud resources.
    - Explore lifecycle management strategies.
    - Consider extended warranties & service contracts for hardware.

- **Adopt Sustainable Practices**
    - Implement policies for reducing waste & promoting sustainability.
    - Educate employees on sustainable cloud practices.
    - Partner with AWS to achieve sustainability goals through programs & initiatives.

AWS Well-Architectured Tool

The Well-Architected Tool is **an auditing tool** to be used to asset your cloud workloads for alignment with the AWS Well-Architected Framework.

It's essentially a checklist, with nearby references to help you assemble a report to share with executives & key stakeholders.

AWS Architecture Center

The AWS Architecture Center is a web portal that contains best practices & reference architectures for a variety of different workloads.

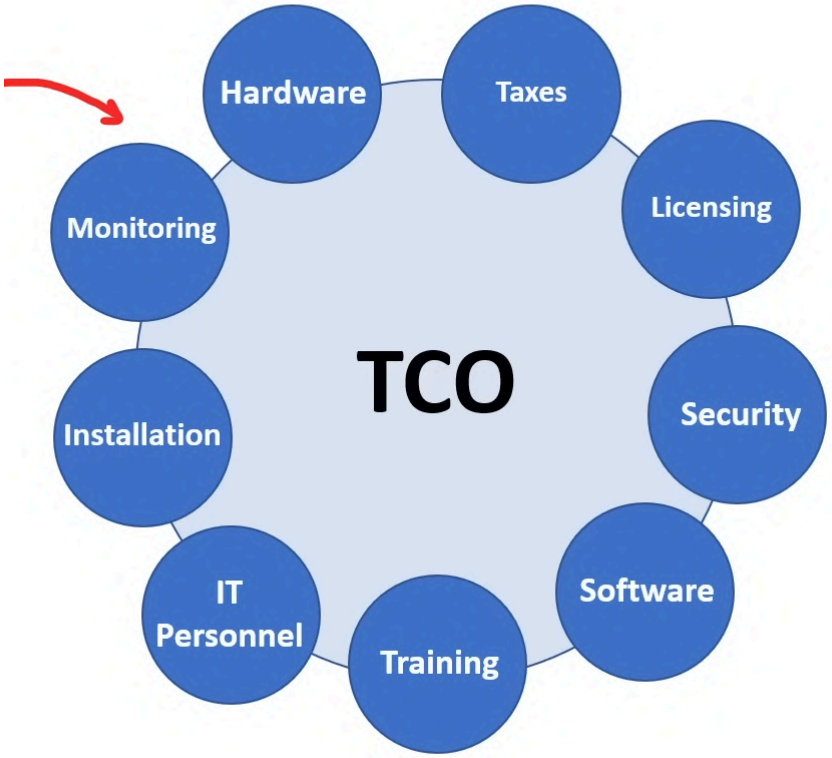
AWS Architecture Centre

Total Cost of Ownership (TCO) & Migration

Total Cost of Ownership (TCO)

TCO is a financial estimate intended to help buyers & owners determine the direct & indirect costs of a product or service.

Creating a TCO report is useful when your company is looking to migrate from on-premise to cloud.



According to research stated by Garter:
"Cloud services can initially be more expensive than running on-premises data centres. [However, it also proves that] cloud services can become cost-effective over time if organisations learn to use & operate them more efficiently."

Capital Expenditure (CAPEX)	Operational Expenditure (OPEX)
On-Premise	**AWS (75% Savings)**
Software Licence Fees	Subscription Fees
• Implementation • Configuration	• Implementation • Configuration

• Training • Physical Security • Hardware • IT Personal • Maintenance	• Training

Capital vs Operational Expenditure

Capital Expenditure (CAPEX)	Operational Expenditure (OPEX)
Spending money upfront on **physical infrastructure** deducts that expense from your tax bill over time.	The costs associated with an on-premises data centre that has shifted the cost to the service provider. The customer only has to be concerned with **non-physical costs**.

• Server Costs. (computers)
• Storage Costs. (hard drives)
• Network Costs. (Routers, Cables, Switches)
• Backup & Archive Costs.
• Disaster Recovery Costs.
• Datacenter Costs. (Rent, Cooling, Physical Security)
• Technical Personal.
	• Leasing Software & Customising features.
	• Training Employees in Cloud Services.
	• Paying for Cloud Support.
	• Billing based on cloud metrics
	  ○ Compute usage.
	  ○ Storage usage.

With Capital Expenses, **you have to guess upfront** what you plan to spend.	With Operational Expenses you can try a product or service **without investing in equipment**.

Shifting IT Personnel

Does Cloud Make IT Personnel Redundant?

A company is considering migrating its workloads from on-premise to the cloud to take advantage of the savings. There is a concern among the staff that there will be mass layoffs.

Shifting your IT Team

- A company needs IT personnel during the migration phase
- A company can transition some roles to new cloud roles:

    - Networking to Cloud Networking
- A company may decide to take a Hybrid approach so they'll always need to have a traditional IT team & a Cloud IT Team
- A company can change employees activities from **Managing Infrastructure to Revenue Generating**.

AWS Pricing Calculator

The **AWS Pricing Calculator** is a **free cost estimate tool** that can be used within your **web browser** without the need for an AWS Account to estimate the cost of various AWS services.

Calculator

The AWS Pricing Calculator contains 100+ services that you can configure for a cost estimate.

To calculate the Total Cost of Ownership an organisation needs to compare their existing cost against the AWS costs & so the AWS Pricing Calculator can be used to determine that cost.

You can export your final estimate to a CSV.

Migration Evaluator

AWS Migration Evaluator (formally known as TSO Logic) is an **estimating tool** used to determine an organisation existing on-premise cost so it can compare it against AWS Costs for planned cloud migration

Migration Evaluator uses an **Agentless Collector** to collect data from your on-premise infrastructure to extract your on-premise costs

EC2 VM Import / Export

VM Import/Export allows users **to import Virtual Machine images into EC2**.

AWS has import instructions for:

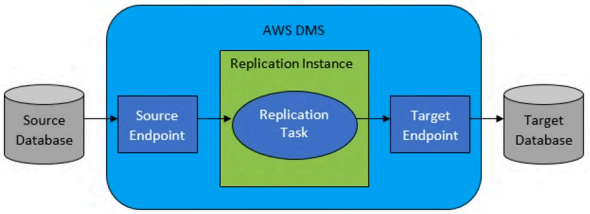
- VMWare
- Citrix
- Microsoft Hyper-V
- Windows VHD from Azure
- Linux VHD from Azure

Database Migration Service

AWS Database Migration Service (DMS) allows you to quickly & securely migrate one database to another.
DMS can be used to migrate your on-premise database to AWS.

Possible Sources

- Oracle Database
- Microsoft SQL
- MySQL
- MariaDB
- PostgreSQL
- MongoDB
- SAP ASE
- IMDB Db2
- Azure SQL Database
- Amazon RDS
- Amazon S3 (database dumps)
- Amazon Aurora
- Amazon DocumentDB



AWS Schema Conversion Tool is used in many cases to automatically convert a source database schema to a target database schema.
Each migration path requires a bit of research since not all combinations of sources & targets are possible.

Possible Targets:

- Oracle Database
- Microsoft SQL
- MySQL
- MariaDB
- PostgreSQL

- Redis
- SAP ASE
- Amazon Redshift
- Amazon RDS
- Amazon DynamoDB
- Amazon S3
- Amazon Aurora
- Amazon OpenSearch Service
- Amazon ElastiCache for Redis
- Amazon DocumentDB
- Amazon Neptune
- Apache Kafka

AWS Cloud Adoption Framework (CAF)

The AWS Cloud Adoption Framework is a whitepaper to help you plan your migration from on-premise to AWS.

At the highest level, the AWS CAF organises guidance into **six focus areas**.

1. **Business Perspective**
    - e.g. Business Managers, Finance Managers, Budget Owners, & Strategy Stakeholders.
    - How to update the staff skills & organisational processes to optimise business value as they move ops to the cloud

2. **People Perspective**
    - e.g. Human Resources, Staffing, People Managers.
    - how to update the staff skills & organisational processes to optimise & maintain their workforce, & ensure competencies are in place at the appropriate time.

3. **Governance Perspective**
    - e.g. CIO, Program Managers, Project Managers, Enterprise Architects, Business Analysts
    - how to update the staff skills & organisational processes that are necessary to ensure business governance in the cloud, & manage & measure cloud investments to evaluate their business outcomes.

4. **Platform Perspective**
    - e.g. CTO, IT Managers, Solution Architects.

- how to update the staff skills & organisational processes that are necessary to deliver & optimise cloud solutions & services.

5. **Security Perspective**
    - e.g. CISO, IT Security Managers, IT Security Analysts.
    - how to update the staff skills & organisational processes that are necessary to ensure that the architecture deployed in the cloud aligns to the organisation's security control requirements, resiliency, & compliance requirements.

6. **Operations Perspective**
    - e.g. IT Operations Managers, IT Support Managers.
    - how to update the staff skills & organisational processes that are necessary to ensure system health & reliability during the move of operations to the cloud & then to operate using agile, ongoing, cloud computing best practices.

Billing & Pricing

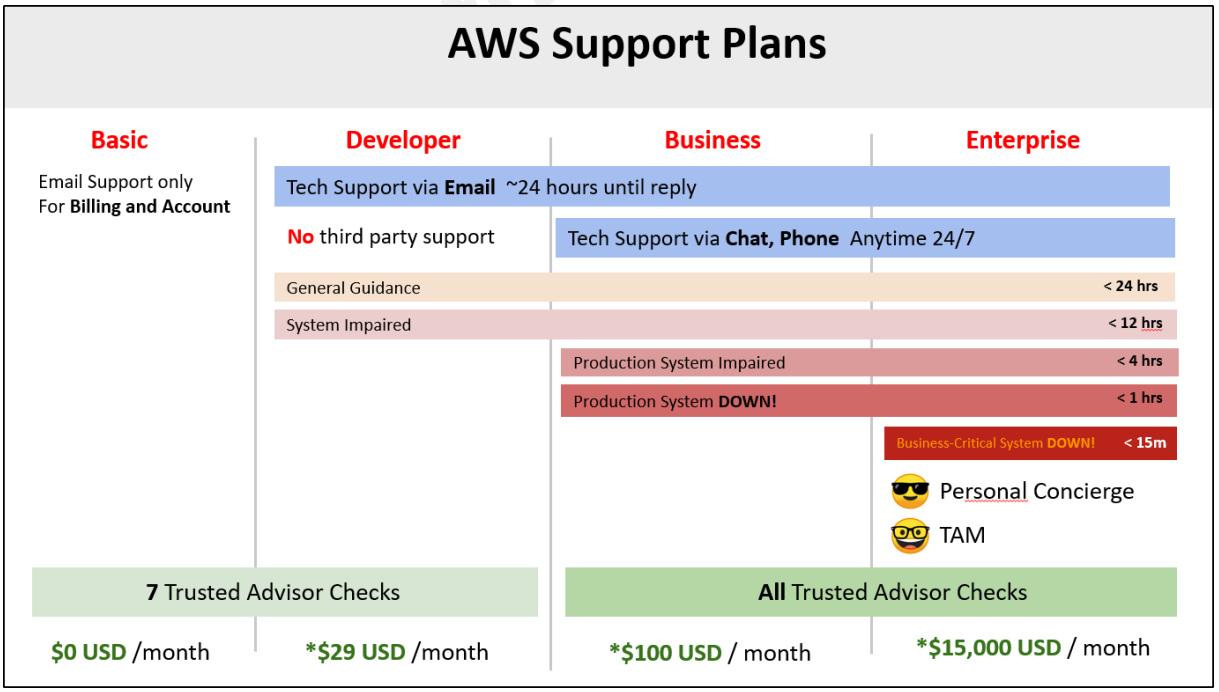
AWS Free Services

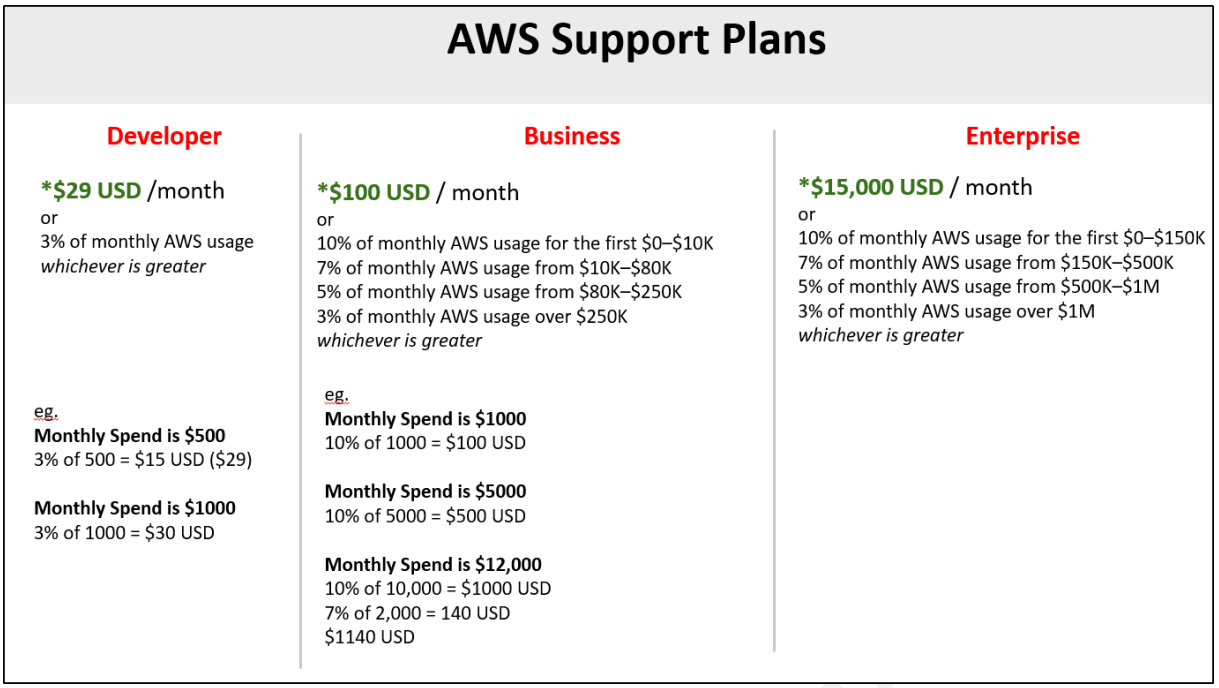
AWS Free services are free forever, unlike the “free-tier” that are up to a point of usage or time.

The AWS services are also free. However these AWS Services provide other services which may cost money.

- IAM - Identity Access Management.
- Amazon VPC.
- Auto Scaling.
- CloudFormation.
- Elastic Beanstalk.
- Opsworks.
- Amplify.
- AppSync.
- CodeStar.
- Organisations & Consolidated Billing.
- AWS Cost Explorer.

AWS Support Plans





Technical Account Manager (TAM)

A Technical Account Manager? (TAM) provides both **proactive guidance and reactive support** to help you succeed with your AWS journey.

What does a TAM do? (Straight from an AWS Job Posting)

- Build solutions, provide technical guidance and advocate for the customer
- Ensure AWS environments remain operationally healthy whilst reducing cost and complexity
- Develop trusting relationships with customers, understanding their business needs and technical challenges
- Using your technical acumen and customer obsession, you'll drive technical discussions regarding incidents, trade-offs, and risk management
- Consult with a range of partners from developers through to C-suite executives
- Collaborates with AWS Solutions Architects, Business Developers, Professional Services Consultants, and Sales Account Managers
- Proactively find opportunities for customers to gain additional value from AWS
- Provide detailed reviews of service disruptions, metrics, detailed pre launch planning
- Being part of a wider Enterprise Support team providing post-sales, consultative expertise

- Solve a variety of problems across different customers as they migrate their workloads to the cloud
- Uplift customer capabilities by running workshops, brown bag sessions, etc.

TAMs follow the Amazon Leadership Principles Especially about being Customer Obsessed!

TAMs are only available at the Enterprise Support tier.

AWS Marketplace

AWS Marketplace is a curated digital catalogue with **thousands** of software listings from independent software vendors.

Easily find, buy, test, and deploy software that already runs on AWS.

The product can be **free** to use or can have an **associated charge**. The charge becomes part of your AWS bill, and once you pay, AWS Marketplace pays the provider.

The sales channel for ISVs and Consulting Partners allows you to **sell your solutions** to other AWS customers.

Products can be offered as

- Amazon Machine Images (AMIs)
- AWS CloudFormation templates
- Software as a service (SaaS) offerings
- Web ACL
- AWS WAF rules

Consolidated Billing

Consolidated Billing is a feature of AWS Organizations that allows you to pay for multiple AWS accounts with **one bill**.

For billing, AWS treats all the accounts in an organisation as if they were one account.

You can designate one **master account that pays the charges** of all the other **member accounts**.

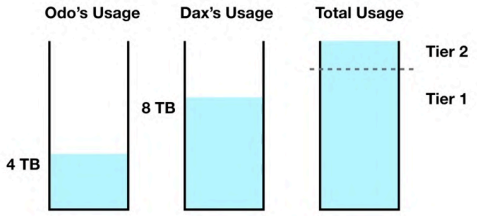
Consolidated billing is offered at no additional cost!

Use **Cost Explorer** to visualise usage for consolidated billing.

You can combine the usage across all accounts in the organisation to share the volume pricing discounts.

Volume Discounts

AWS has **Volume Discounts** for many services.
The more you use, the more you save.
Consolidated Billing lets you take advantage of Volume Discounts.
Consolidate Billing is a feature of AWS Organizations.



Data Transfer	
First 10 TB	$0.17 per GB
Next 40 TB	$0.13 per GB

Odo	(4*1024)*0.17	= $696.32
Dax	(8*1024)*0.17	= $1392.64
Unconsolidated	696.32+1392.64	= **$2088.96**
Consolidated	((10*1024)*0.17)+((2*1024)*0.13)	= **$2007.04**

1 TB = 1024 GB

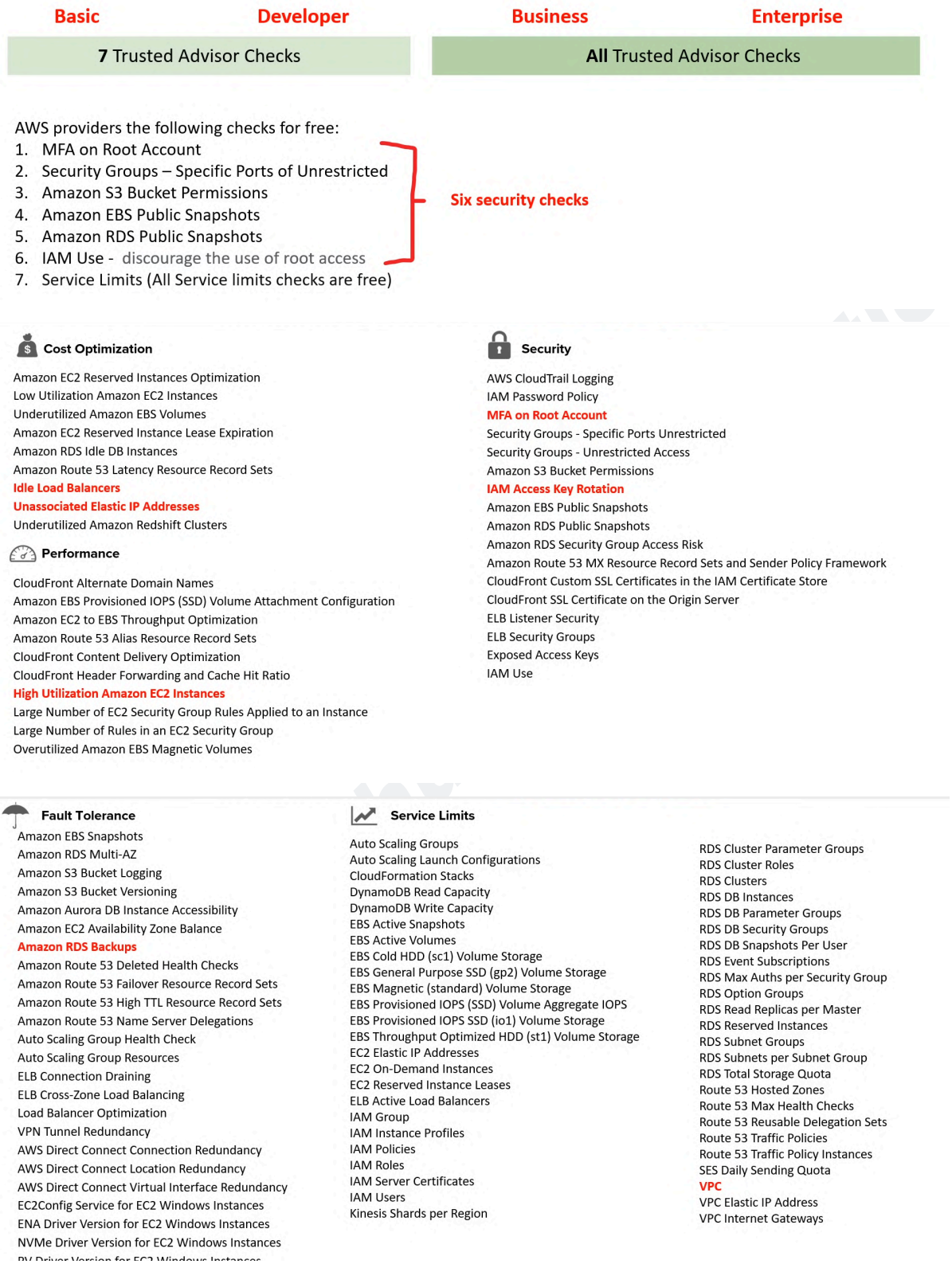
AWS Trusted Advisor

AWS Trusted Advisor is a **recommendation tool** that automatically and actively monitors your AWS account to provide **actionable recommendations** across a series of categories.

Think of AWS Trusted Advisor like an automated checklist of best practices on AWS

The 5 categories of AWS Trusted Advisor

- Cost Optimization – How can we save money?
- Performance – How can we improve performance?
- Security – How can we improve security?
- Fault Tolerance – How can we prevent a disaster or data loss?
- Service Limits – Are we going to hit the maximum limit for a service?



Service Level Agreements

What is a Service Level Agreement (SLA)?

An SLA is a **formal commitment** about the **expected level of service** between a customer and provider.

When a service level is not met and if the Customer meets its obligations under the SLA, Customer will be eligible to receive the compensation eg. **Financial or Service Credits**.

What is a Service Level Indicator (SLI)?
A **metric / measurement** that indicates what measure of performance a customer is receiving at a given time
An SLI metric could be uptime, performance, availability, throughput, latency, error rate, durability, correctness

What is a Service Level Objective (SLO)?
The objective that the provider has agreed to meet
SLOs are represented as a specific **target percentage** over a period of time.

Availability SLA of **99.99%** in a period of **3 months**

Target percentages

- 99.95%
- 99.99%
- 99.999999999% (commonly called **Nine nines**)
- 99.99999999999% (commonly called **Nine elevens**)

Service Health Dashboard

The **Service Health Dashboard** shows the general status of AWS services.

An **icon** and **details** will indicate the status of each AWS Service.

AWS Personal Health Dashboard

AWS Personal Health Dashboard provides **alerts and guidance** for AWS events that might affect your environment.

All AWS customers can access the Personal Health Dashboard.

The Personal Health Dashboard shows recent events to help you manage active events, and shows proactive notifications so that you can plan for scheduled activities

Use these alerts to get notified about changes that can affect your AWS resources, and then follow the guidance to diagnose and resolve issues.

AWS Abuse

AWS Trust & Safety is a team that specifically deals with abuses occurring on the AWS platform for the following issues:

- **Spam -** You are receiving unwanted emails from an AWS-owned IP address, or AWS resources are used to spam websites or forums.
- **Port scanning -** Your logs show that one or more AWS-owned IP addresses are sending packets to multiple ports on your server. You also believe this is an attempt to discover unsecured ports.
- **Denial-of-service (DoS) attacks -** Your logs show that one or more AWS-owned IP addresses are used to flood ports on your resources with packets. You also believe that this is an attempt to overwhelm or crash your server or the software running on your server.
- **Intrusion attempts -** Your logs show that one or more AWS-owned IP addresses are used to attempt to log in to your resources.
- **Hosting prohibited content -** You have evidence that AWS resources are used to host or distribute prohibited content, such as illegal content or copyrighted content without the consent of the copyright holder.
- **Distributing malware -** You have evidence that AWS resources are used to distribute software that was knowingly created to compromise or cause harm to computers or machines that it's installed on.

AWS Support does not deal with Abuse tickets. You need to contact abuse@amazonaws.com or fill out the Report Amazon AWS abuse form.

AWS Free-Tier

AWS has a free-tier that allows you to use AWS at no cost

- For the first 12 months of signup.
- Free usage up to a certain monthly limit forever.

AWS Credits

AWS Promotional Credits (or AWS Credits for short) are the equivalent to USD dollars on the AWS platform.
AWS Credits can be earned in several ways such as

- Joining the AWS Activate startup program.
- Winning Hackathons.
- Participating in Surveys.
- ...

AWS Credits generally have an expiry date attached to them.

AWS Credits can be used for most services but there are exceptions where AWS Credits cannot be used eg. Purchasing a domain via Route53.

AWS Partner Network (APN)

The AWS Partner Network (APN) is a global partner program for AWS. Joining the APN will open your organisation up to business opportunities and allow exclusive training and marketing events

When joining the APN you can either be a:
Consulting Partner – You help companies utilise AWS.
Technology Partner – You build technology on top of AWS as a service offering.

- A partner belongs to a specific Tier: Select, Advanced, or Premier.
- Different tiers have different Annual fee commitments.
- Different tiers have different Knowledge requirements.
- AWS Certification.
- AWS APN-Exclusive Certifications.
- You can get back Promotional AWS Credits.
- You can have unique speaking opportunities in the official AWS marketing channels. Eg blogs, webinars.
- Being part of the APN is a requirement to be a Sponsor with a vendor booth at AWS Events.

AWS Budgets

AWS Budgets give you the ability to set up alerts if you **exceed** or are **approaching** your defined budget.

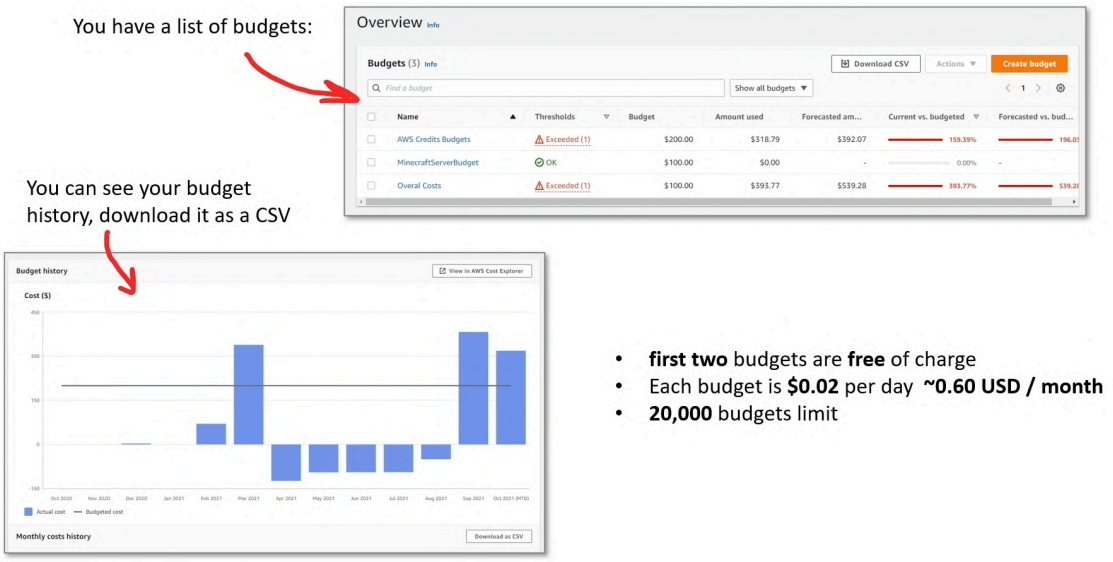
Create **Cost**, **Usage**, or **Reservation** Budgets.

It can be tracked at the **monthly**, **quarterly**, or **yearly levels**, with customizable start and end dates.

Alerts support **EC2**, **RDS**, **Redshift**, and **ElastiCache** reservations.

AWS Budgets can be used to Forecast costs but are limited compared to Cost Explorer or doing your analysis with AWS Cost and Usage Reports along with a Business Intelligence tool.

Budget based on a fixed cost or plan your upfront based on your chosen level
Can be easily managed from the **AWS Budgets** dashboard or via the **Budgets API**.
Get Notified by providing an email or **Chatbot** and threshold how close to the current or forecasted budget.



- **first two** budgets are **free** of charge
- Each budget is **$0.02** per day **~0.60 USD / month**
- **20,000** budgets limit

AWS Budget Reports

AWS Budget Report is used alongside AWS Budgets to create and send daily, weekly, or monthly reports to monitor the performance of your AWS Budget that will be emailed to specific emails.

AWS Budget Reports serve as a more convenient way of staying on top of reports since they are delivered to your email instead of logging into the AWS Management Console.

AWS Cost and Usage Reports (CUR)

Generate a **detailed spreadsheet**, enabling you to **better analyse and understand your AWS costs**.

Places the reports into S3
Use Athena to turn the report into a queryable database
Use QuickSight to visualise your billing data as graphs

Choose the granularity of your data by selecting hourly, daily, or monthly
The report will contain Cost Allocation Tags.
CUR data is stored in a CSV (GZIP) or Parquet format in your selected S3 bucket.

Cost Allocation Tags

Cost Allocation Tags are optional metadata that can be attached to AWS resources so when you generate a Cost and Usage Report you can use that data to better analyse your data.

There are **two types** of tags

- User-Defined : Eg – Project.
- AWS Generated : E.g – aws:createdBy.

You have to **activate** the tags you want to show up in the report.

Billing Alerts/Alarms

You can create your own Alarms in CloudWatch Alarms to monitor spending. They are commonly called “Billing Alarms”.

You first need to turn on **Billing Alerts**.

To create a CloudWatch Alarm, you can choose Billing as your Metric.

Billing Alarms are much more flexible than AWS Budgets and ideal for more complex use-cases for monitoring spending and usage.

AWS Cost Explorer

AWS Cost Explorer lets you **visualise, understand, and manage** your AWS costs and usage **over time**.

Default reports help you gain insight into your cost drivers and usage trends.

Use **forecasting** to get an idea of future costs.

Choose if you want to view your data at a **monthly** or **daily** level of granularity.

Use **filter** and **grouping** functionalities to dig even deeper into your data!

Cost Explorer shows up in **US-East-1.**

AWS Pricing API

With AWS you can programmatically access pricing information to get the latest price offering for services.

There are two versions of this API:

- Query API – The Pricing Service API via **JSON**
    - https://api.pricing.us-east-1.amazonaws.com
- Batch API – The Price List API via **HTML**
    - https://pricing.us-east-1.amazonaws.com/offers/v1.0/aws/index.json

You can also subscribe to Amazon Simple Notification Service (Amazon SNS) notifications to get alerts when prices for the services change.

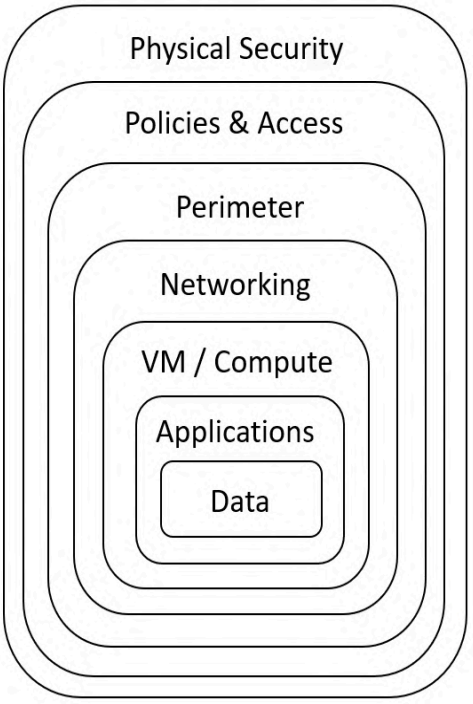
AWS prices change periodically, such as when AWS cuts prices when new instance types are launched, or when new services are introduced.

Security

Defence-In-Depth

The 7 Layers of Security

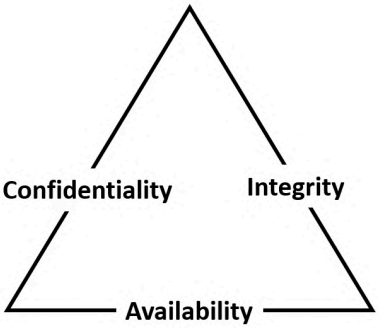
1. **Data -** Access to business and customer data and encryption to protect data.
2. **Application -** Applications are secure and free of security vulnerabilities.
3. **Compute -** Access to virtual machines (ports, on-premise, cloud)
4. **Network -** Limit communication between resources using segmentation and access controls.
5. **Perimeter -** Distributed denial of service (DDoS) protection to filter large-scale attacks before they can cause a denial of service for users.
6. **Identity and Access -** Controlling access to infrastructure and change control.
7. **Physical -** Limiting access to a data centre to only authorised personnel.



Confidentiality, Integrity, Availability (CIA)

Confidentiality, Integrity, and Availability (CIA) triad is a model describing the foundation of security principles and their trade-off relationship.

Confidentiality
Confidentiality is a component of privacy that is implemented to protect our data from unauthorised viewers. In practice, this can be using cryptographic keys to encrypt our data and using keys to encrypt our keys (envelope encryption)



Integrity
Maintaining and assuring the accuracy and completeness of data over its entire lifecycle. In

Practice utilising ACID-compliant databases for valid transactions. Utilising tamper-evident or tamper-proof Hardware security modules. (HSM)

Availability
Information needs to be made be available when needed In Practice: High Availability, Mitigating DDoS, Decryption access

The CIA triad was first mentioned in a **NIST publication from 1977**.

There have been efforts to expand and modernise or suggest alternatives to CIA triad:

- (1998) Six Atomic Elements of Information i.e. confidentiality, possession, integrity, authenticity, availability, and utility
- (2004) NIST Engineering Principles for Information Technology Security — 33 security principles

Vulnerabilities

What is a vulnerability?

A hole or a weakness in the application, which can be a design flaw or an implementation bug, that allows an attacker to cause harm to the stakeholders of an application.

Allowing Domains or Accounts to Expire
Buffer Overflow
Business logic vulnerability
CRLF Injection
CSV Injection
Catch NullPointerException
Covert storage channel
Deserialization of untrusted data
Directory Restriction Error
Doubly freeing memory
Empty String Password
Expression Language Injection
Full Trust CLR Verification issue
Heartbleed Bug
Improper Data Validation
Improper pointer subtraction
Information exposure through query strings
Injection problem
Insecure Compiler Optimization
Insecure Randomness
Insecure Temporary File
Insecure Third Party Domain Access
Insecure Transport
Insufficient Entropy
Insufficient Session-ID Length
Least Privilege Violation
Memory leak
Missing Error Handling
Missing XML Validation
Multiple admin levels
Null Dereference
OWASP .NET Vulnerability Research
Overly Permissive Regular Expression
PHP File Inclusion
PHP Object Injection
PRNG Seed Error
Password Management Hardcoded Password
Password Plaintext Storage
Poor Logging Practice
Portability Flaw
Privacy Violation
Process Control
Return Inside Finally Block
Session Variable Overloading
String Termination Error
Unchecked Error Condition
Unchecked Return Value Missing Check against Null
Undefined Behavior
Unreleased Resource
Unrestricted File Upload
Unsafe JNI
Unsafe Mobile Code
Unsafe function call from a signal handler
Unsafe use of Reflection
Use of Obsolete Methods
Use of hard-coded password
Using a broken or risky cryptographic algorithm
Using freed memory
Vulnerability template
XML External Entity (XXE) Processing

Encryption

What is cryptography?

The practice and study of techniques for secure communication in the presence of third parties called adversaries.

What is encryption?

The process of encoding (scrabbling) information **using a key** and a **cypher** to store sensitive data in an unintelligible format as a means of protection. An encryption takes in plaintext and produces **ciphertext**.

The **enigma machine** was used during WW2. A different key for each day was used to set the position of the rotors. It relied on simple cypher substitution.

Cypher

What is a cypher?
An algorithm that performs encryption or decryption. Cipher is synonymous with “code”.

A Cipher (also spelled "cypher") is a cryptographic technique that encrypts information to ensure its confidentiality and integrity. It scrambles data, making it unreadable without the correct decryption key or algorithm.

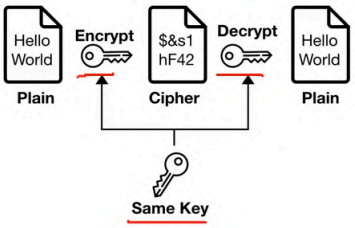
What is ciphertext?
Ciphertext is the result of encryption performed on plaintext via an algorithm. Ciphertext is the encrypted form of a message or data resulting from applying encryption to plaintext.

A **codebook** is a type of document used for gathering and storing cryptography codes.

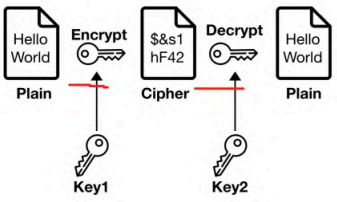
Cryptographic Keys

What is a Cryptographic Key?
A key is a variable used in conjunction with an encryption algorithm in order to encrypt or decrypt data.

1. **What is Symmetric Encryption?**
   The same key is used for encoding and decoding
   eg. **Advanced Encryption Standard (AES)**



2. **What is Asymmetric Encryption?**
   Two keys are used. One to encode and one to decode eg. **Rivest–Shamir–Adleman (RSA)**



Hashing and Salting

What is hashing function?
A function that accepts arbitrary size value and maps it to a fixed-size data structure. Hashing can reduce the size of the store value.

Hashing is a **one-way process** and is **deterministic**.

A deterministic function always returns the same output for the same input.

Hashing Passwords
Hashing functions are used to store passwords in the database so that a password does not reside in a plaintext format.

To authenticate a user, when a user inputs their password, it is hashed, and the hash is compared to the store hashed. If they match then the user has successfully logged in.

Popular hashing functions are **MD5, SHA256, and Bcrypt**.
If an attacker knows what function you are using and stole your database, they could enumerate a dictionary of passwords to determine the password.

Salting Passwords
A salt is a random string not known to the attacker that the hash function accepts to mitigate the deterministic nature of hashing functions.

Digital Signatures and Signing

What is a digital signature?
A mathematical scheme for verifying the authenticity of digital messages or documents.

A Digital signature gives us **tamper-evidence**.

- Did someone mess (modify) the data?
- Is this data not from the expected sender?

There are three algorithms for digital signatures:

- **Key Generation** – generates a public and private key.
- **Signing** – the process of generating a digital signature with a private key and inputted message.
- **Signing Verification** – verify the authenticity of the message with a public key.

SSH uses a public and private key to authorise remote access into a remote machine e.g. Virtual Machine. It is common to use RSA.
ssh-keygen is a **well-known command** to generate a public and private key.

What is Code Signing?
When you use a digital signature to ensure **computer code** has not been tampered.

In-Transit vs At-Rest Encryption

Encryption In-Transit	Encryption At-Rest
Data that is secure when moving between locations	Data that is secure when residing on storage or within a database
Algorithms: **TLS, SSL**	Algorithms: **AES, RSA**

Transport Layer Security (TLS)
An encryption protocol for data integrity between two or more communicating computer applications.

- TLS 1.0, 1.1 are deprecated. **TLS 1.2** and **1.3** is the current best practice.

Secure Sockets Layers (SSL)
An encryption protocol for data integrity between two or more communicating computer applications.

- SSL 1.0, 2.0 and 3.0 are deprecated.

Compliance Programs

Compliance Programs
A set of internal policies and procedures of a company to comply with laws, rules, and regulations or to uphold the business reputation.

Common Compliance Programs

1. **Health Insurance Portability and Accountability Act** (1996) is United States legislation that provides data privacy and security provisions for safeguarding medical information.

2. **The Payment Card Industry Data Security Standard (PCI DSS)** for selling things online and handling credit card information.

3. **International Organization for Standardization (ISO) / International Electrotechnical Commission**
    - ISO/IEC 27001 — control implementation guidance.
    - ISO/IEC 27017 — enhanced focus on cloud security.
    - ISO/IEC 27018 — protection of personal data in the cloud. eg. PII.
    - ISO/IEC 27701 — Privacy Information Management System (PIMS) framework.

   Outlines controls and processes to manage data privacy and protect PII.

4. **System and Organization Controls (SOC)**
    - SOC 1 — 18 standard and report on the effectiveness of internal controls (SSAE) at a service organisation.
        - Relevant to their client's internal control over financial reporting (ICFR).
    - SOC 2 — evaluates internal controls, policies, and procedures that directly relate to the security of a system at a service organisation.
    - SOC 3 — A report based on the Trust Services Criteria that can be freely distributed.

5. **Payment Card Industry Data Security Standard (PCI DSS)**
   A set of security standards designed to ensure that ALL companies that accept, process, store or transmit credit card information maintain a secure environment.

6. **Federal Information Processing Standard (FIPS) 140-2**
   US and Canadian government standard that specifies the security requirements for cryptographic modules that protect sensitive information.

7. **Personal Health Information Protection Act (PHIPA)**
   An Ontario provincial law (Canada) that regulates patient Protected Health Information.

8. **Health Insurance Portability and Accountability Act (HIPAA)**
   US federal law that regulates patient Protected Health Information.

9. **Cloud Security Alliance (CSA) STAR Certification**
   Independent third-party assessment of a cloud provider's security posture.

10. **Federal Risk and Authorization Management Program (FedRAMP)**

US government standardised approach to security authorizations for Cloud Service Offerings

11. **Criminal Justice Information Services (CJIS)**
    Any US state or local agency that wants to access the FBI's CJIS database is required to adhere to the CJIS Security Policy.

12. **General Data Protection Regulation (GDPR)**
    A European privacy law. Imposes new rules on companies, government agencies, non-profits, and other organisations that offer goods and services to people in the European Union (EU) or collect and analyse data tied to EU residents. I

Penetration Testing

What is PenTesting?
An authorised simulated cyberattack on a computer system, performed to evaluate the security of the system.

Pen Testing **is allowed** to be performed on AWS!

Permitted Services

- Amazon EC2 instances.
- NAT Gateways.
- Elastic Load Balancers.
- Amazon RDS.
- Amazon CloudFront.
- Amazon Aurora.
- Amazon API Gateways.
- AWS Lambda and Lambda Edge functions.
- Amazon Lightsail resources.
- Amazon Elastic Beanstalk environments.

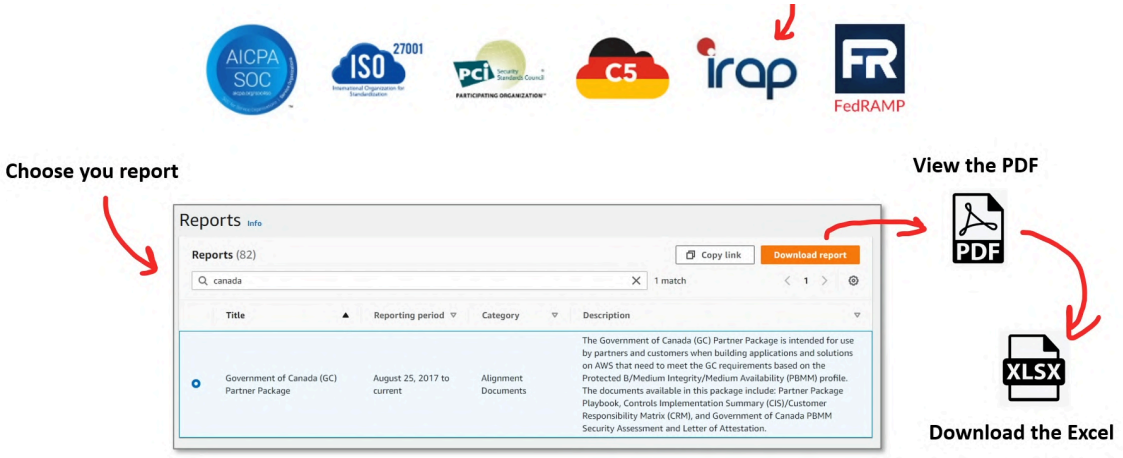
Prohibited Activities

- **DNS zone walking via Amazon Route 53 Hosted Zones**.
- Subject to the **DDoS Simulation Testing policy**.
    - Denial of Service (DoS).
    - Distributed Denial of Service (DDoS).
    - Simulated DoS, Simulated DDoS.
- Port flooding.
- Protocol flooding.
- Request flooding (login request flooding, API request flooding)

For Other Simulated Events you will need to submit a request to AWS. A reply could take up to 7 days.

AWS Artifact

AWS Artifact is a self-serve portal for on-demand access to **AWS compliance reports**.



AWS Inspector

What is Hardening?

The act of eliminating as many **security** risks as possible. Hardening is common for Virtual Machines where you run a collection of security checks known as a security benchmark

AWS Inspector runs a **security benchmark** against specific EC2 instances.
You can run a variety of security benchmarks.
Can perform both **Network** and **Host** Assessments.

- Install the AWS agent on your EC2 instances.
- Run an assessment for your assessment target.
- Review your findings and remediate security issues.

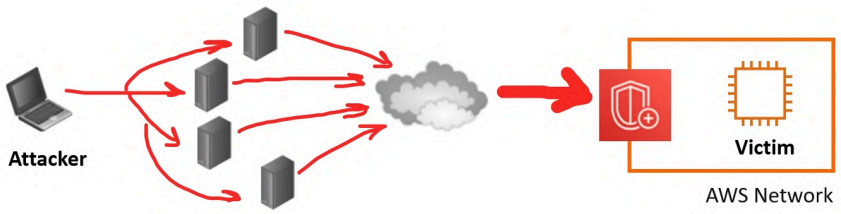
One very popular benchmark you can run is by CIS which has 699 checks!

CIS - Center for Internet Security

Distributed Denial of Service (DDoS)

What is a DDoS (Distributed Denial of Service) Attack?

A malicious attempt to disrupt normal traffic by flooding a website with large amounts of fake traffic.



AWS Shield

AWS Shield is a **managed** DDoS (Distributed Denial of Service) protection service that safeguards applications running on AWS.

When you route your traffic through Route 53 or CloudFront you are using AWS Shield Standard.

Protects you against **Layer 3**, **4** and **7** attacks

- 7 Physical.
- 4 Transport.
- 3 Network.

Shield Standard vs Advanced

Shield Standard (FREE)	Shield Advanced (*3000 USD / Year)
Protection against most common DDoS attacks	Additional protection against larger and more sophisticated attacks
• Access to tools and best practices to build a DDoS resilient architecture. • Automatically available on all AWS services.	Available on : • Amazon Route 53 • Amazon CloudFront • Elastic Load Balancing • AWS Global Accelerator • Elastic IP (Amazon EC1 and Network Load Balancer)

	Notable Features
	• Visibility and Reporting on Layer 3,4 and 7
	  ○ 7 Application
	  ○ 4 Transport
	  ○ 3 Network

	• Access to Team and Support (with Business or Enterprise Support) • DDoS Cost Protection • Comes with SLA

Both plans integrate with AWS Web Application Firewall (WAF) to give you Layer 7 (Application) protection.

AWS Guard Duty

What is an Intrusion Detection System(IDS) / Intrusion Protection System (IPS)?
A device or software application that monitors a network or systems for malicious activity or policy violations.

Guard Duty is a **threat detection service** that continuously monitors for malicious, suspicious activity and unauthorised behaviour.
It uses Machine Learning to analyse the following AWS logs:

- **CloudTrail Logs**
- **VPC Flow Logs**
- **DNS logs**

It will alert you of Findings which **you can automate an incident response via CloudWatch Events or with 3rd Party Services.**

Amazon Macie

Macie is a fully managed service that continuously monitors **S3 data access**
It's a piece of hardware designed to store encryption keys.
HSM holds keys in memory and never writes them to disk.

Federal Information Processing Standard (FIPS)
US and Canadian government standard that specifies the security requirements for cryptographic modules that protect sensitive information.

HSM's that are **multi-tenant** is **FIPS 140-2 Level 2 Compliant**
(multiple customers virtually isolated on an HSM)


eg. AWS KMS

HSM's that are **single-tenant** are **FIPS 140-2 Level 3 Compliant**
(single customer on a dedicated HSM)


eg. AWS CloudHSM

AWS Key Management System

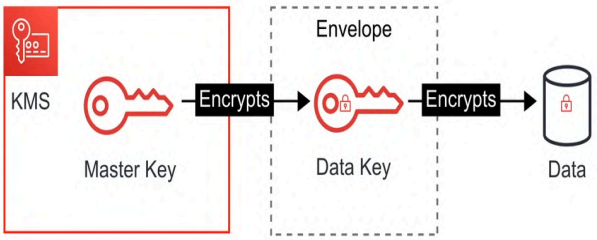
AWS Key Management Service (KMS) is a managed service that makes it easy for you to create and control the encryption keys used to encrypt your data.

- KMS is a multi-tenant HSM (hardware security module)
- Many AWS services are integrated to use KMS to encrypt your data with a simple checkbox.
- KMS uses Envelope Encryption.

Envelope Encryption

When you encrypt your data, your data is protected, but you have to protect your encryption key.

When you encrypt your data key with a master key as an additional layer of security.



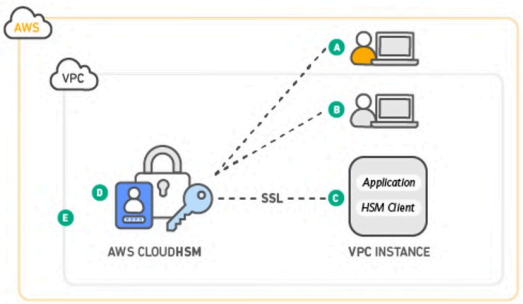
CloudHSM

CloudHSM is a **single-tenant HSM** as a service that automates hardware provisioning, software patching, high availability, and backups.

AWS CloudHSM enables you to generate and use your encryption keys on FIPS 140-2 Level 3 validated hardware.

Built on Open HSM industry standards to integrate with:

- PKCS#11.
- Java Cryptography Extensions (JCE).
- Microsoft CryptoNG (CNG) libraries.



You can also transfer your keys to other commercial HSM solutions to make it easy for you to migrate keys on or off of AWS.

Configure AWS KMS to use AWS CloudHSM cluster as a custom key store rather than the default KMS key store.

Variation Study

Know Your Initialisms

- **IAM** - Identity and Access Management
- **S3** - Simple Storage Service
- **SWF** - Simple Workflow Service
- **SNS** - Simple Notification Service
- **SQS** - Simple Queue Service
- **SES** - Simple Email Service
- **SSM** - Simple Systems Manager
- **RDS** - Relational Database Service
- **VPC** - Virtual Private Cloud
- **VPN** - Virtual Private Network
- **CFN** - CloudFormation
- **WAF** - Web Application Firewall
- **MQ** - Amazon ActiveMQ
- **ASG** - Auto Scaling Groups
- **TAM** - Technical Account Manager
- **ELB** - Elastic Load Balancer
- **ALB** - Application Load Balancer
- **NLB** - Network Load Balancer
- **GWLB** - Gateway Load Balancer
- **CLB** - Classic Load Balancer
- **EC2** - Elastic Cloud Compute
- **ECS** - Elastic Container Service
- **ECR** - Elastic Container Repository
- **EBS** - Elastic Block Storage
- **EFS** - Elastic File Storage
- **EMR** - Elastic MapReduce
- **EB** - Elastic Beanstalk
- **ES** - Elasticsearch
- **EKS** - Elastic Kubernetes Service
- **MSK** - Managed Kafka Service
- **RAM** - AWS Resource Manager
- **ACM** - Amazon Certificate Manager
- **PoLP** - Principle of Least Privilege
- **IoT** - Internet of Things
- **RI** - Reserved Instances

AWS Config vs AWS AppConfig

AWS Config	AWS AppConfig
AWS Config is a governance tool for Compliance as Code (CoC).	AWS App Config is used to automate the process of deploying application configuration variable changes to your web application (s).
You can create rules that will check to see if resources are configured the way you expect them to be.	You can write a validator to ensure the changed variable will not break your web-app
If a resource drifts from the expected configuration you are notified or AWS Config can auto-remediate (correct) the configuration back to the expected state	You can monitor deployments and automate integrations to catch errors or rollback.

Simple Notification Service vs Simple Queue Service

Simple Notification Service	Simple Queue Service
Pass Along Messages eg. PubSub	Queue Up Messages, Guaranteed Delivery
Send notifications to subscribers of topics via multiple protocols. eg, HTTP, Email, SQS, SMS	Place messages into a queue. Applications pull queue using AWS SDK • Can retain a message for up to 14 days • Can send them in sequential order or in parallel • Can ensure only one message is sent • Can ensure messages are delivered at least once
SNS is generally used for sending plain text emails which are triggered via other AWS Services. The best example of this is billing alarms.	
Can retry sending in case of failure for HTTPS	
Really good for webhooks, simple internal emails, triggering Lambda functions	Really good for delayed tasks, queueing up emails

Simple Notification Service vs Simple Email Service vs PinPoint vs Workmail

Simple Notification Service	Simple Email Service	Amazon PinPoint	Amazon Workmail
Practical and Internet Emails	Transactional Emails	Promotional Emails	Email Web Client
Send notifications to **subscribers** of **topics** via multiple protocols. eg, HTTP, **Email**, SQS, SMS	Emails that should be triggered based on in-app actions: Signup, Reset Password, Invoices... • A cloud based email service. eg. **SendGrid** • SES sends **HTML emails**, SNS cannot. • SES can receive inbound emails • SES can create Email Templates • Custom domain name email • Monitor your email reputation	Emails for marketing • Create email campaigns • Segment your contacts • Create customer journeys via emails • A/B emailing testing	Similar to Gmail and Outlook. Create company emails, read, write and send emails from a Web Client within AWS Management Console
SNS is generally used for sending **plain text emails** which are triggered via other AWS Services. The best example of this is billing alarms.			
Most exam questions are going to be talking about SNS because lots of services can trigger SNS for notifications. You Need to Know What are **Topics** and **Subscriptions** regarding **SNS**.			
Really good for webhooks, simple internal emails,			

triggering Lambda functions			

Amazon Inspector vs AWS Trusted Advisor

Amazon Inspector	AWS Trusted Advisor
Audits **a single EC2 instance** that you've selected	Trusted Advisor **doesn't generate a PDF** report.
	Gives you a holistic view of recommendations across multiple services and best practices
Generates a report from a long list of security checks i.e 699 checks.	eg. You have open ports on these security groups
	You should enable MFA on your root account when using Trusted advisor.

Connect Names Services

They all have **"Connect"** in the name but they are not related or similar in functionality.

AWS Direct Connect

- A Dedicated Fibre Optics Connection from your DataCenter to AWS.
- Intended for large enterprises with their own data centre and they need an insanely fast and private connection directly to AWS.
- If you need a **secure connection** you need to apply an AWS VPN connection on top of Direct Connect.

Amazon Connect

- Call Center as a Service.
- Get a toll-free number, accept inbound and outbound calls, set up automated phone systems.
- Interactive Voice System. (IVS)

Media Connect

- New Version of Elastic Transcoder, Converts Videos to Different Video Types.

- You have 1000 videos of you and you need to transcode them into different videos format, maybe you need to apply watermarks, or insert an introduction video in front of every video.

AWS Elemental MediaConnect

- A high-quality transport service for live video.
- Get the reliability and security of satellite and fibre combined with the flexibility, agility, and economics of IP-based networks using AWS Elemental MediaConnect.

Elastic Transcoder vs MediaConvert

Both services transcode videos.

Elastic Transcoder	AWS Elemental MediaConvert
The Old Way	The New Way
Elastic Transcoder was the original transcoding service. It may have programmatic APIs or workflows not available in MediaConvert.	MediaConvert is a more robust transcoding service that can perform various operations during transcoding.
It exists due to legacy customers still using the platform.	
• Transcodes videos to streaming formats.	• Transcodes videos to streaming formats • Overlays images • Insert videos clips • Extracts captions data • Robust UI

AWS Artifact vs Amazon Inspector

Both Artifact and Inspector compile out PDFs.

AWS Artifact	Amazon Inspector
Why should an enterprise trust AWS?	How do we know this EC2 instance is Secure? Prove It?

Generates a security report that's based on **Global Compliance Frameworks**	Runs a script that analyses your EC2 instance, then generates a PDF report telling you which security checks passed.
Global Compliance Frameworks like : • Service Organization Control. (SOC) • Payment Card Industry. (PCI)	**Audit tool for the security of EC2 instances.**

Elastic Load Balancer Variants

Elastic Load Balancer (ELB) has 4 different types of possible load balancers.

ALB vs NLB vs GWLB vs CLB

<table>
<tr><th>Application Load Balancer (ALB)</th><th>Network Load Balancer (NLB)</th><th>Gateway Load Balancer (GWLB)</th><th>Classic Load Balancer (CLB)</th></tr>
<tr><td>Layer 7 - HTTP/S.</td><td>Layers 3 and 4 – TCP and UDP.</td><td rowspan="4">When you need to deploy a fleet of third-party virtual appliances that support GENEVE.</td><td>Layers 3,4 and 7.</td></tr>
<tr><td rowspan="2">Routing Rules
• Create rules to change routing based on information found in an HTTP/S request.</td><td>Where extreme performance is required for TCP and TLS traffic.</td><td>Intended for applications that were built within the EC2-Classic network.</td></tr>
<tr><td>Capable of handling millions of requests per second while maintaining ultra-low latencies.</td><td>Doesn't use Target Groups.</td></tr>
<tr><td>Can attach an AWS WAF.</td><td>Optimised for sudden and volatile traffic patterns while using a single static IP address per Availability Zone.</td><td>Retires on Aug 15, 2022.</td></tr>
</table>