

Agenda and Day Wise plan

Day 1 Agenda

- Definitions and Characteristics
- Deployment Models
- Service Models
- Benefits and Challenges of Cloud Computing
- Cloud Service Providers (CSP) and AWS

Day 2 Agenda

- AWS Global Infrastructure
- Compute and Serverless Computing
- Storage Services
- Database Services
- Networking Services
- Sample Architecture

Day 3 Agenda

- Shared Responsibility Models
- User vs CSP Rights and Responsibilities
- AWS Identity and Access Management
- IAM users, Groups and Roles
- KMS and Cloud Trail
- Compliance and Governance

Day 4 Agenda

- Application Deployments
- Elastic Beanstalk
- Auto scaling and Load Balancing
- Health checks, Events and Alarms
- Monitoring
- Getting Started with CLI Tools

Agenda and Day Wise plan

Day 5 Agenda

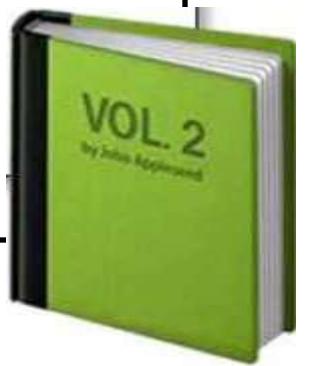
- Capstone Project
- Build and Deploy Web Applications from Code to Infra to Deployments
- Integrations with ELB, CloudMonitoring for secure Web application
- Using S3 as static Web Host
- Custom Metrics and monitoring
- Advance Topics: IoT, ML/AI, Code Pipeline/CodeDeploy

What is Cloud Computing?

cloud com-put-ing

noun

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



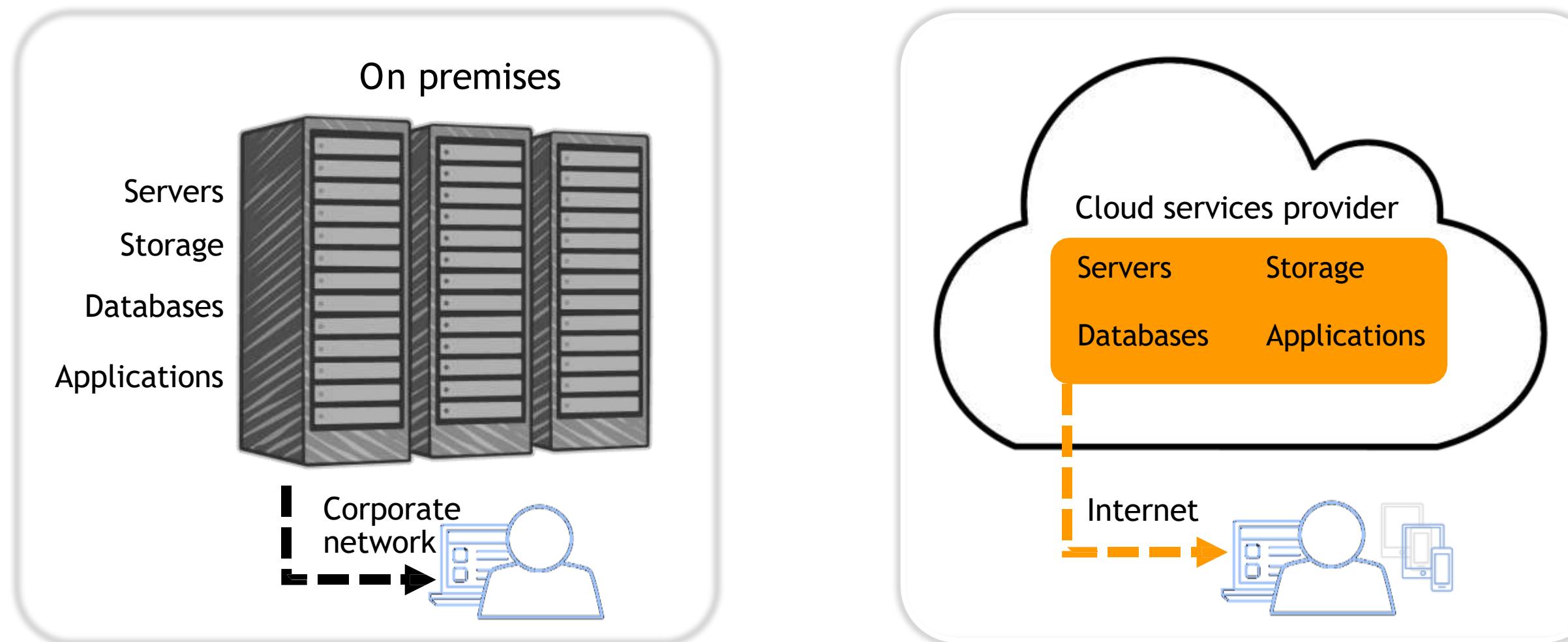
On-Premise

- You own the servers
- You hire the IT people
- You pay or rent the real-estate
- You take all the risk

Cloud Providers

- Someone else owns the servers
- Someone else hires the IT people
- Someone else pays or rents the real-estate
- You are responsible for your configuring cloud services and code, someone else takes care of the rest.

What is Cloud Computing?



What is Cloud Computing?



Everything is virtualized SW

Everything accessible over network

Use and pay as Required.

Services and Resources Managed for you.

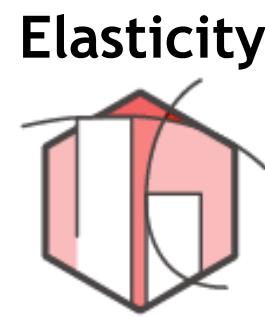
What is Cloud Computing?

With cloud computing, you can stop thinking of your infrastructure as hardware, and instead think of it (and use it) as software.



Agility

On-demand self-service



Elasticity

Scale rapidly to meet demand



Cost Savings

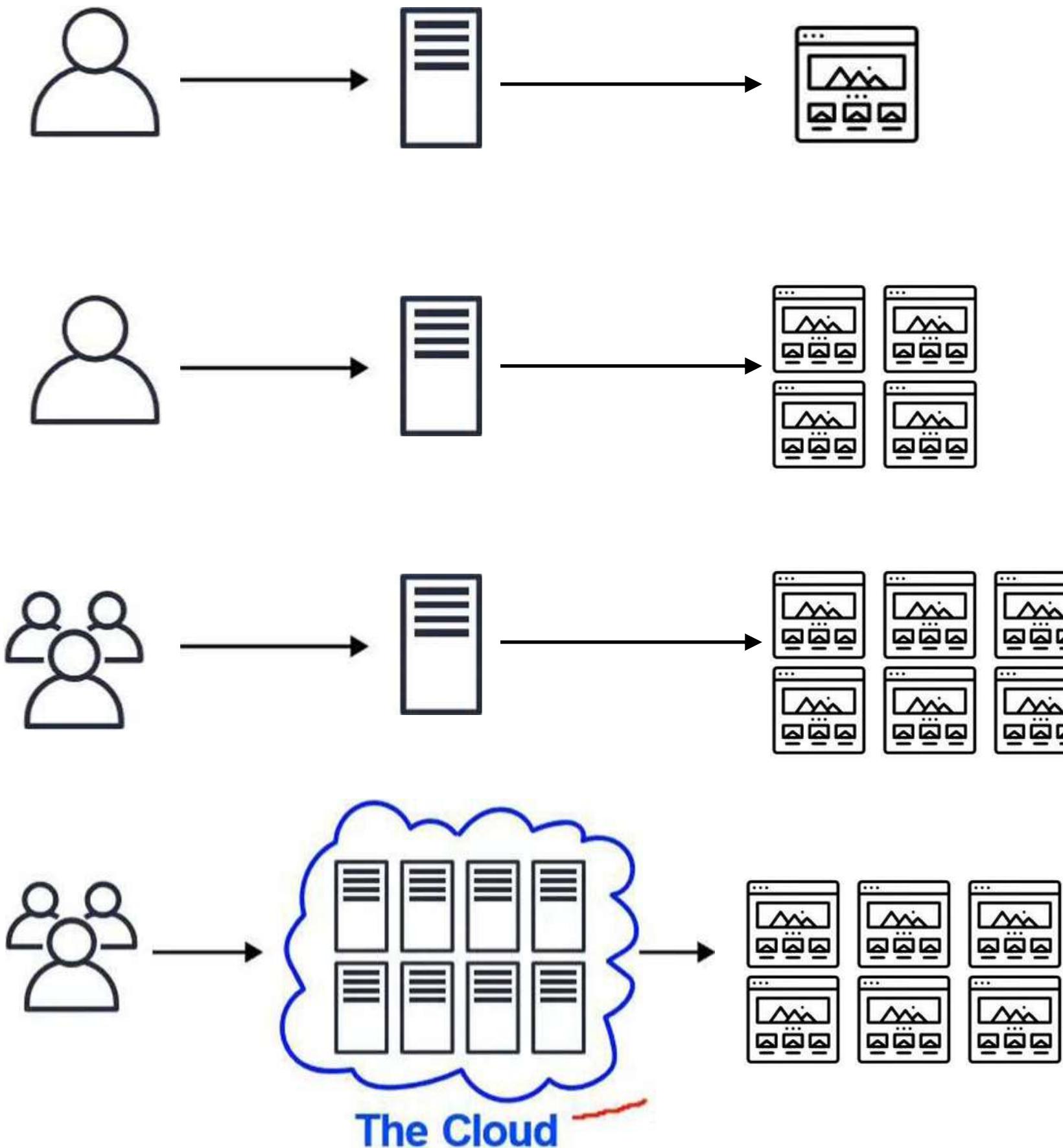
Only pay for IT as you consume it



Deploy globally

Broad network access

The Evolution of Cloud Hosting



Dedicated Server

One physical machine dedicated **to single a business**.

Runs a single web-app/site.

Very Expensive, High Maintenance, *High Security

Virtual Private Server (VPS)

One physical machine dedicated **to a single business**.

The physical machine is virtualized **into sub-machines**

Runs multiple web-apps/sites

Better Utilization and Isolation of Resources

Shared Hosting

One physical machine, shared by **hundred of businesses**

Relies on most tenants under-utilizing their resources.

Very Cheap, Limited functionality, Poor Isolation

Cloud Hosting

Multiple physical machines that act as one system.

The system is abstracted into multiple **cloud services**

Flexible, Scalable, Secure, Cost-Effective, High Configurability

The Evolution of Computing

* Dedicated



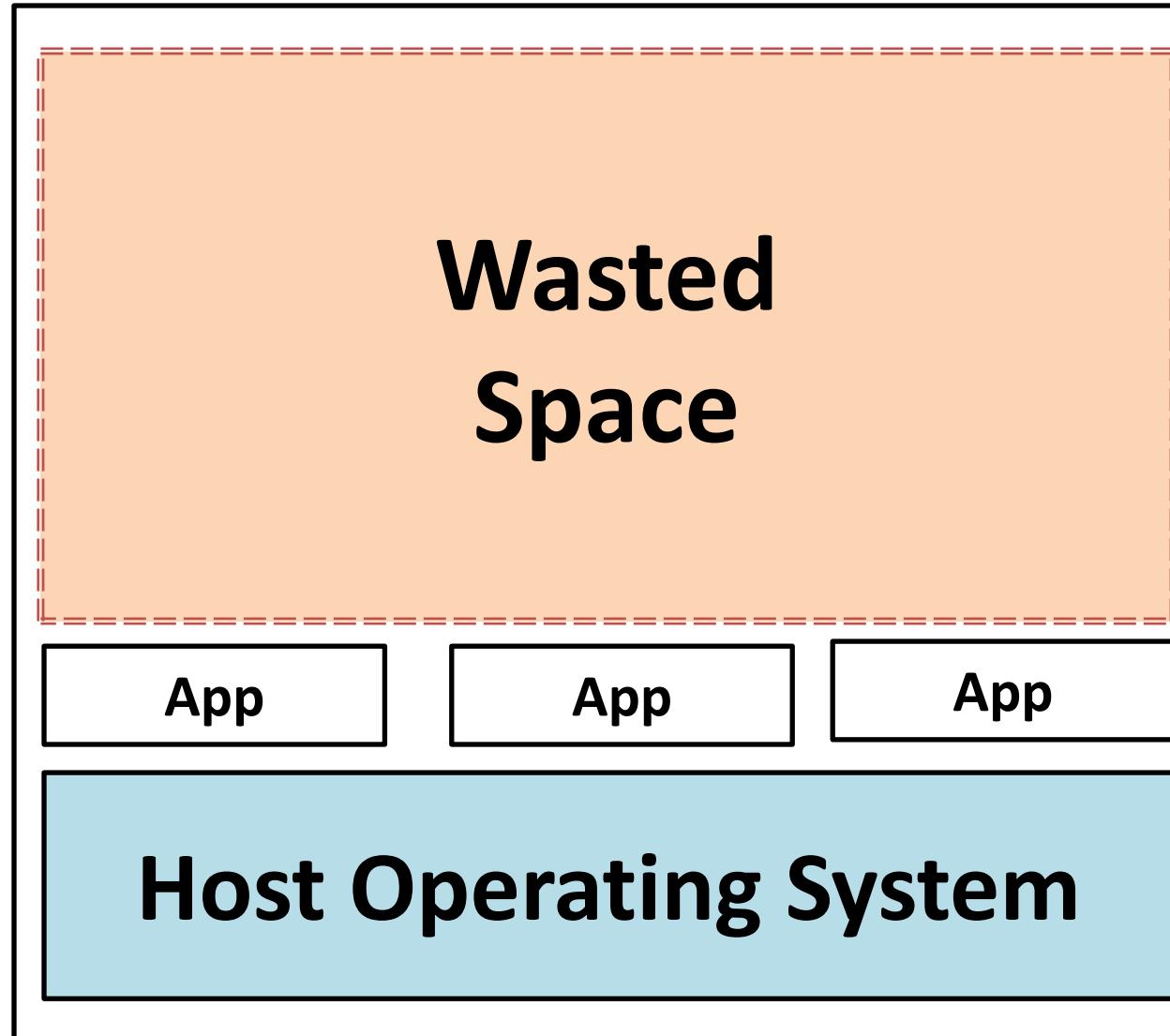
VMs



Containers



Functions



physical server

- A physical server **wholly utilized by a single customer.**
- You have to guess your capacity
- you'll overpay for an underutilized server
- You can't "vertical" scale, you need a manual migration
- Replacing a server is very difficult
- You are limited by your Host Operating System
- Multiple apps can result in conflicts in resource sharing
- You have a ***guarantee of security, privacy, and full utility of underlying resources**

The Evolution of Computing

*Dedicated



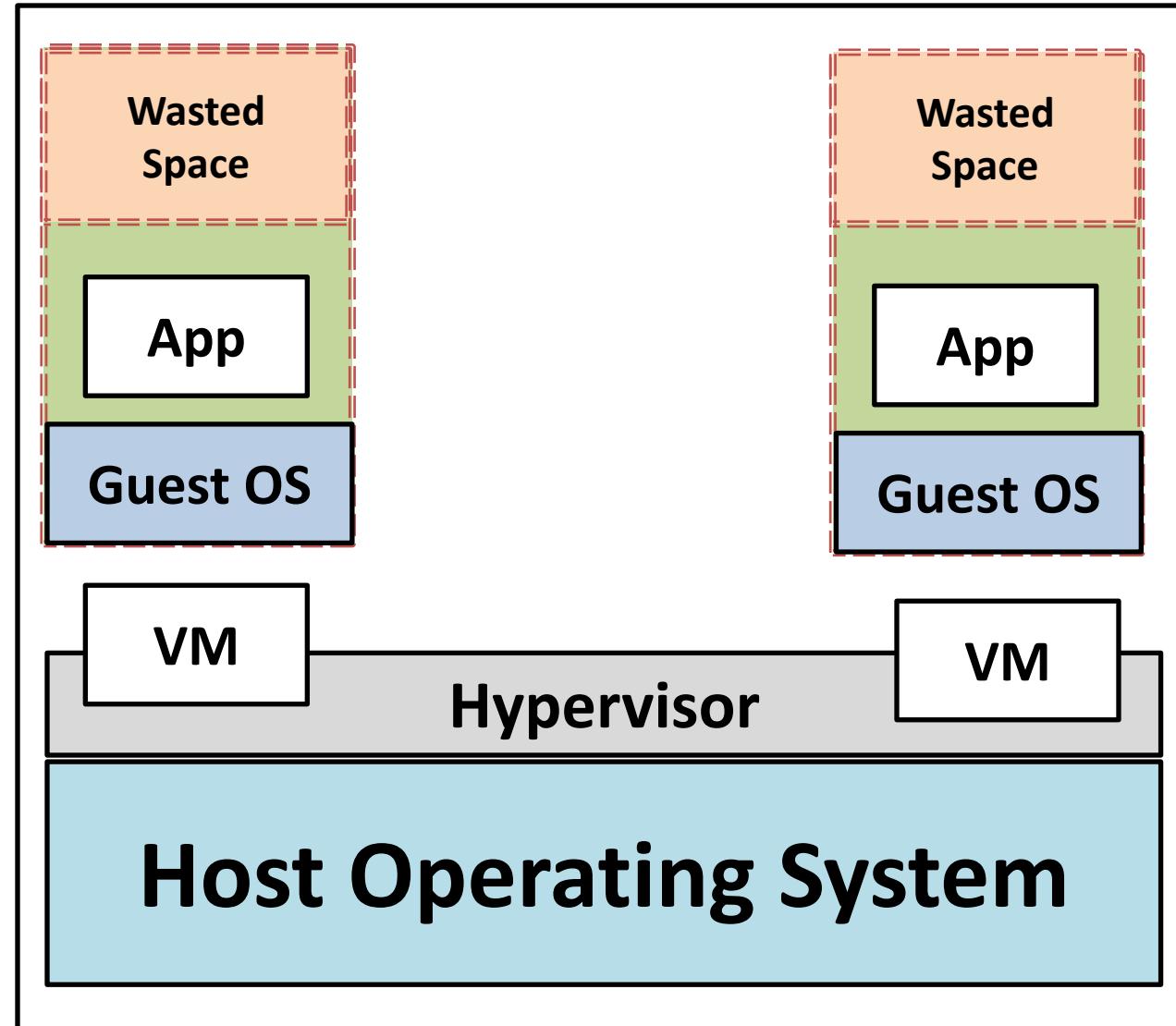
VMs



Containers



Functions



physical server

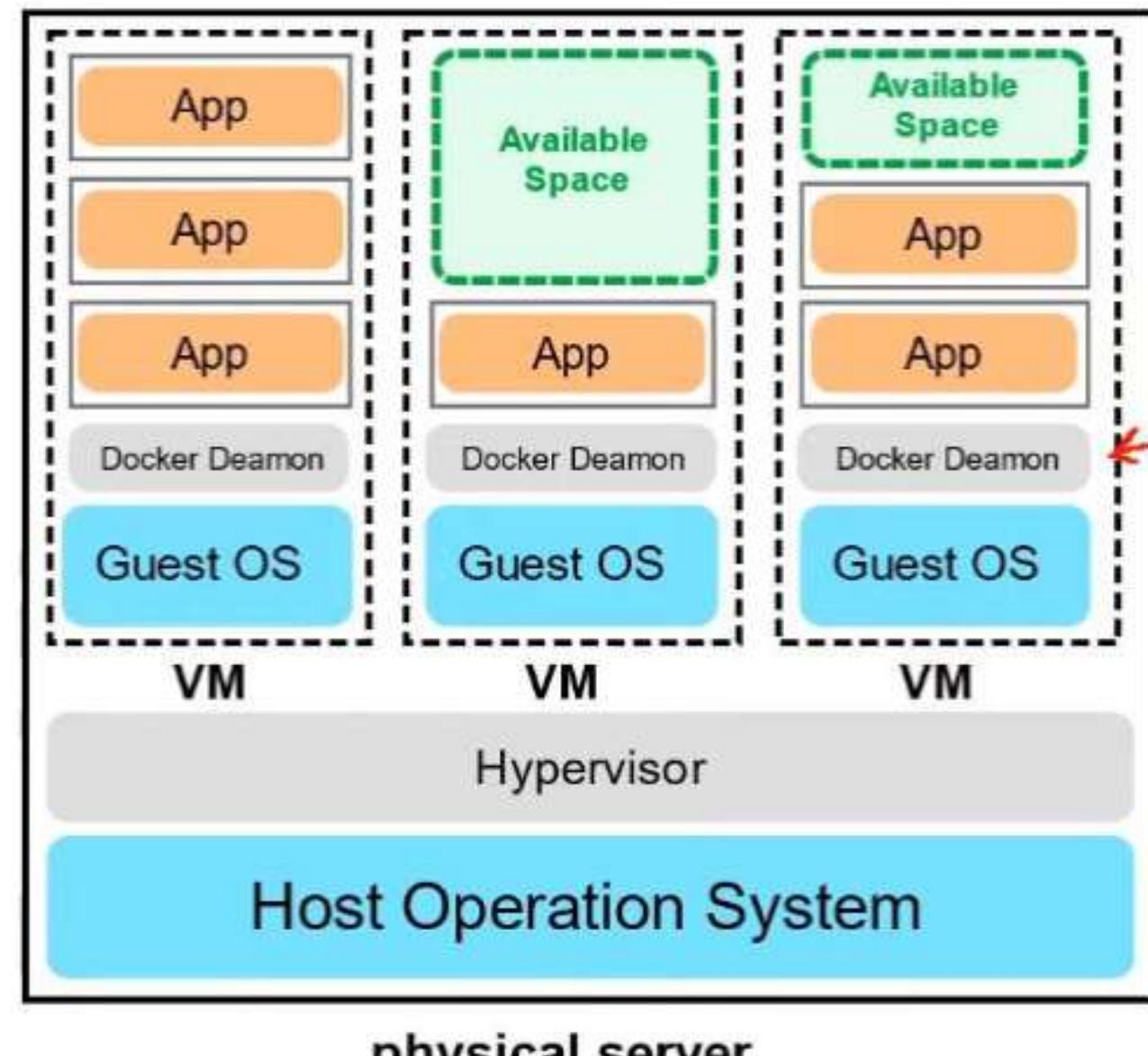
- You can run **multiple Virtual Machines on one machine**.
- **Hypervisor** is the software layer that lets you run the VMs
- A physical server shared by multiple customers
- You are to pay for a fraction of the server
- You'll overpay for an underutilized Virtual Machine
- You are limited by your Guest Operating System
- Multiple apps on a single Virtual Machine can result in conflicts in resource sharing
- Easy to export or import images for migration
- Easy to Vertical or Horizontally scale

The Evolution of Computing

* Dedicated



VMs



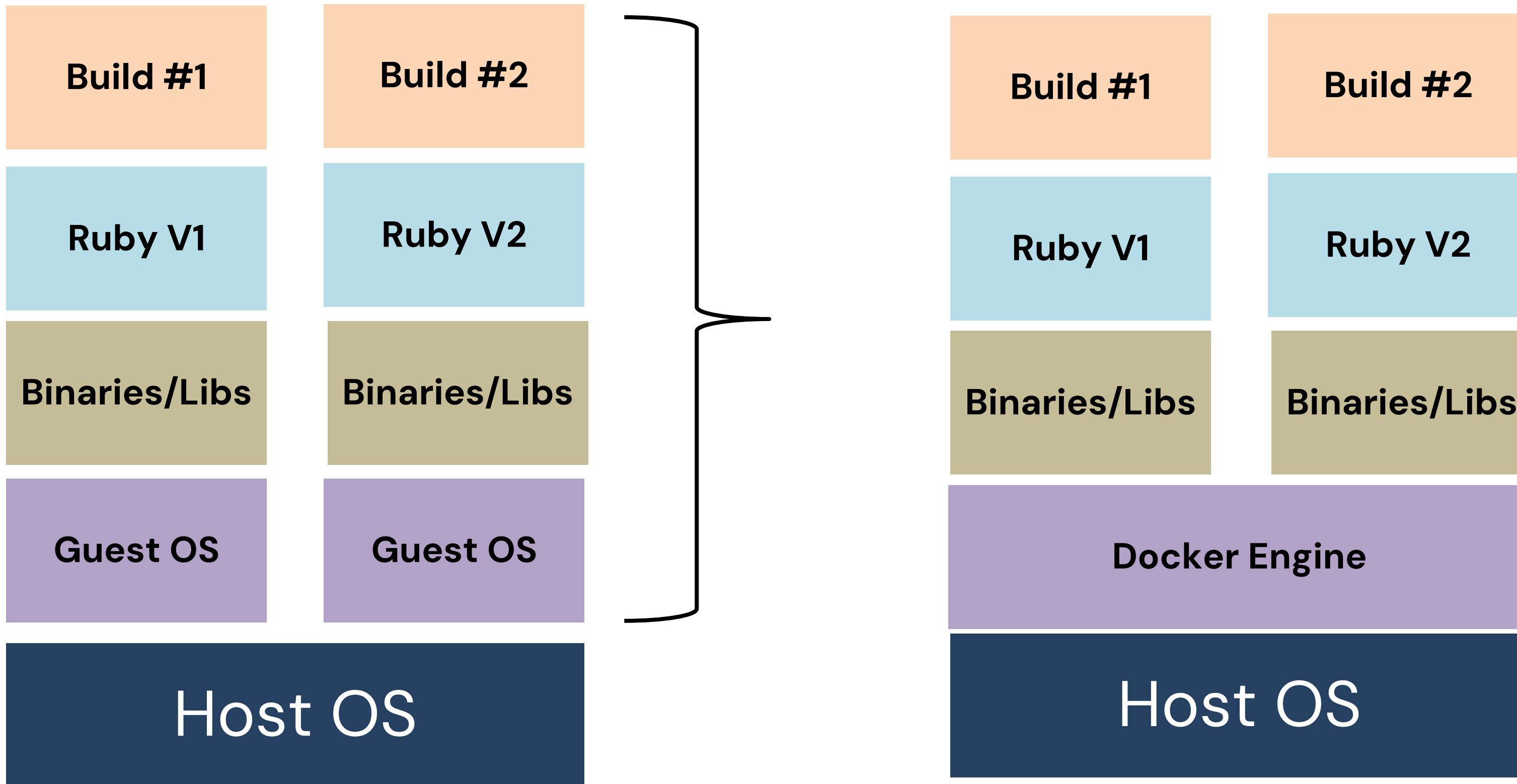
Containers



Functions

- Virtual Machine running multiple VM and containers
- You can maximize the utilize of the available capacity which is more cost-effective
- Your containers share the same underlying OS so containers are more efficient than multiple VMs
- **Multiple apps can run side by side without being limited to the same OS requirements and will not cause conflicts during resource sharing**

A Comparison of VMs and Docker



The Evolution of Computing

*Dedicated



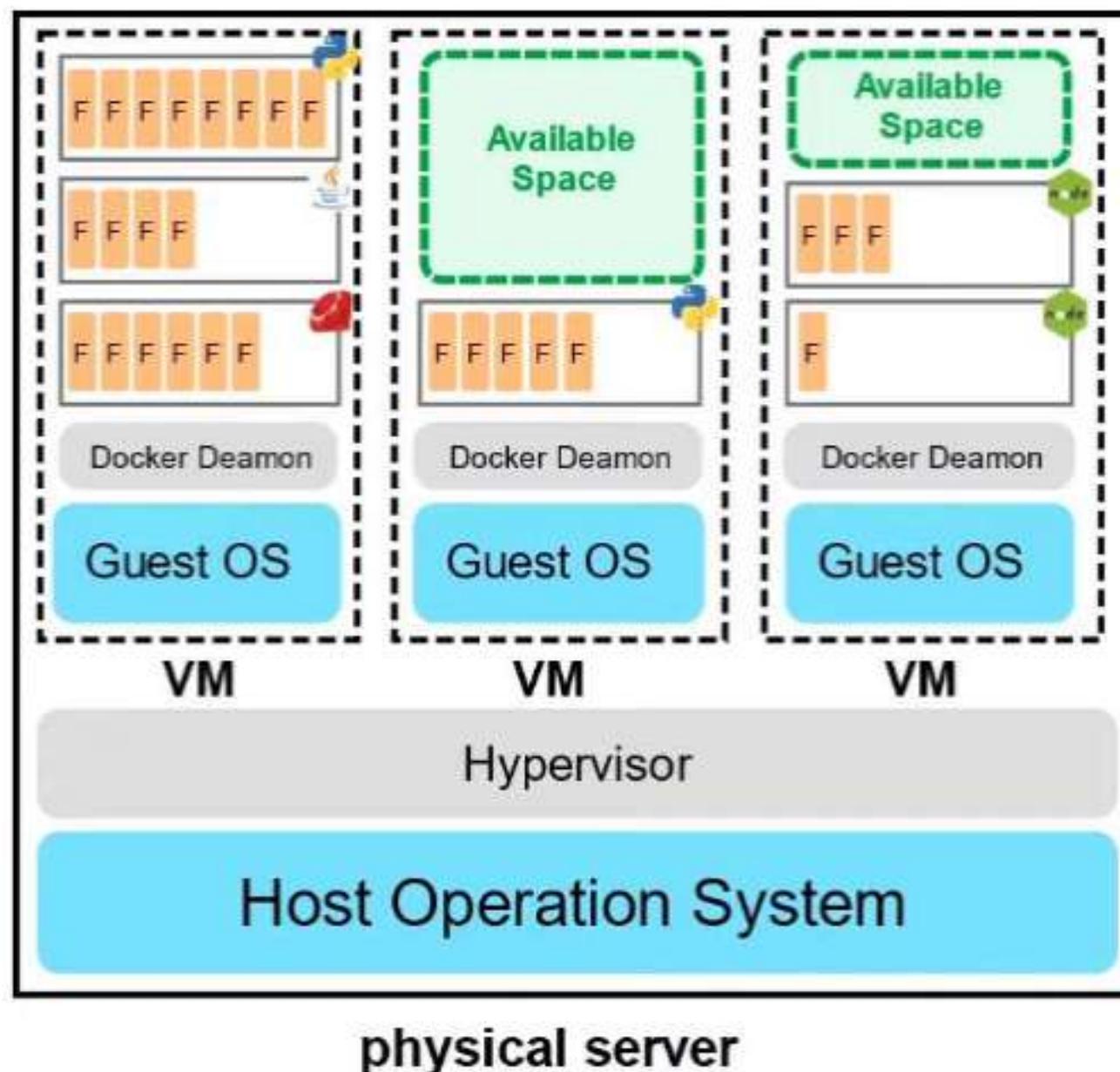
VMs



Containers



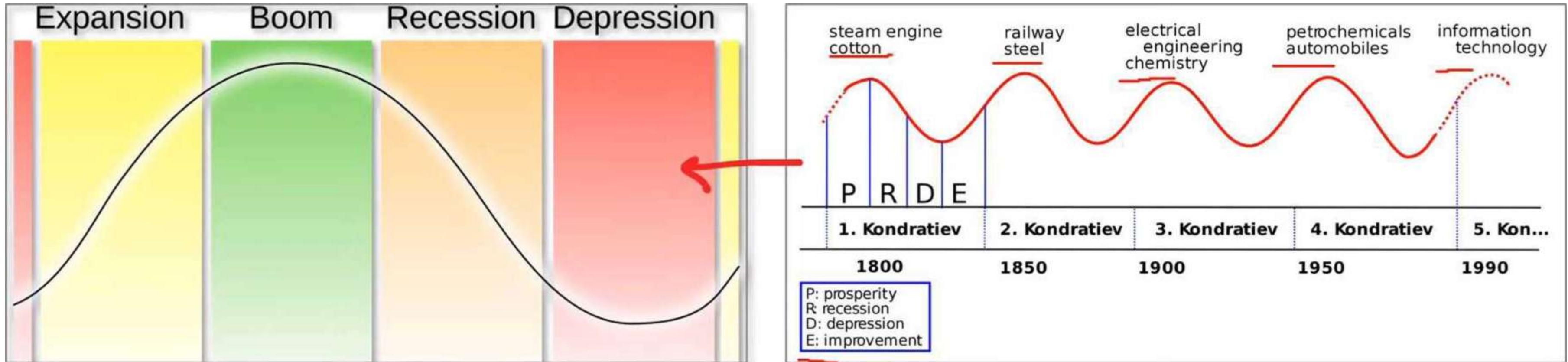
Functions



- Are managed VMs running managed containers.
- Known as **Serverless Compute**
- You upload a piece of code, choose the amount of memory and duration.
- Only responsible for code and data, nothing else. Very cost-effective, only pay for the time code is running, VMs only run when there is code to be executed
- Cold Starts is a side-effect of this setup

Innovation Waves

Kondratiev waves (aka Innovation Waves or K-Waves) are hypothesized cycle-like phenomena in the global world economy.
The phenomenon is closely connected with Technology life cycles.



A common pattern of a wave
change of **supply** and **demand**

Each wave irreversibly changes the
society on a global scale.
The latest wave is **Cloud Technology**

Burning Platform

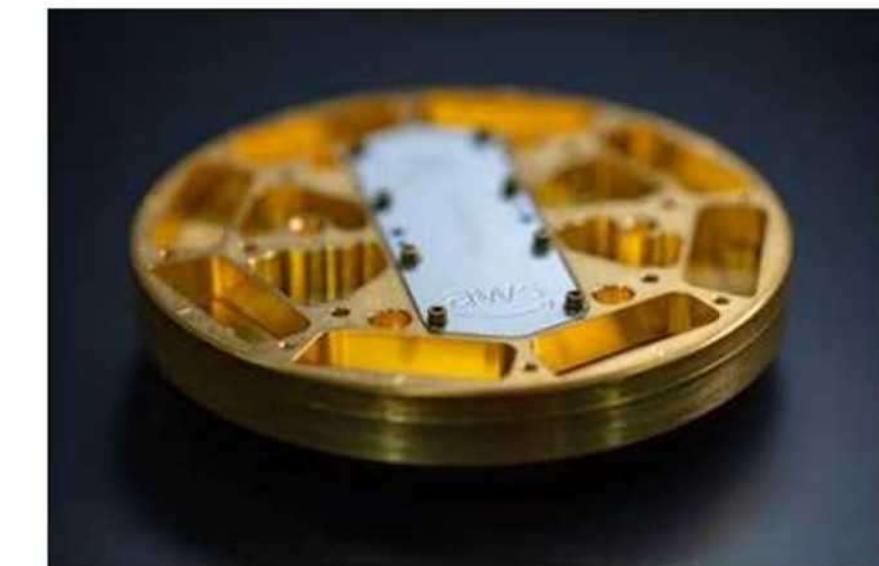
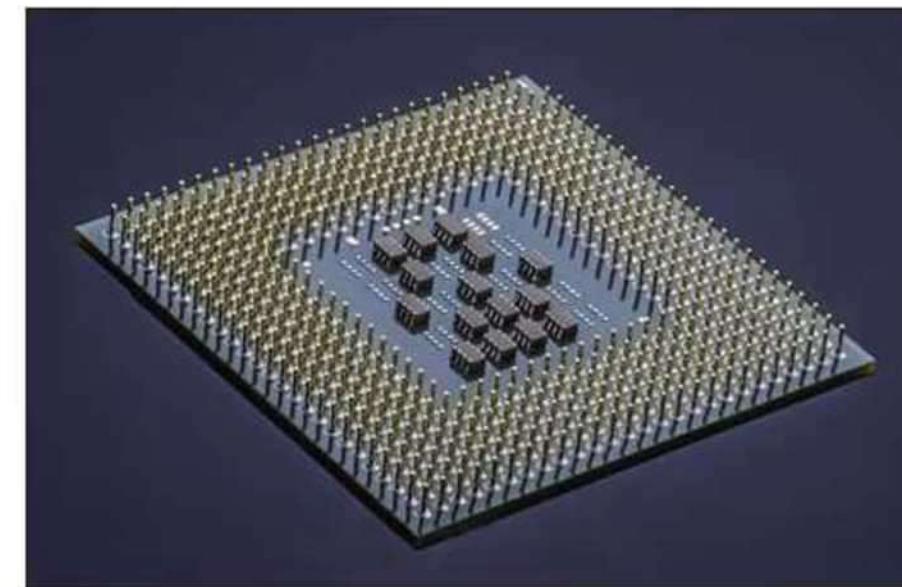
Burning platform is a term used **when a company abandons old technology for new technology** with the uncertainty of success and can be motivated by fear that the organization future survival hinges on its **digital transformation**



Evolution of Computing Power

What is Computing Power?

The throughput measured at which a computer can complete a computational task.



General Computing

Xeon CPU Processor

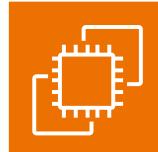
GPU Computing

*50x faster than traditional CPUs

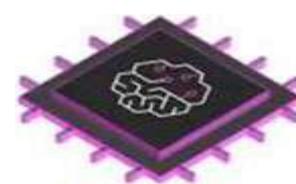
Quantum Computing

- D-Wave 2000Q
- **Rigetti 16Q Aspen-4**
- IonQ linear ion trap
- 100 Million times faster

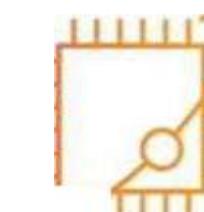
AWS Service Offering



AWS Elastic Compute Cloud EC2



AWS Inferentia (Infl)



AWS Bracket

Types of Cloud Computing



SaaS Software as a Service For Customers
A product that is run and managed by the service provider
***Don't worry about how the service is maintained.
It just works and remains available.***

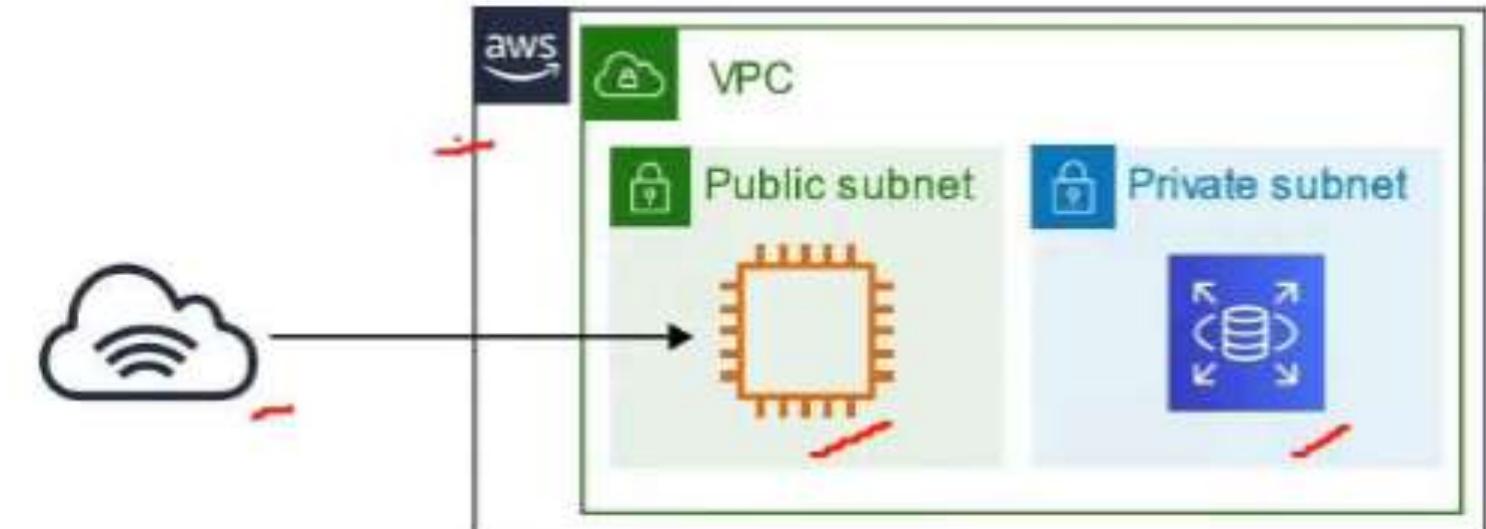
PaaS Platform as a Service For Developers
Focus on the deployment and management of your apps.
Don't worry about, provisioning, configuring or understanding the hardware or OS.

IaaS Infrastructure as a Service For Admins
The basic building blocks for cloud IT. Provides access to networking features, computers and data storage space.
Don't worry about IT staff, data centers and hardware

Cloud Computing Deployment Models

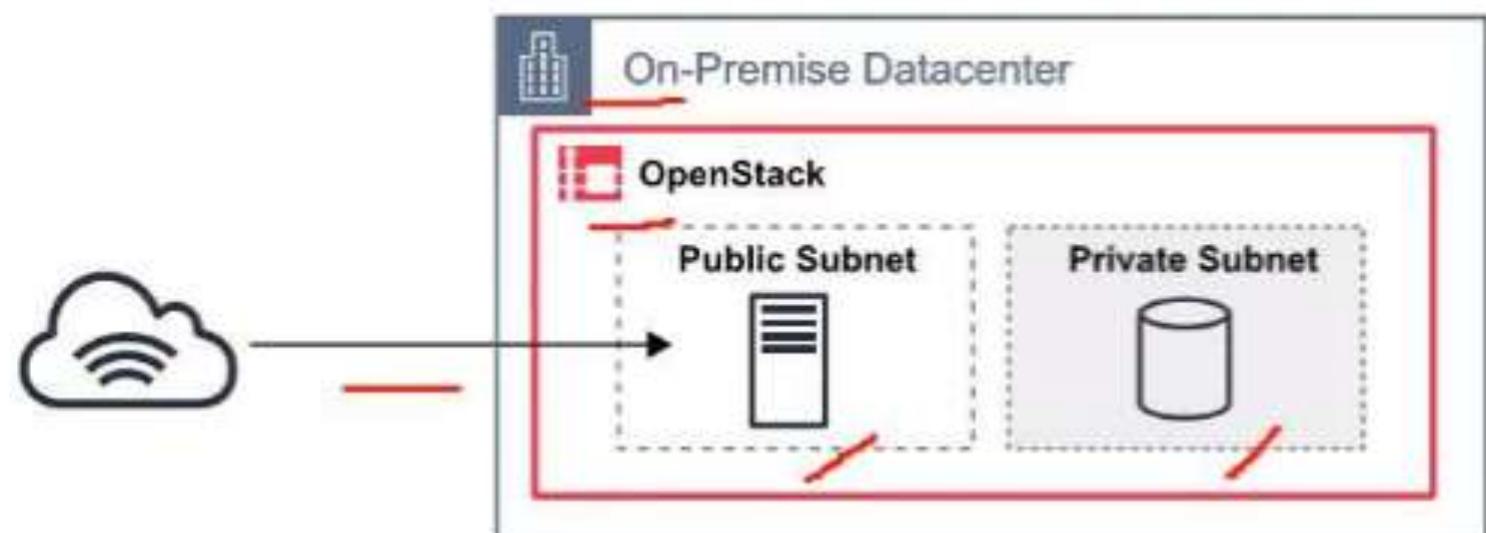
Public Cloud

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



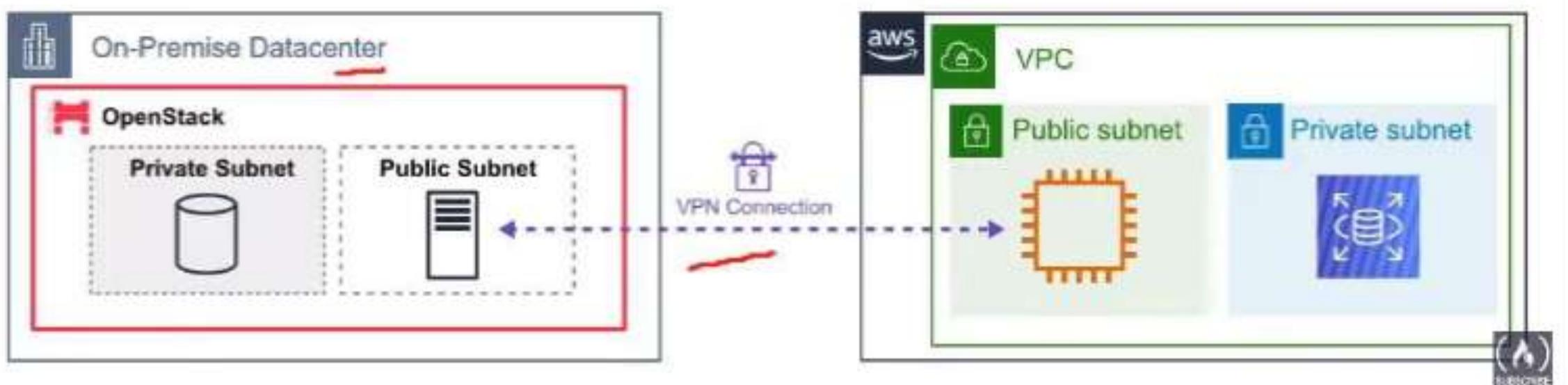
Private Cloud

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**



Cloud Computing Deployment Models

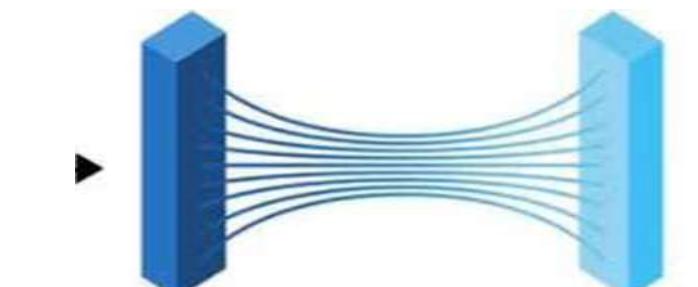
Cross-Cloud

Using **Multiple Cloud Providers**

Aka multi-cloud, "-hybrid-cloud"



Amazon EKS



Azure Arc



GCP Kubernetes Engine



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

Cloud Computing Deployment Models

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



Deloitte.

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".



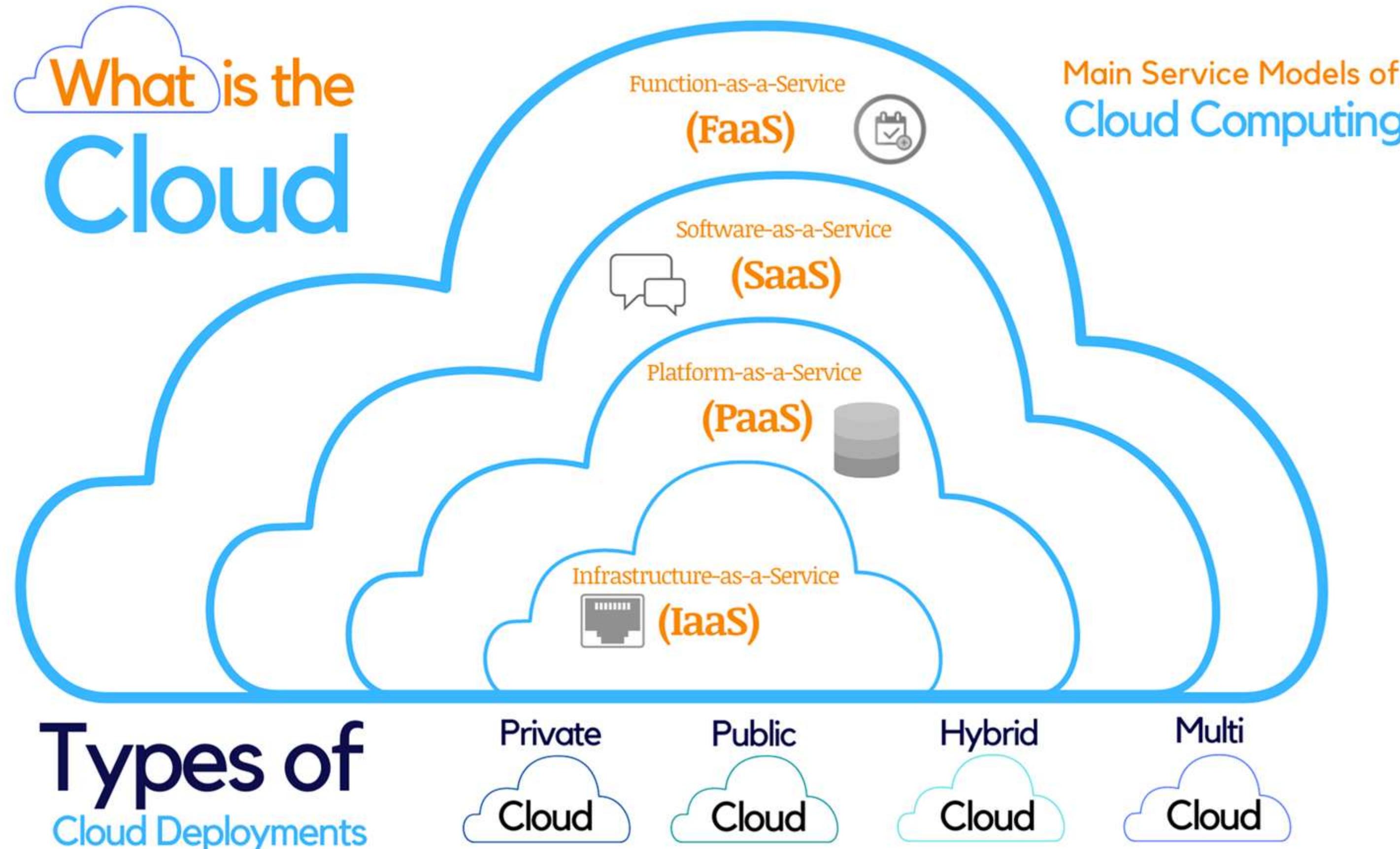
Canada

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 & Types

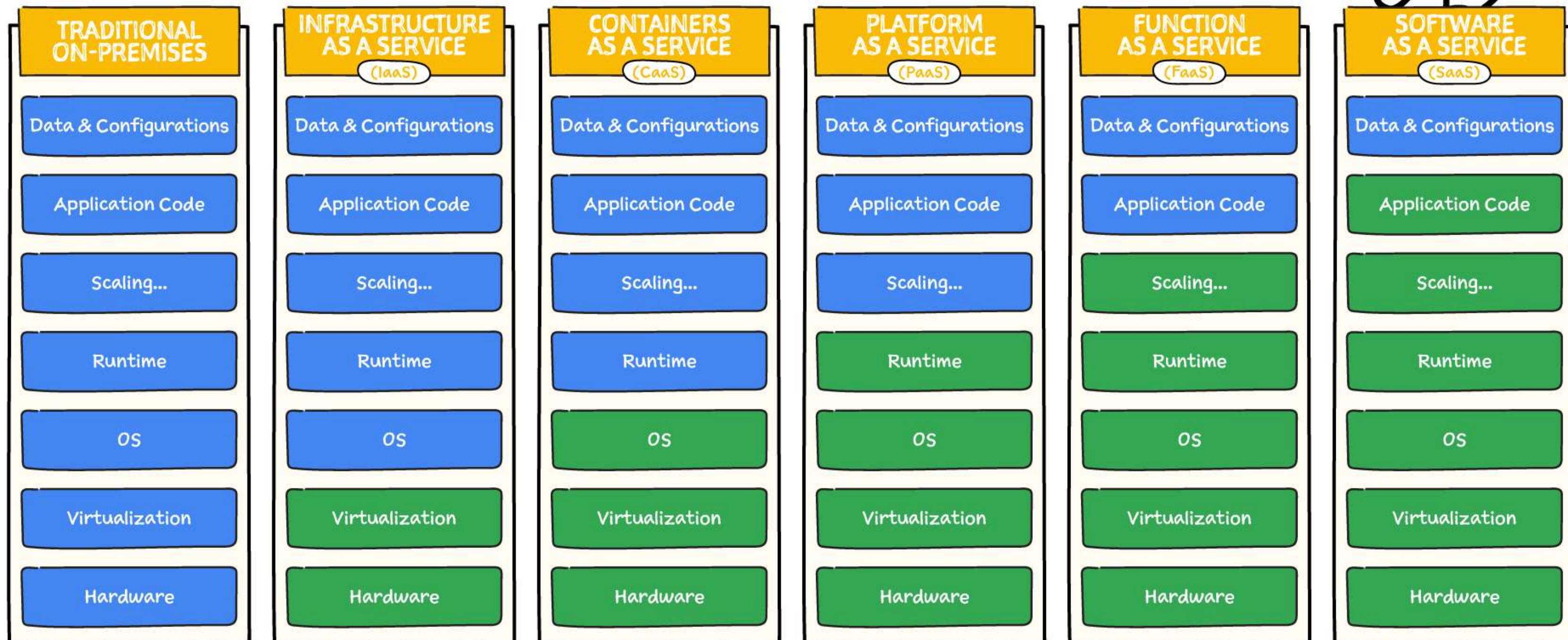


Cloud Computing Deployment Models & Types

#GCPSketchnote
@PVERGADIA
THECLOUDGIRL.DEV
08.II.2021



Wait... what is **Cloud** again?



You Manage



Cloud Provider Manages

Cloud Computing Deployment Models & Types

You want to	Use
Provision HW e.g. servers	IAAS
Want to Run an app without servers	FAAS
Want to build and deploy applications	PAAS
Want to re-use Existing SW from others	SAAS
Want to sell your SW so that others can use	SAAS
Type	Common Examples

SaaS	Google Workspace, Dropbox, Salesforce, Cisco WebEx, Concur, GoToMeeting
PaaS	Amazon Web Services (AWS) Elastic Beanstalk, Windows Azure, Heroku, Force.com, Google App Engine, Apache Stratos, Red Hat OpenShift
IaaS	DigitalOcean, Linode, Rackspace, AWS, Cisco Metapod, Microsoft Azure, Google Compute Engine (GCE)

The Benefits of Cloud



Here is a summary of reasons why an organization would consider adopting or migrating to utilizing public cloud.

Agility

- Increase speed and 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

Security

Reliability

- Stop spending money on running and maintaining data centers

High Availability

Scalability

- Benefit from massive economies of scale

Elasticity

Cloud Architecture

Missing:

- Fault Tolerance
- Disaster Recovery



The Benefits of Cloud is a reworking and expansion of the Six Advantages of Cloud

Six. Advantages to Cloud



- 1 Trade capital expense for variable expense

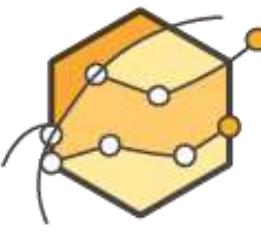
You can **Pay On-Demand** meaning there is no upfront-cost and you pay for only what you consume or pay by the hour, minutes or seconds.

Instead of paying for upfront costs of data centers and servers



2. Benefit from massive economies of scale

You are **sharing the cost with other customers** to get unbeatable savings. **Hundreds of thousands of customers utilizing a fraction of a server**



3. Stop guessing capacity

Scale up or down to meet the current need. Launch and destroy services whenever
Instead of paying for idle or underutilized servers.



4. Increase speed and agility

Launch resources within a few clicks in minutes
instead of waiting days or weeks of your IT to implement the solution on-premise



5. Stop spending money on running and maintaining data centers

Focus on your own customers, developing and configuring your applications
instead of operations such as of racking, stacking, and powering servers



6. Go global in minutes

Deploy your app in multiple regions around the world with a few clicks.

Provide lower latency and a better experience for your customers at minimal cost.



The Six Advantages of Cloud was AWS *original description* of Cloud Benefits

Seven Advantages to Cloud

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.

Cloud Characteristics

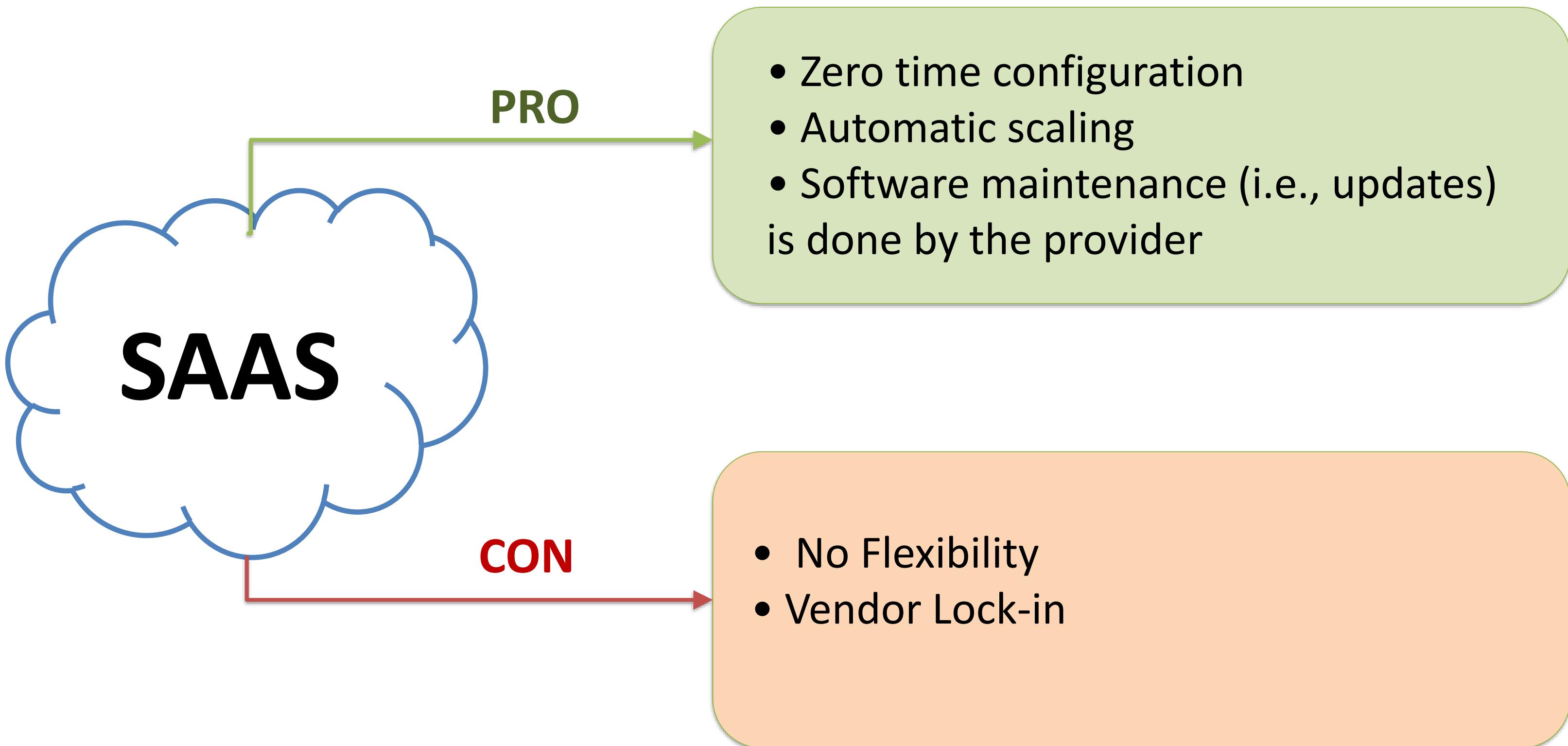
Business Characteristics

- No up-front investment:
- On-demand usage for resources, there's no need to make a large one-time investment before actual demand occurs
- Fixed costs become variable; CAPEX becomes OPEX:
- Resource consumption changes in real time
- Allocation is fine-grained:
- Minimal usage amounts for both time and resources •
- Business Gains flexibility
- Increased customer and business focus

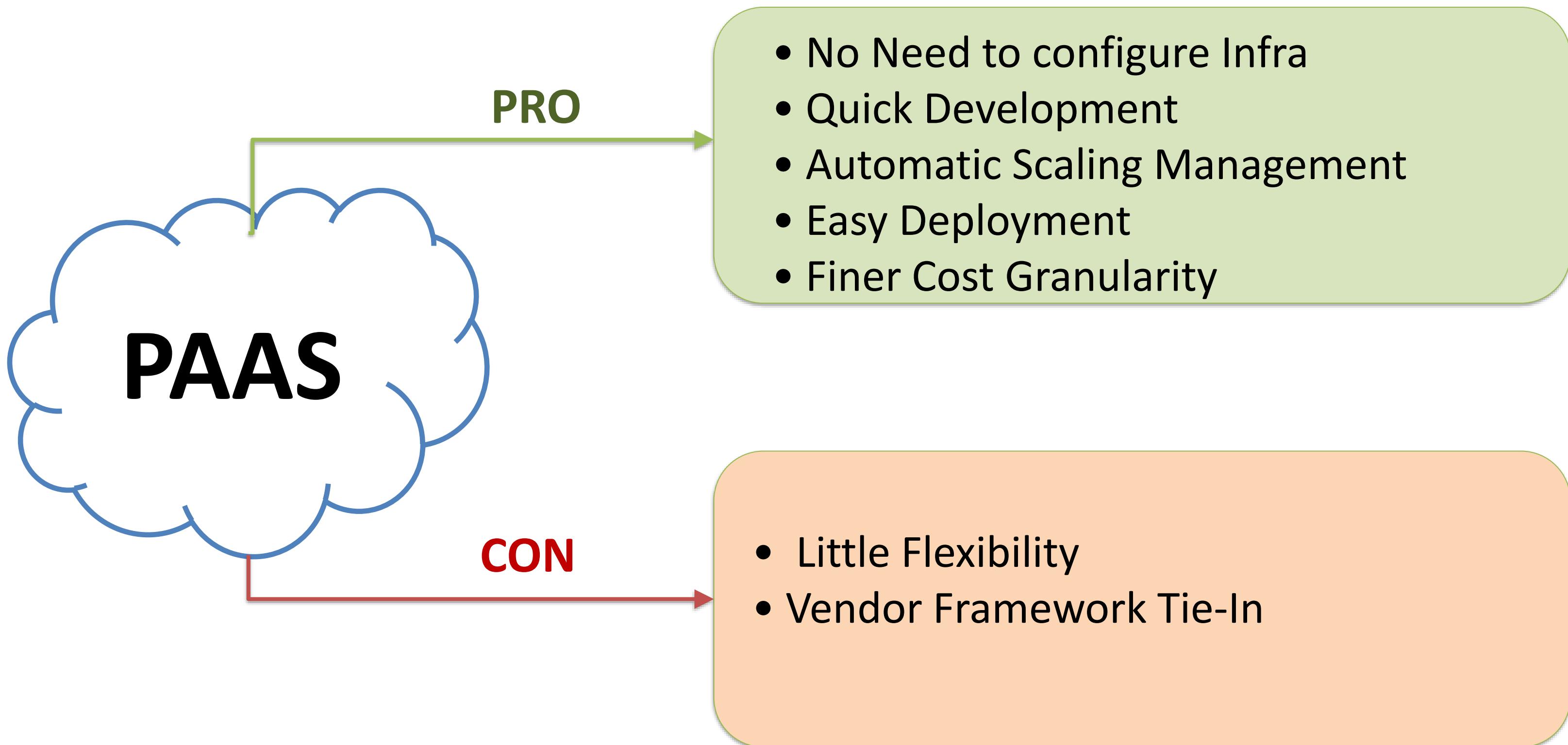
Technology Characteristics

- Quick Scaling of resources
- Infinite scalability illusion
- Abstract resources – results and performance
- Building blocks for additional functionality
- Cheap Experimentation

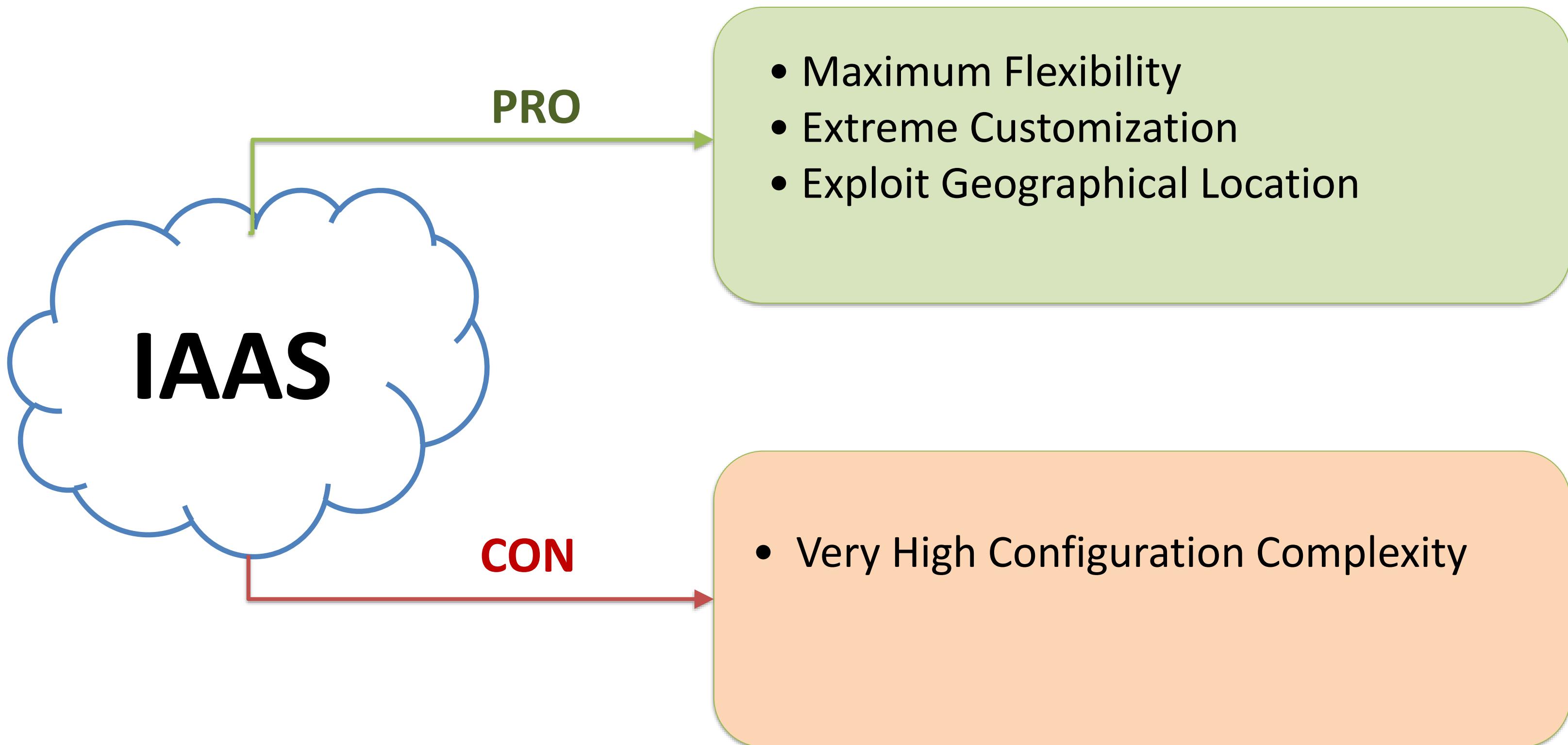
Pros and Cons of Cloud



Pros and Cons of Cloud



Pros and Cons of Cloud



Top 10 Obstacles



1. Availability of Service
2. Data Lock In
3. Data Confidentiality/Auditability
4. Data Transfer Bottlenecks
5. Performance Unpredictability
6. Scalable Storage
7. Bugs in large scale distributed systems
8. Scaling quickly while maintaining QoS
9. Reputation Fate sharing
10. Software Licensing

Cloud Architecture Terminologies

What is a Solutions Architect?

A role in a technical organization that architects a technical solution using multiple systems via researching, documentation, experimentation.

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

- (Security) How secure is this solution?
- (Cost) How much is this going to cost?

Cloud Architecture Terminologies

What is a Cloud Architect?

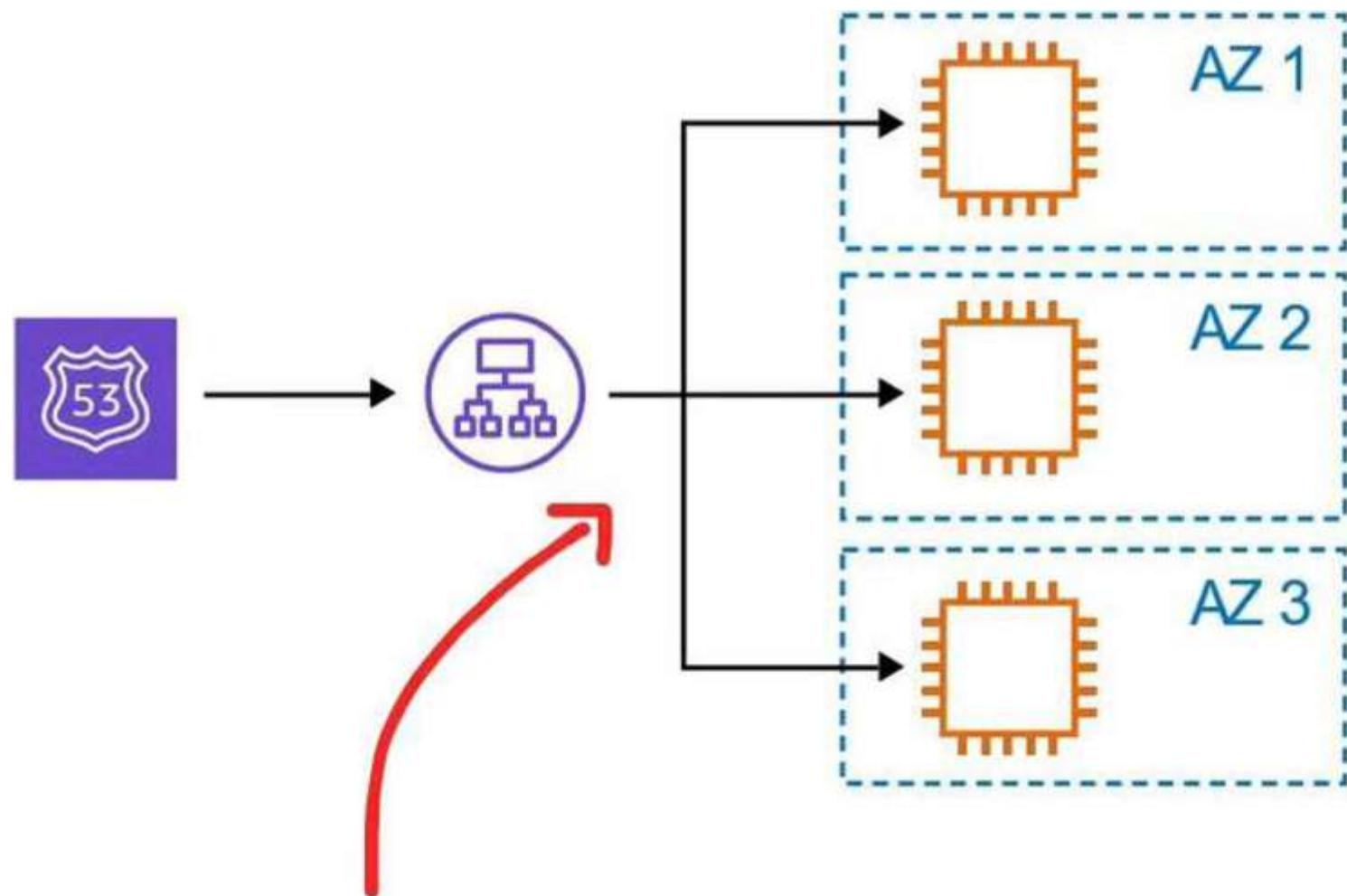
A solutions architect that is focused solely on architecting technical solutions using cloud services.

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

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

High Availability

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



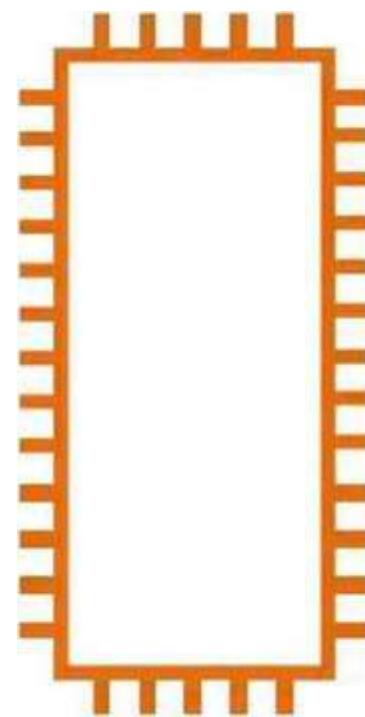
Elastic Load Balancer

A load balancer allows you to evenly distribute traffic to multiple servers in one or more datacenter. If a datacenter or server becomes unavailable (unhealthy) the load balancer will route the traffic to only available datacenters with servers.

Running your workload across multiple **Availability Zones** ensures that if 1 or 2 AZs become unavailable your service / applications remains available.

High Scalability

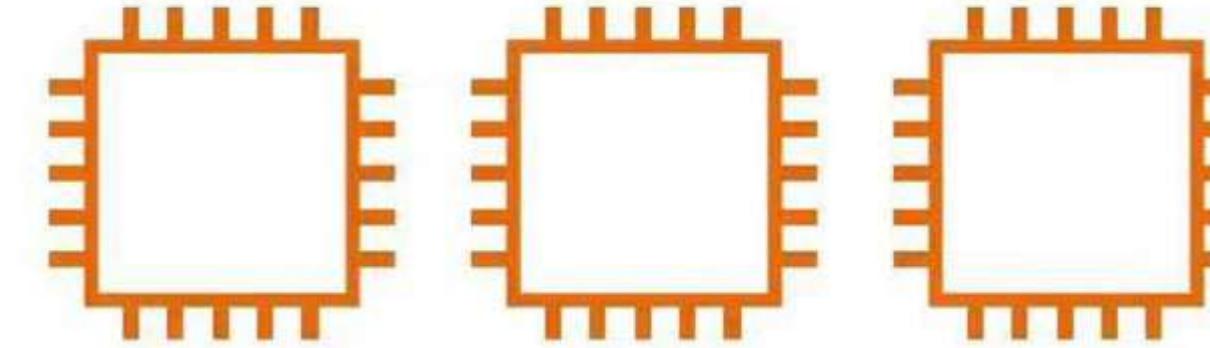
Your ability to **increase your capacity** based on the increasing demand of traffic, memory and computing power



Vertical Scaling

Scaling Up

Upgrade to a bigger server



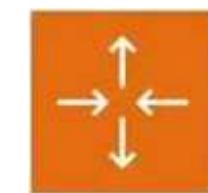
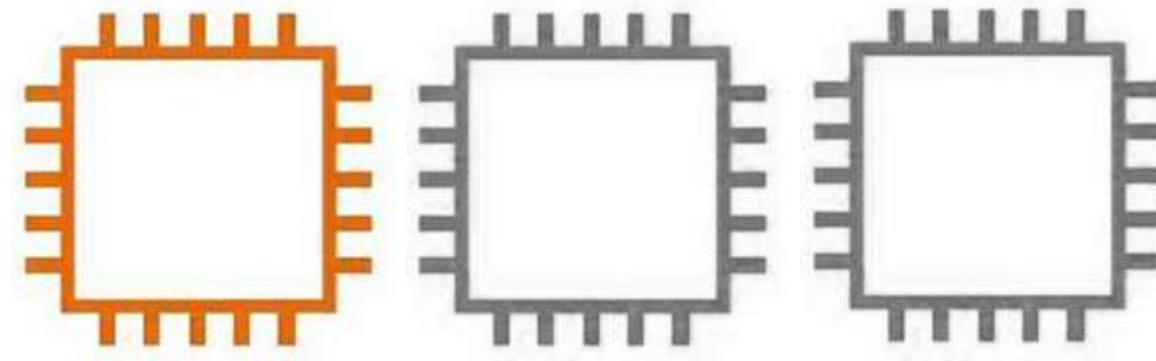
Horizontal Scaling

Scaling Out

Add more servers of the same size

High Elasticity

Your ability to **automatically** increase or decrease your capacity based on the current demand of traffic, memory and computing power



Auto Scaling Groups (ASG) is an AWS feature that will automatically add or remove servers based on scaling rules you define based on metrics

Horizontal Scaling

Scaling Out — Add more servers of the same size

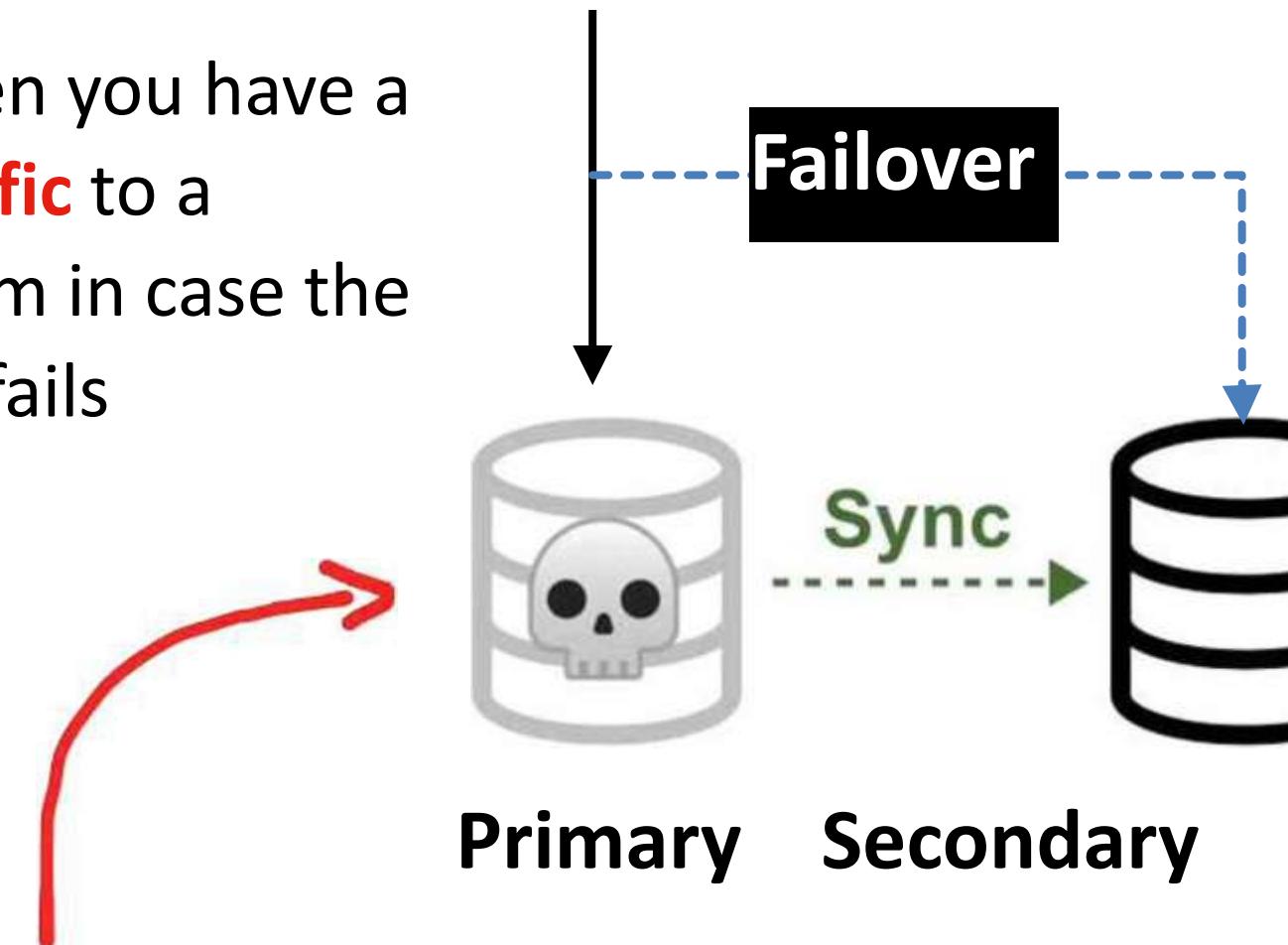
Scaling In — Removing underutilized servers of the same size

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

Highly Fault Tolerant

Your ability for your service to ensure there is no **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 Availability Zone 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 fail over occurs and it becomes the primary database.

High Durability

Your ability to **recover** from a disaster and 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 and preferred Region enabling fast and reliable recovery in case of IT data center failures.

Business Continuity Plan (BCP)

A business continuity plan (BCP) is a document that outlines how a business will continue operating **during an unplanned disruption in services**

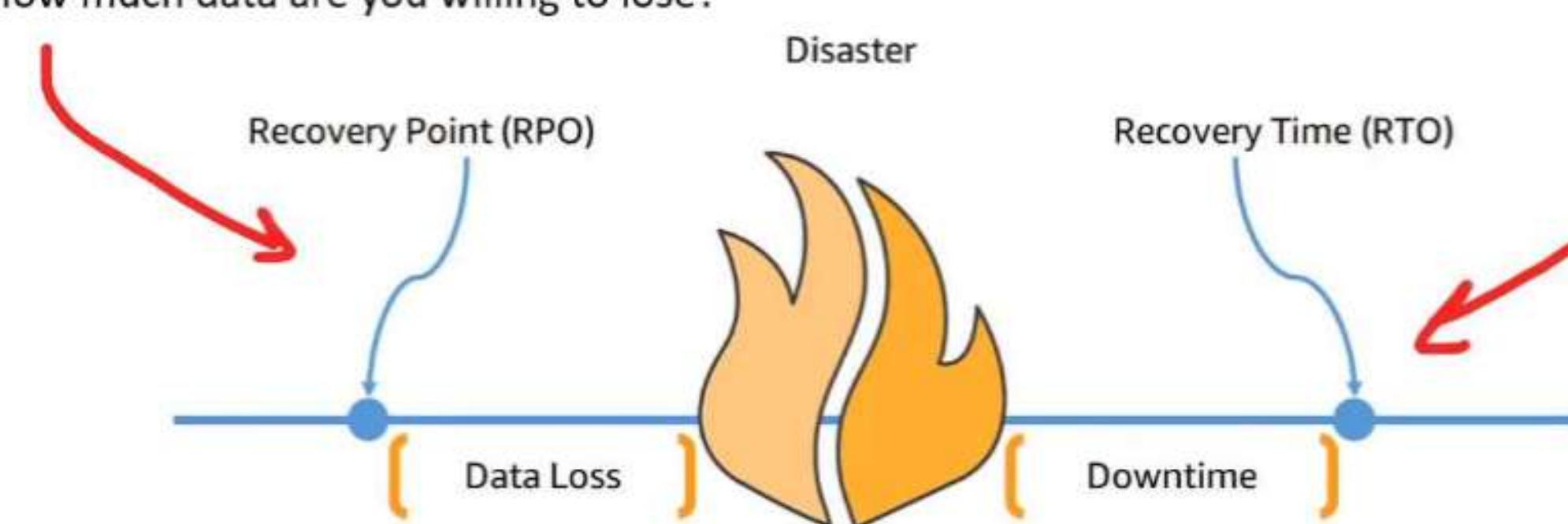
Recovery Point Objective (RPO)

the maximum acceptable amount of data loss after an unplanned data-loss incident, expressed as an amount of time

Recovery Time Objective (RTO)

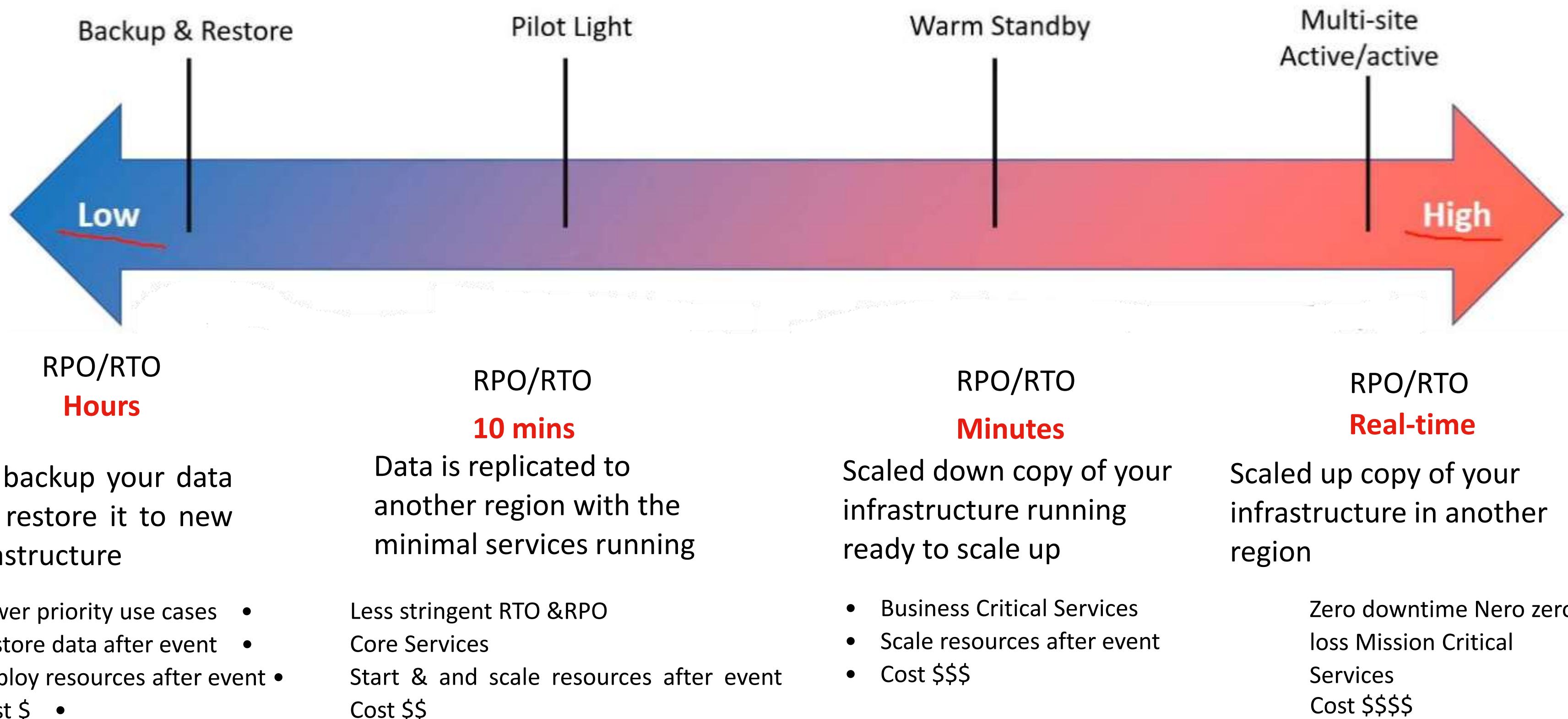
the maximum amount of downtime your business can tolerate without incurring a significant financial loss

How much data are you willing to lose?



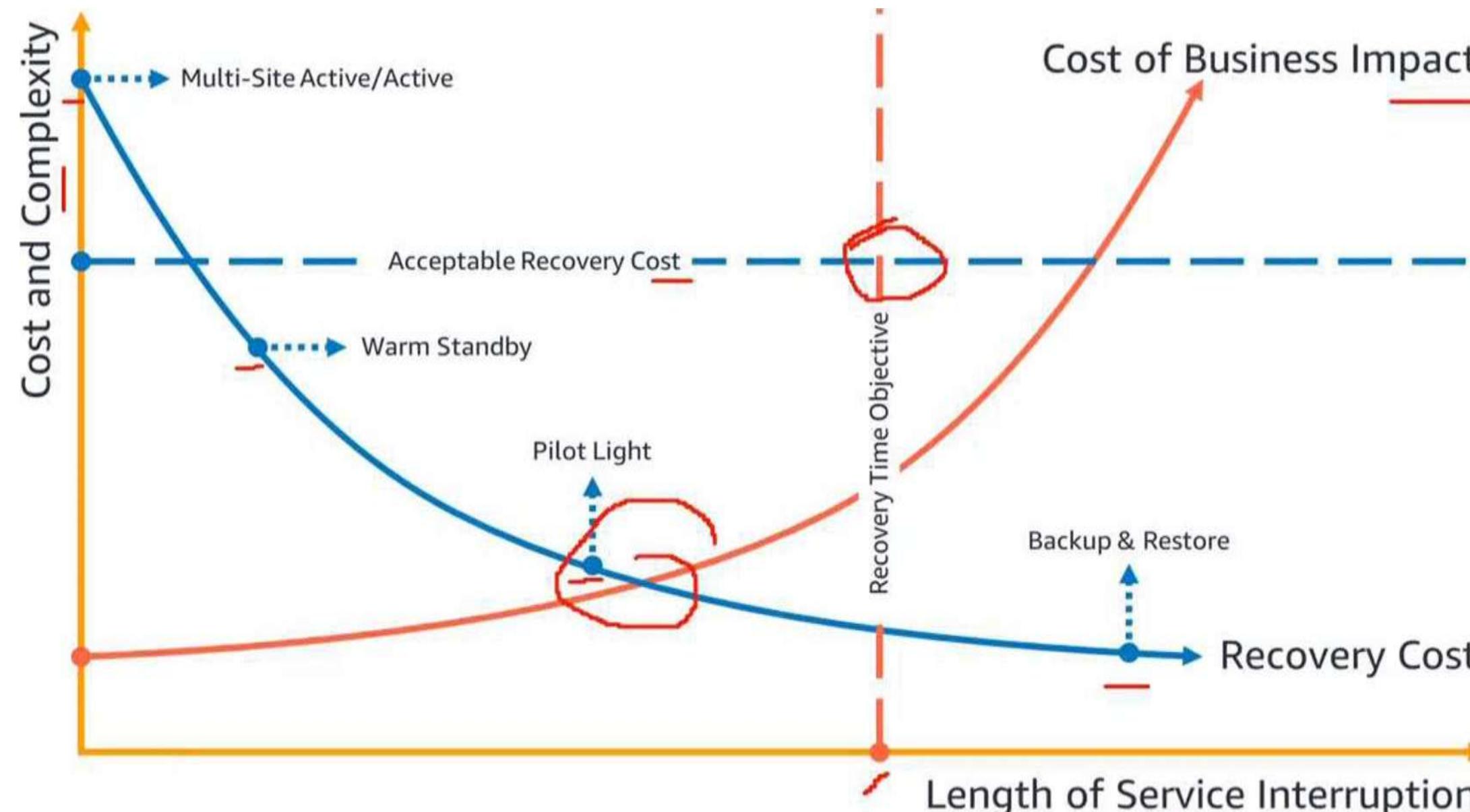
Disaster Recovery Options

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



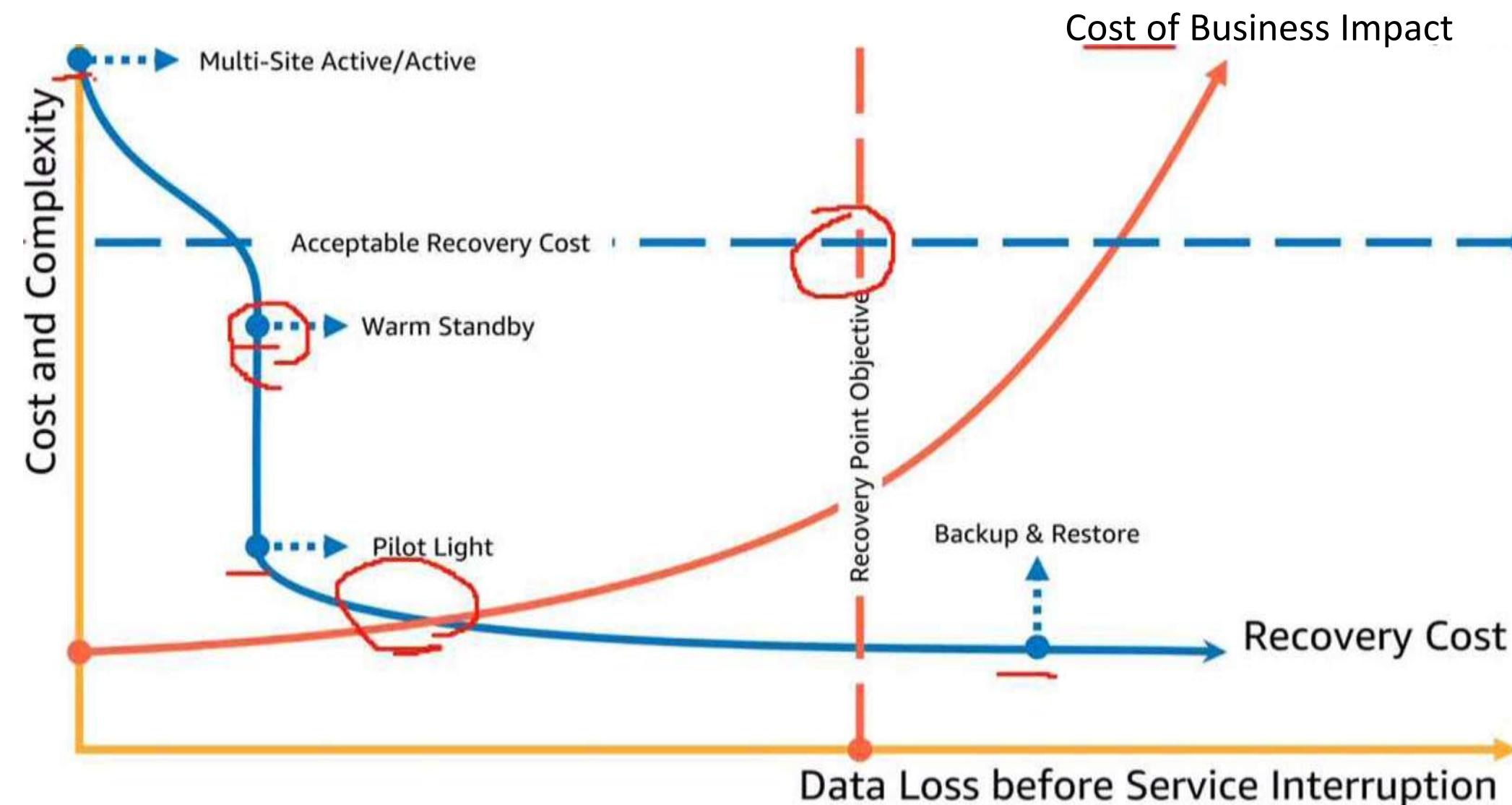
RTO

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



RPO

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



What is Amazon?



An American multinational computer technology corporation headquartered in **Seattle, Washington**



Amazon was founded in 1994 by **Jeff Bezos** and the company started as an online store for books and expanded to other products.

@timothyeberry on Unsplash

What is Amazon?



Jeff Bezos **today**

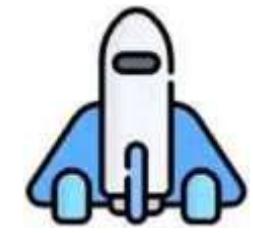
Amazon has expanded beyond just an online e-commerce store into:

- **cloud computing** (Amazon Web Services)
- digital streaming
- Amazon Prime Video
- Amazon Prime Music
- Twitch.tv
- Grocery Stores (Whole Foods Market)
- artificial intelligence
- Low orbit satellites (Kuiper Systems)
- And more!



Andy Jassy is the current CEO of Amazon.
Previously the CEO of AWS.

So Jeff Bezos can focus on space travel.



What is Amazon Web Services (AWS)?



Adam Selipsky

Former CEO of AWS

Former CTO of Tableau, spent a decade with AWS as VP of Marketing, Sales and Support



Matt Garman

CEO of AWS



Jeff Barr

Chief Evangelist

Werner Vogels

CTO of AWS "Everything fails, all the time!"



What is Amazon Web Services (AWS)?

Amazon calls their cloud provider service

Amazon Web Services

Commonly referred to just **AWS**



Old Logo



New Logo

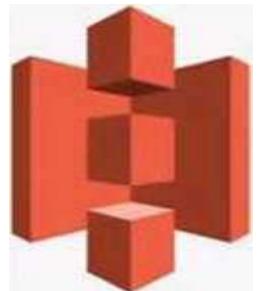
AWS was launched in ***2006** is the **leading cloud service provider** in the world.

Cloud Service Providers can be initialized as **CSPs**

What is Amazon Web Services (AWS)?



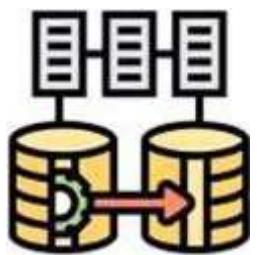
Simple Queue Service (SQS) was the first AWS service launched for public use in 2004



Simple Storage Service (S3) was launched in March of 2006



Elastic Compute Cloud (EC2) was launched in August of 2006



In November 2010, it was reported that all of Amazon.corn's retail sites had migrated to AWS

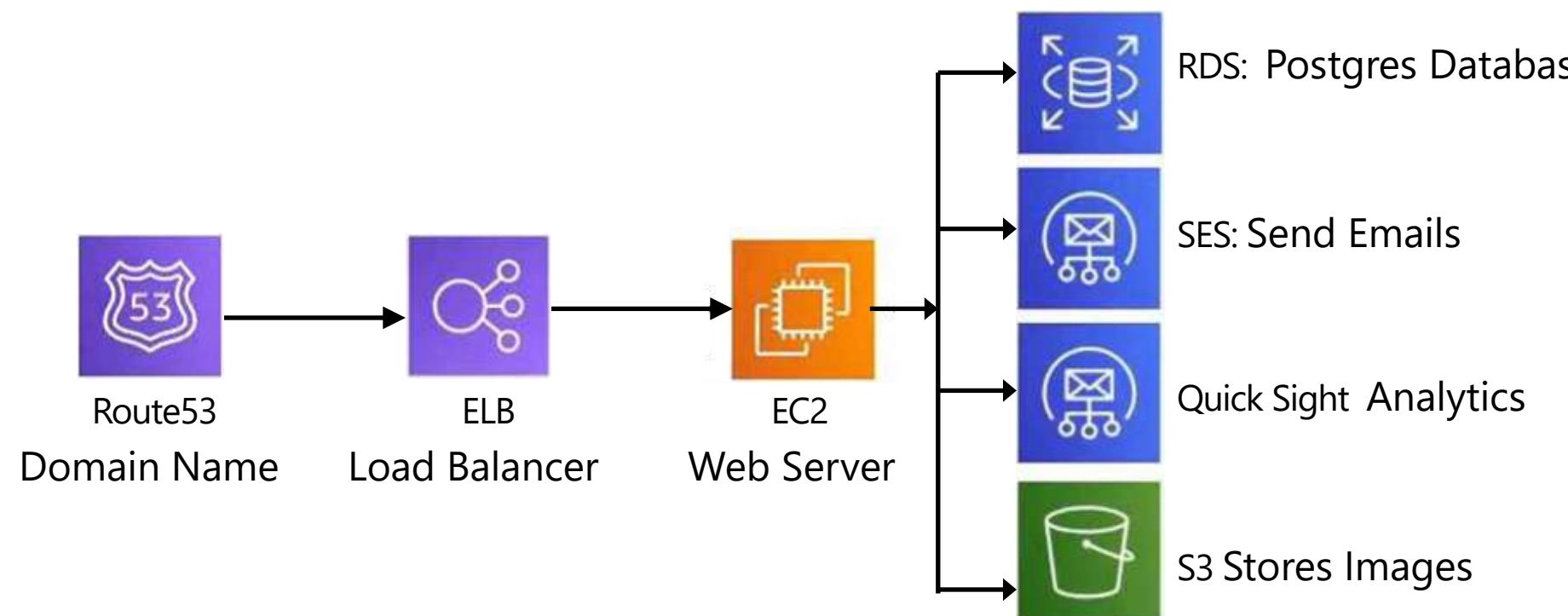


To support industry-wide training and skills standardization, AWS began offering a certification program for computer engineers, on April, 2013

What is a 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

Landscape of CSPs

Tier-1 (Top Tier) - Early to market, wide offering, strong synergies between services, well recognized in the industry



Amazon Web Services (AWS)



Microsoft Azure

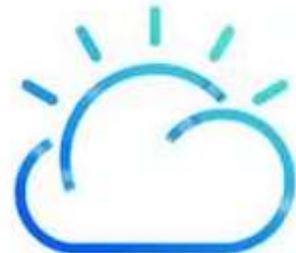


Google Cloud Platform (GCP)



Alibaba Cloud

Tier-2 (Mid Tier) - Backed by well-known tech companies, slow to innovate and turned to specialization



IBM Cloud



Oracle Cloud

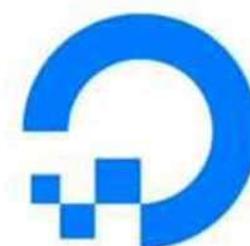


Rackspace (OpenStack)

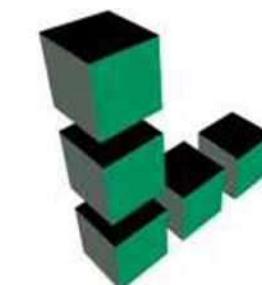
Tier-3 (Light Tier) - Virtual Private Servers (VPS) turned to offer core IaaS offering. Simple, cost-effective



Vultr



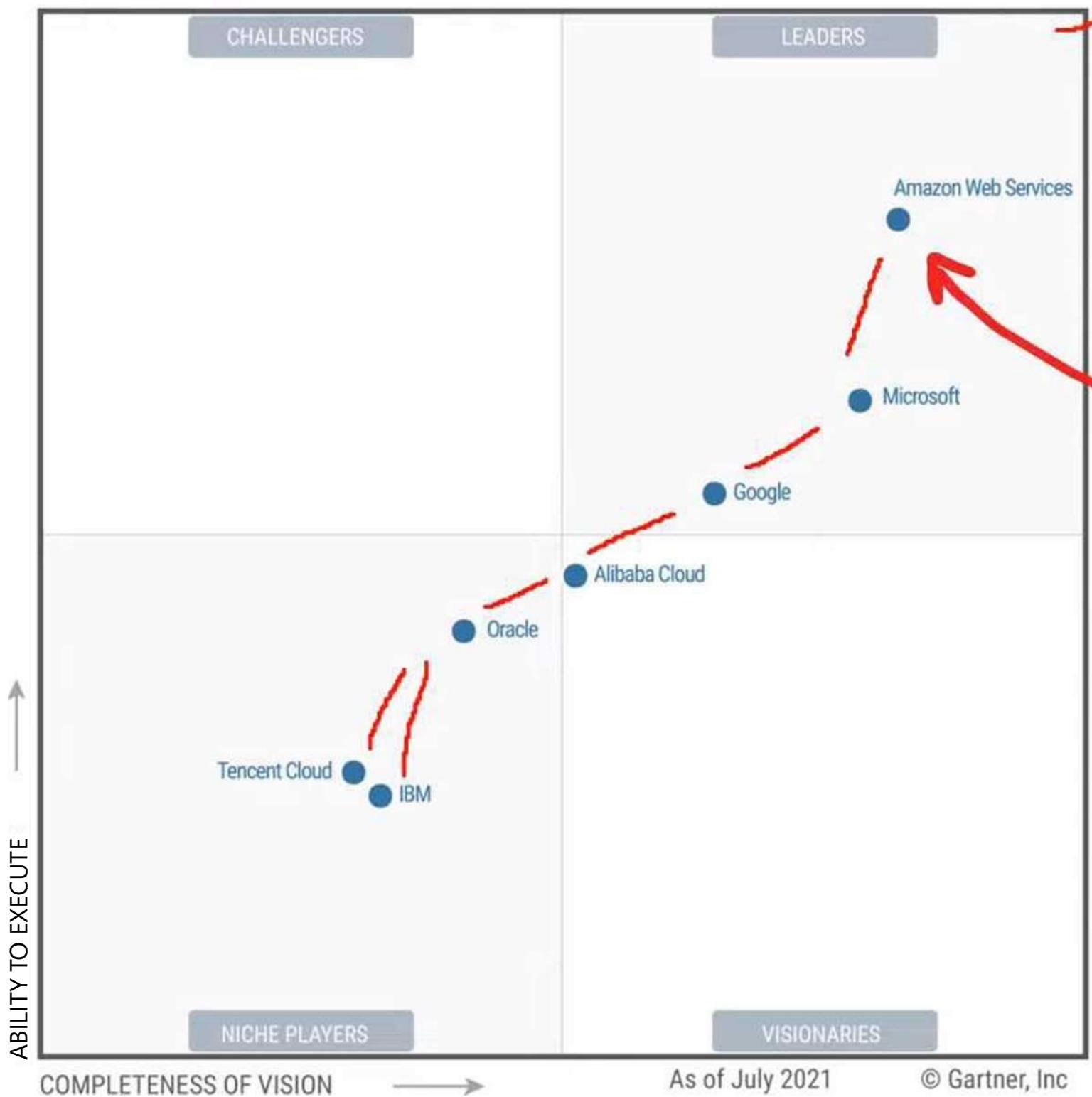
Digital Ocean



Linode

Gartner Magic Quadrant for Cloud

Figure 1: Magic Quadrant for Cloud Infrastructure and Platform Services



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 and participants.



Common Cloud Services

A cloud service provider **can have hundreds of cloud services** that are grouped into various types of services. The four most common types of cloud services (*the 4 core*) for Infrastructure as a Service (IaaS) would be:



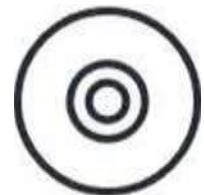
Compute

Imagine having a virtual computer that can run application, programs and code.



Networking

Imagine having virtual network defining internet connections or network isolations between services or outbound to the internet



Storage

Imagine having a virtual hard-drive that can store files



Databases

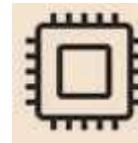
Imagine a virtual database for storing reporting data or a database for general purpose web-application

AWS has over **200+** cloud services

The term "Cloud Computing" can be used to refer to all categories, even though it has "compute" in the name.

Technology Overview

Cloud Service Provider (CSPs) that are **Infrastructure as a Service (IaaS)** will always have **4 core cloud service** offerings:



Compute



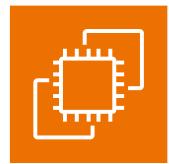
Storage



Database



Networking



||
Amazon Elastic Compute
Cloud (Amazon EC2)



Amazon Elastic Block Store
(Amazon EBS)



Amazon Relational Database
Service (Amazon RDS)



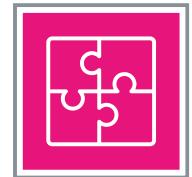
Amazon Virtual Private Cloud
(Amazon VPC)

Technology Overview

AWS offers more ☺



Analytics



Application Integration



AR & VR



AWS Cost Management



Block Chain



Business Applications



Containers



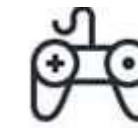
Customer Engagement



Developer Tools



End User Computing



Game Tech



Internet of Things



Machine Learning



Management & Governance



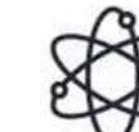
Media Services



Migration & Transfer



Mobile



Quantum Technologies



Robotics



Satellites



**Security, Identity &
Compliance**

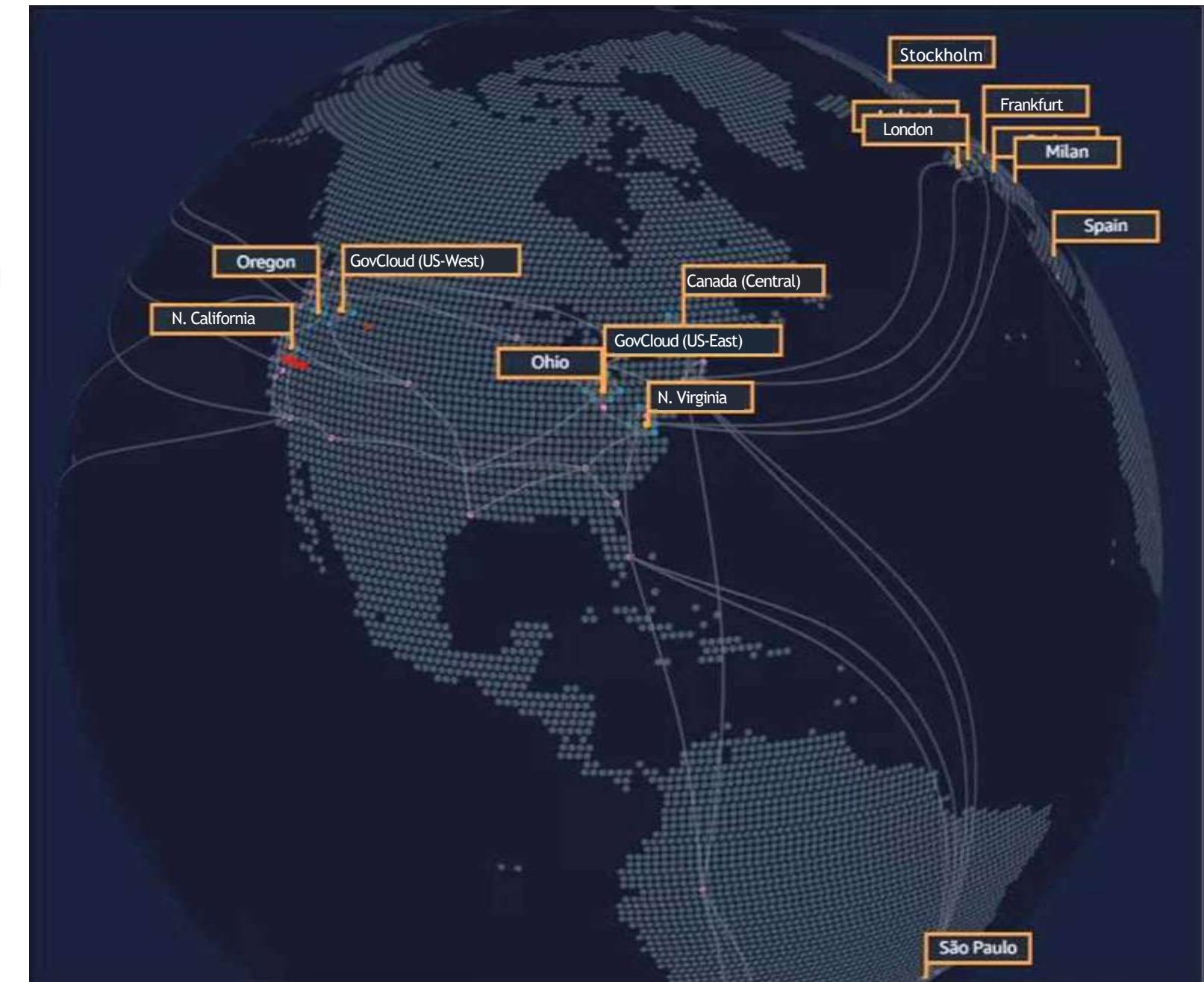
AWS Global Infrastructure

What is the AWS Global Infrastructure?

The AWS Global Infrastructure is **globally distributed hardware and datacenters** that **are physically networked together** to act as one large resource for the end customer.

The AWS Global Infrastructure is made up of the following resources:

- 245+ Countries and territories
- 36 Launched Regions
- 114 Availability Zones
- 135 Direct Connection Locations
- 700+ Points of Presence
- 43 Local Zone
- 31 Wavelength Zones



AWS has **millions** of active customers and **tens of thousands** of partners globally

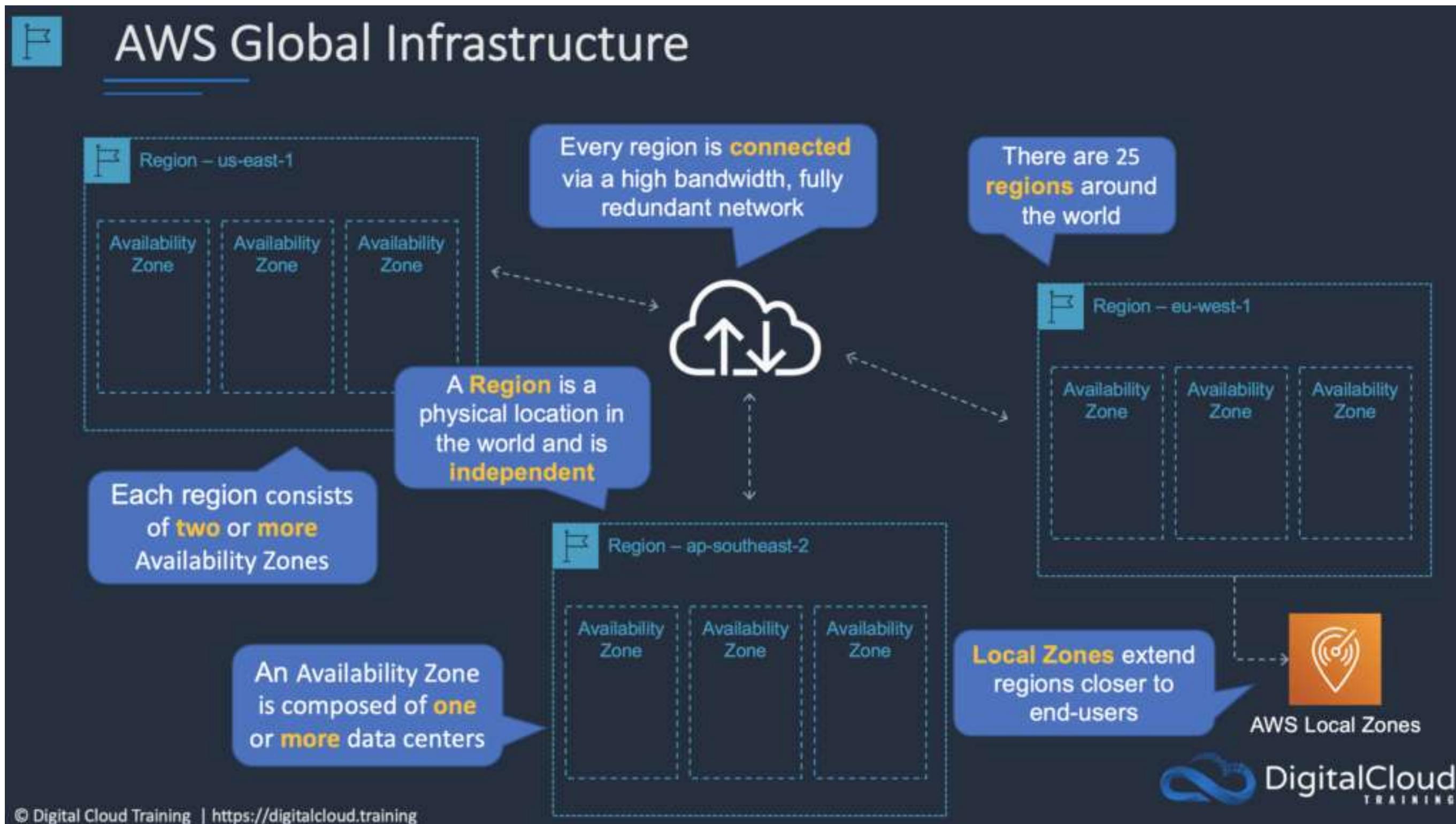
Global Infrastructure - Regions

Regions are **geographically distinct locations** consisting of one or more Availability Zones.

Every region is **physically isolated** from and independent of every other region in terms of **location, power, water supply**

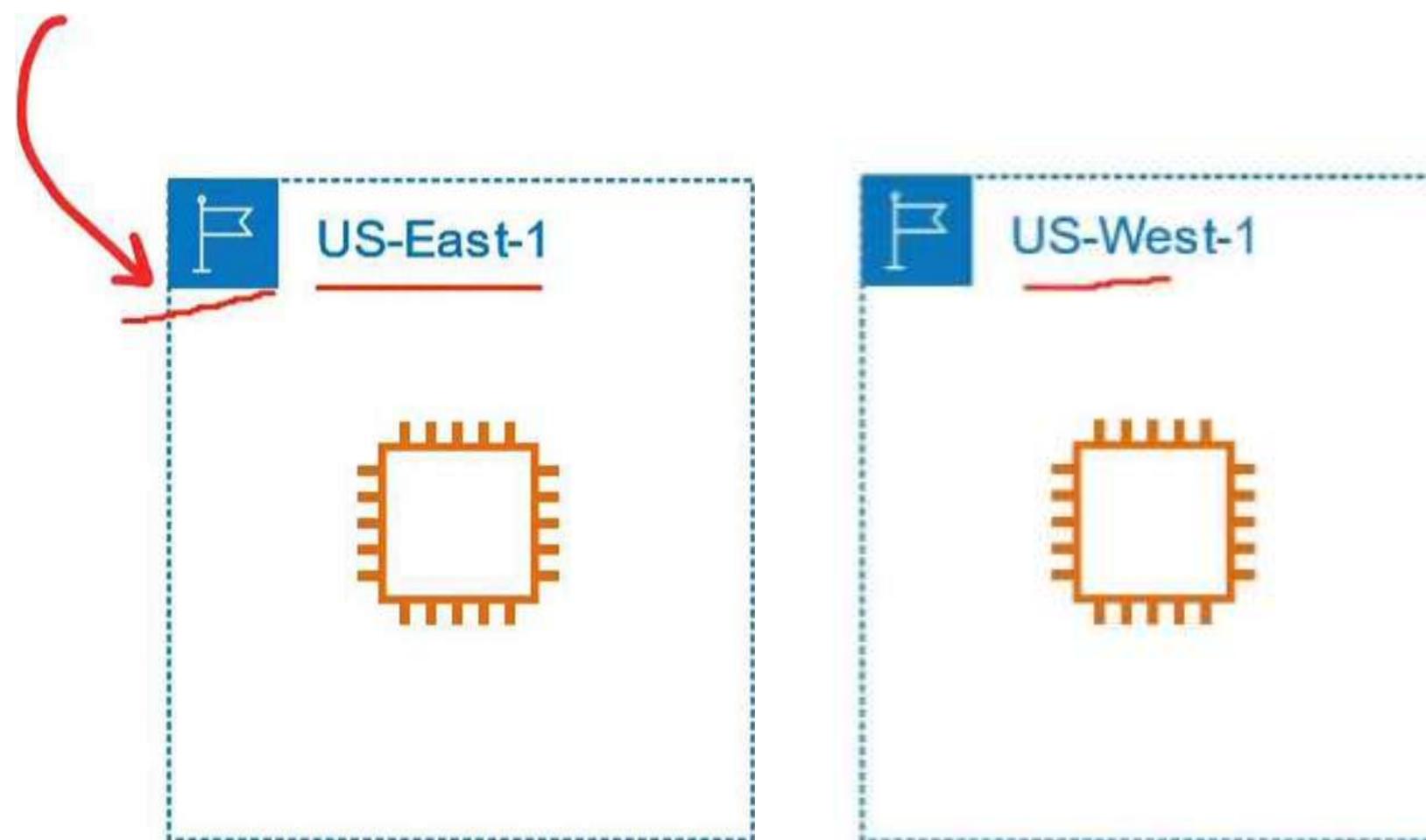


Global Infrastructure - Regions



Global Infrastructure - Regions

This is what a **region** will look like represented in an architectural diagram.



Global Infrastructure - Regions

Each region generally has three Availability Zones

- Some new users are limited to two eg. US-West

New services almost always become available first in

US-EAST

Not all AWS Services are available in all regions

All your billing information appears in **US-EAST-1**
(North Virginia)

The cost of AWS services vary per region

When you choose a region there are four factors you need to consider:

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 or latency to my end-users?



US East (Ohio) Region
Availability Zones: 3 Launched 2016

US West (Oregon) Region
Availability Zones: 4 Launched 2011 Local Zones: 3 Launched 2019

US West (Northern California) Region
Availability Zones: 3* Launched 2009

GovCloud (US-West) Region
Availability Zones: 3 Launched 2011

GovCloud (US-East) Region
Availability Zones: 3 Launched 2018

Canada (Central) Region**
Availability Zones: 3 Launched 2016

US East (Northern Virginia) Region
Availability Zones: 6 Launched 2006 Local Zones: 8 Launched 2020

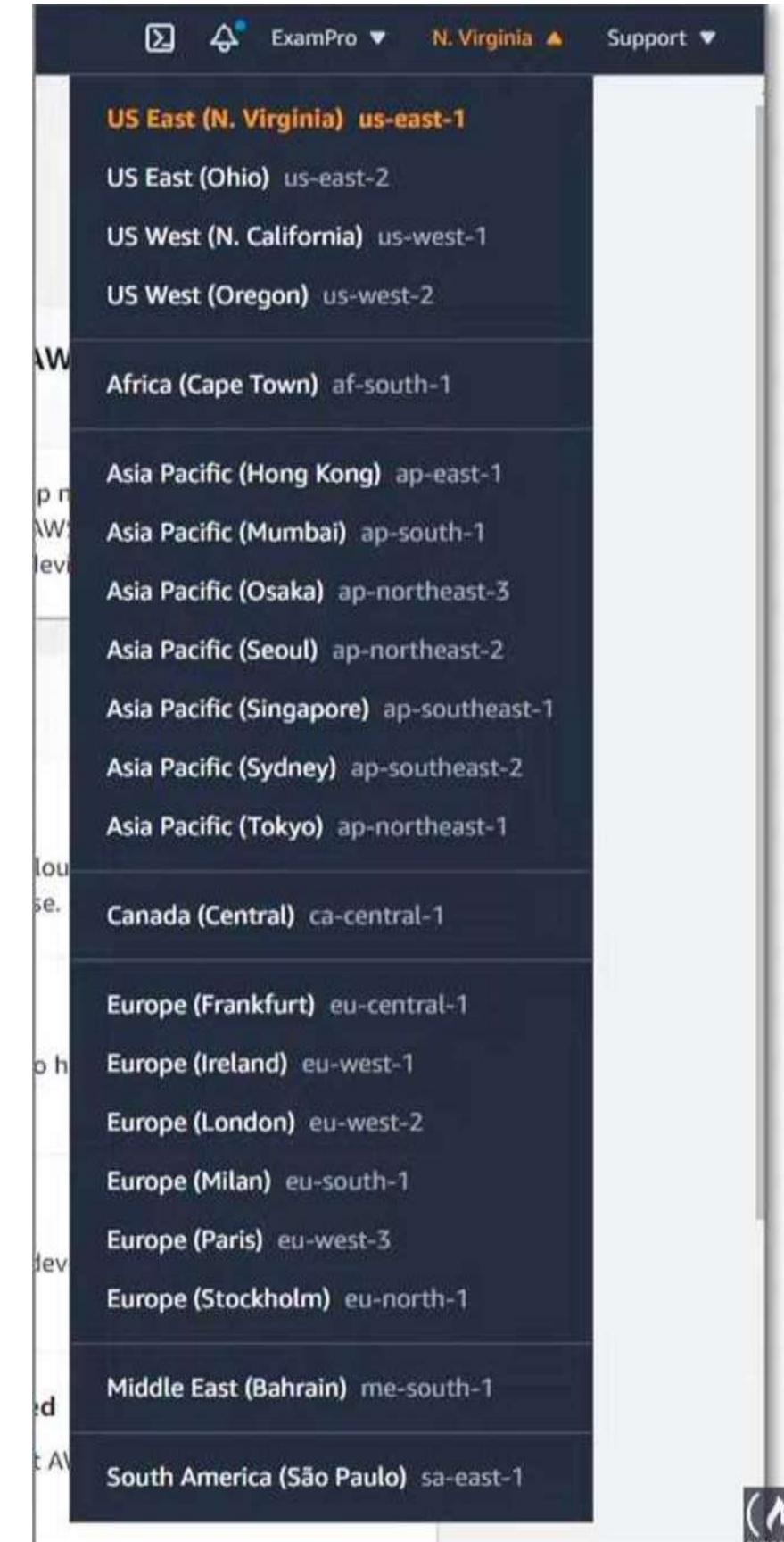
Global Infrastructure - Regional vs Global Services

Regional Services

AWS **scopes** their AWS Management Console on a selected Region.

This will determine where an AWS service will be launched and what will be seen within an AWS Service's console.

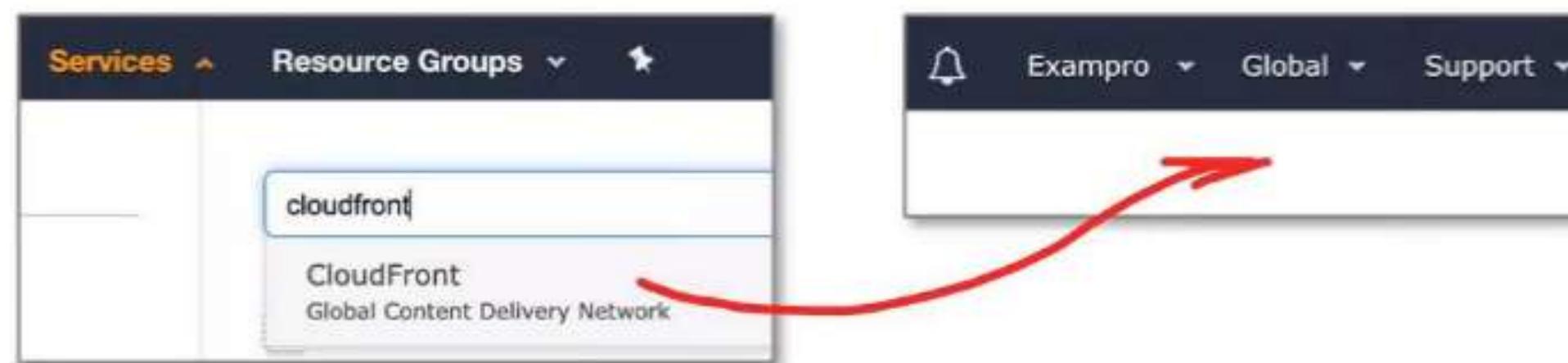
You generally don't explicitly set the Region for a service at the time of creation



Global Services

Some AWS Services operate across multiple regions and the region will be fixed to "Global"

E.g. Amazon S3, CloudFront, Route53, IAM



For these global services at the time of creation:

- There is 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

Global Infrastructure - Availability Zones

An **Availability Zone** (AZ) is physical location made up of one or more datacenter.

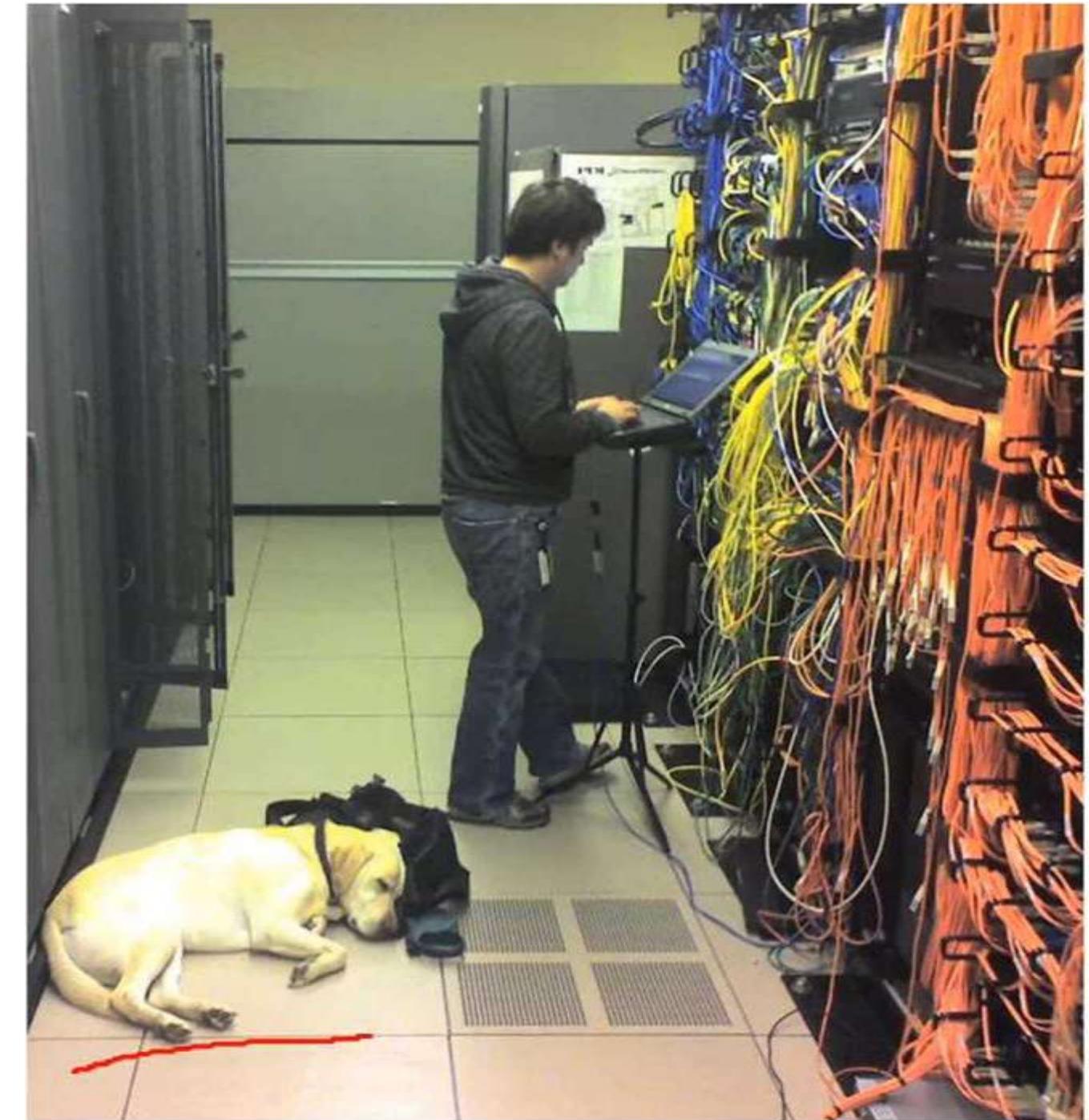
A datacenter is a secured building that contains hundreds of thousands of computers.

A region will ***generally contain 3 Availability Zones**

Datacenters within a region will be isolate from each other (different buildings). But they will be close enough to provide low-latency (< 10ms).

Its common practice to run workloads in at least 3 AZs to ensure services remain available in case one or two datacenters fail. (High Availability)

AZs are represented by a Region Code, followed by a letter identifier eg. **us-east-1a**



Global Infrastructure - Availability Zones

A Subnet is associated with an Availability Zone.

You never choose the AZ when launching resources. You choose the Subnet which is associated to the AZ.

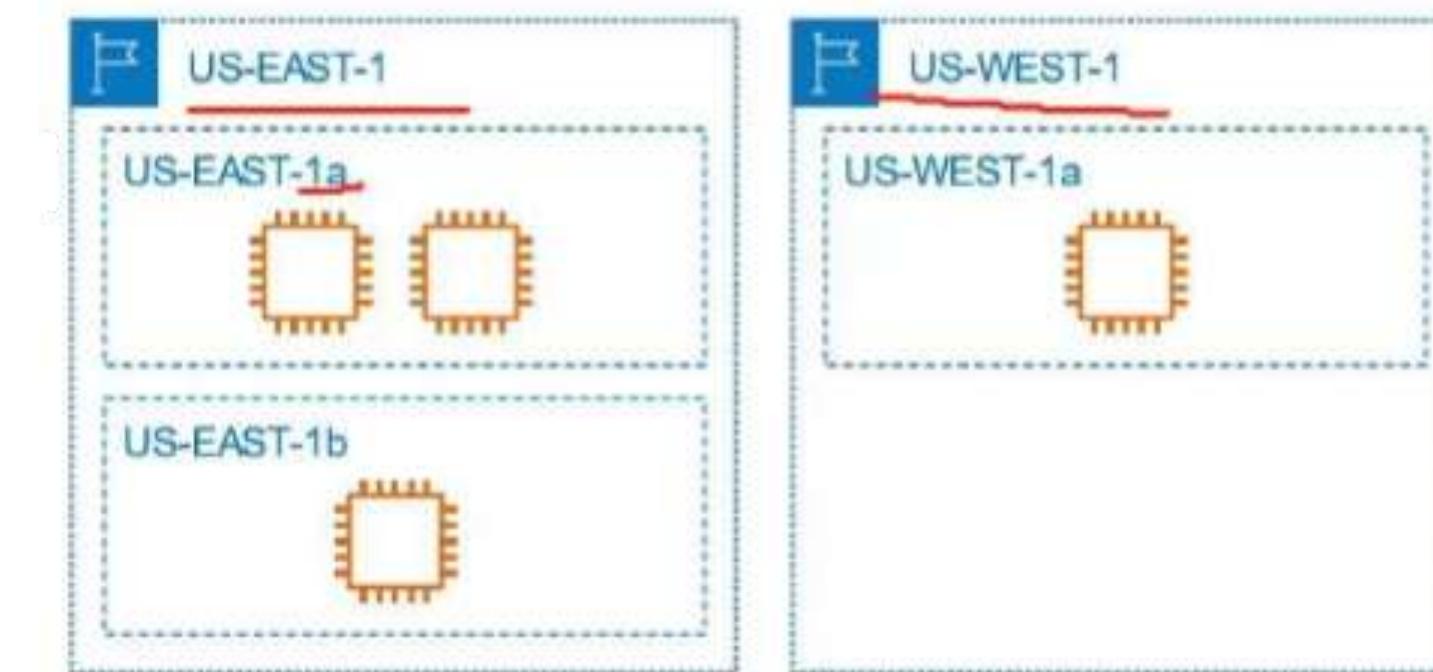


Subnet
Public IP
Placement group
Reservation

> No preference (default subnet in any Availability Zone)
subnet-d9de91f7 | Default in us-east-1c subnet-d0c28f8c |
Default in us-east-1a subnet-349fdf53 | Default in us-east-1b subnet-a8c2f8a7 | Default in us-east-1f subnet-b9db4c87 | Default in us-east-1e subnet-13869659 | Default in us-east-1d

Example of an architectural diagram, representing two AZs, the Subnets associated with those AZs, and EC2 instances (Virtual Machines) launched in those subnets

The US-EAST-1 region has 6 AZs
(the most Availability Zones of any region)



Global Infrastructure - Availability Zones

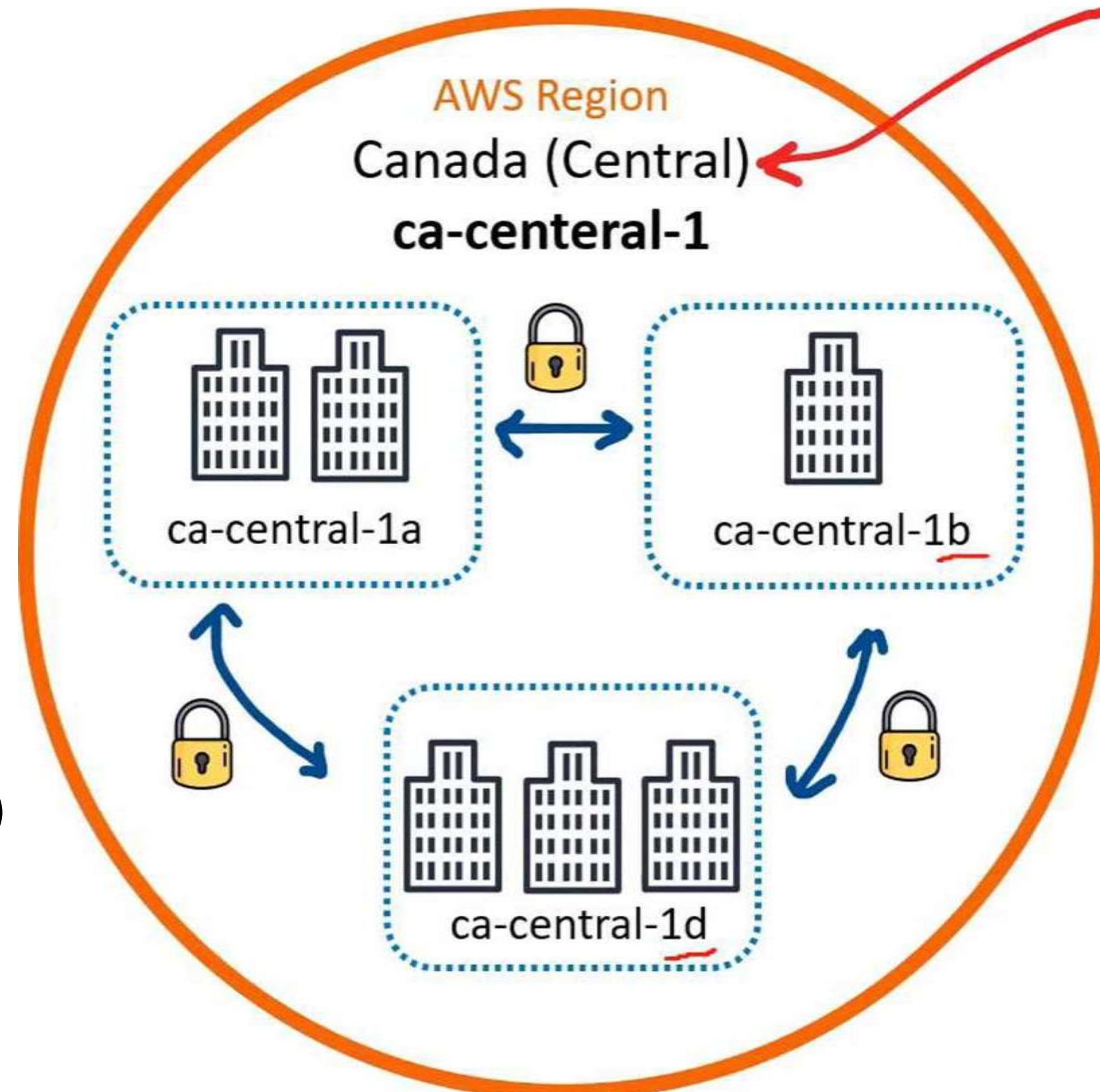
A region has multiple Availability Zones

An Availability Zone is made up of **one or more** datacenters

All AZs in an AWS Region are interconnected with high-bandwidth, low-latency networking, over fully redundant, dedicated metro fiber providing high-throughput, low-latency networking between

All traffic between AZs is encrypted

AZs are within 100 km (60 miles) of each other.



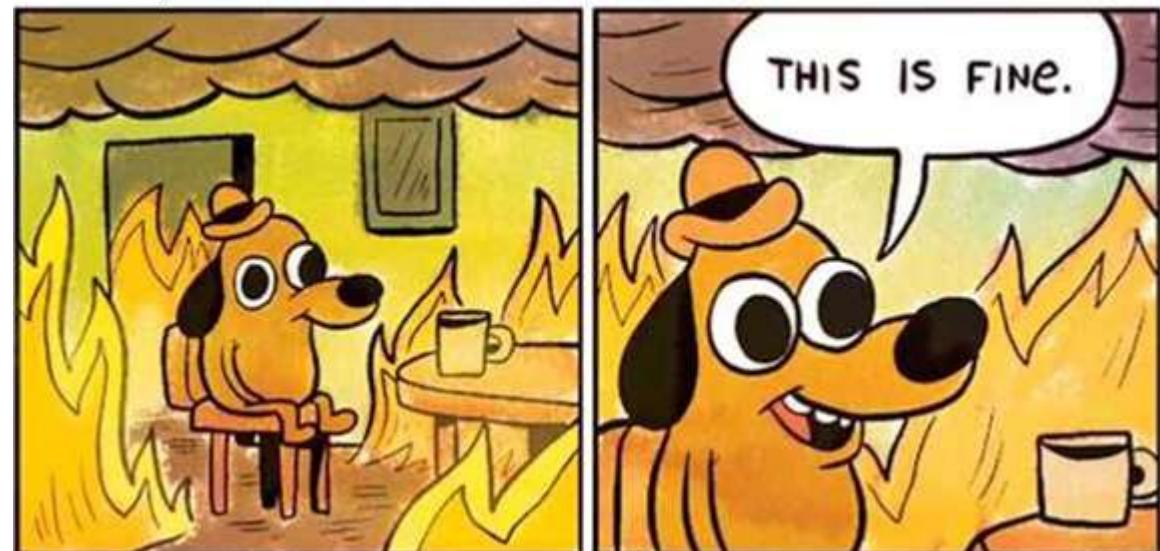
Montreal



(©stevenwright Upsplash

(A)

Global Infrastructure - Fault Tolerance



What is a fault domain?

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.

What is a fault level?

A fault level is a collection of fault domains.

Fault Domain

The scope of a fault domain could be:

- specific servers in a rack
- an entire rack in a datacenter
- an entire room in a datacenter
- the entire data center building

Its up to the Cloud Service Provider (CSPs) to define the boundaries of a domain

An AWS Region would be a **Fault Level**

A Availability Zone would be a **Fault Domain i**

Fault Level us-east-1 (Region)

Fault Domain us-east-1a (AZ)

Fault Domain us-east-1b (AZ)

Global Infrastructure - Fault Tolerance

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

- This achieves the greatest possible fault tolerance and stability

Each Availability Zone is **isolated**, but the Availability Zones in a Region are connected through low-latency links

Each Availability Zone is designed as an **independent failure zone**

- **A "Failure Zone" is AWS describing a Fault Domain.**

Failure Zone

- Availability Zones are physically separated within a typical metropolitan region and are located in lower risk flood plains
- discrete uninterruptible power supply (UPS) and onsite backup generation facilities
- data centers 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 Availability Zone.
- Availability Zones are all redundantly connected to multiple tier-1 transit providers



Multi-AZ for High Availability

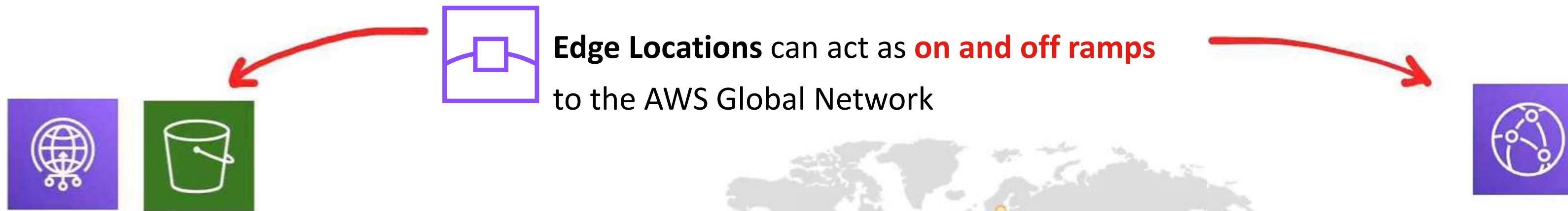
If an application is partitioned across AZs, companies are better isolated and protected from issues such as **power outages, lightning strikes, tornadoes, earthquakes**, and more.

AWS Global Network

The AWS Global Network represent the **interconnections between AWS Global Infrastructure**.

Commonly referred to as the "The Backbone of AWS"

Think of it as private expressway, where things can move very fast between datacenters.



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



VPC Endpoints

Ensuring your resources stay within the AWS Network and do no traverse over the public Internet.

Amazon CloudFront (CDN)

Uses Edge Locations as an off-ramp, to provide at the Edge storage and compute near the end user.

Global Infrastructure - Point of Presence (PoP)

Points of Presence (PoP) is an intermediate location between an AWS Region and the end user, and this location could be a datacenter or collection of hardware.

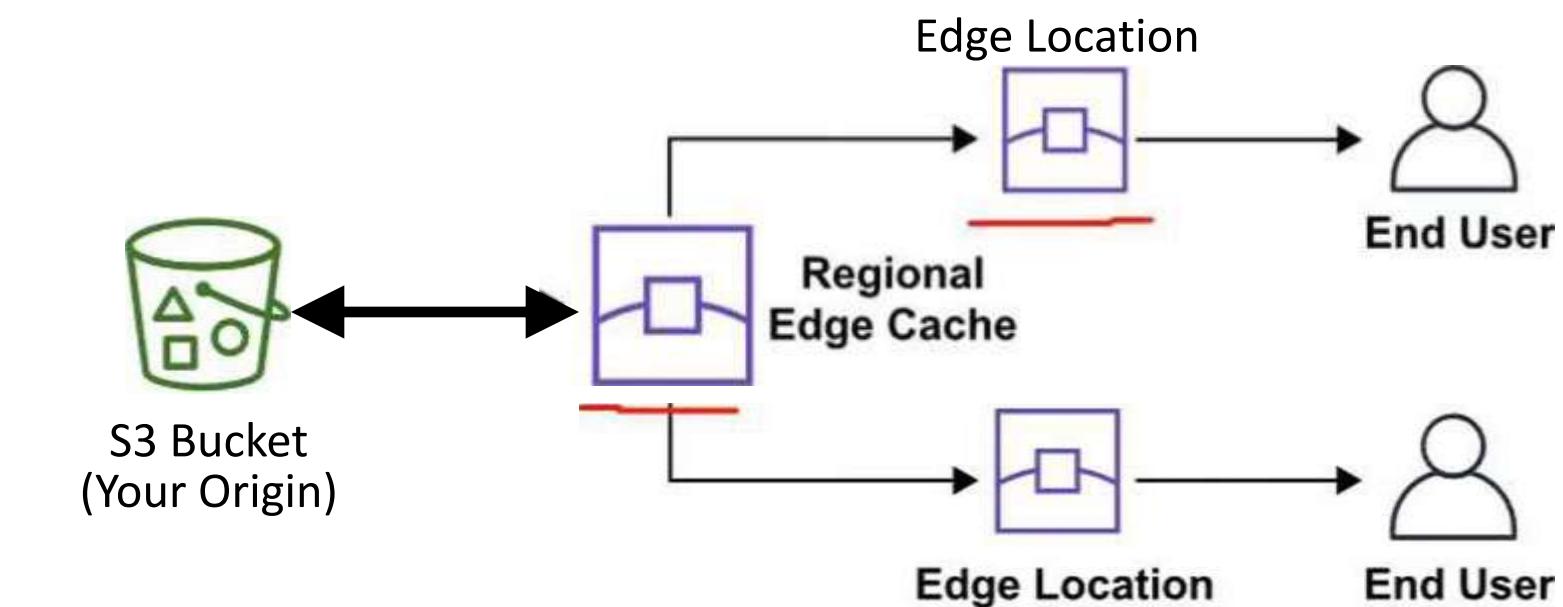
For AWS a Point of Presence is a data center **owned by AWS or a trusted partner** that is utilized by AWS Services related **for content delivery or expediated upload**.

PoP resources are:

- **Edge Locations**
- **Regional Edge Caches**

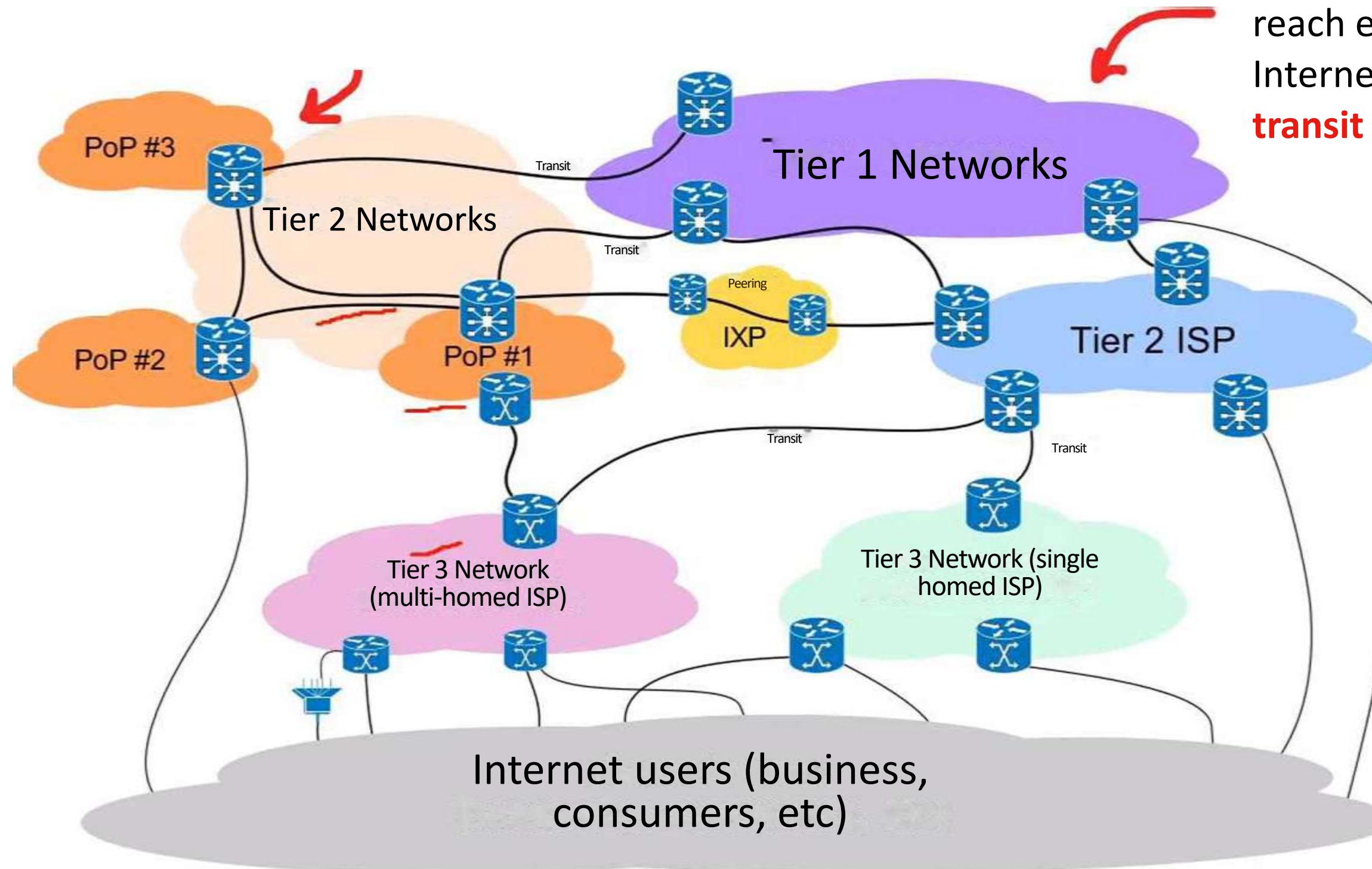
Edge Locations are datacenters that hold cached (copy) on the most popular files (eg. web pages, images and videos) so that the delivery of distance to the end users are reduced.

Regional Edge Locations are datacenters that hold much larger caches of less-popular files to reduce a full round trip and also to reduce the cost of transfer fees.



Global Infrastructure - Point of Presence (PoP)

PoPs live at the **edge/intersection** of two networks



Tier 1 network is a network that can reach every other network on the Internet **without purchasing IP transit or paying for peering**.

AWS Availability Zones are all redundantly connected to multiple **tier-1 transit providers**

Global Infrastructure - Point of Presence (PoP)

The following AWS Services use PoPs **for content delivery or expedited upload**.



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

- You point your website to CloudFront so that it will route requests to nearest Edge Location cache
- allows you to choose an **origin** (such as a web-server or storage) that will be source of cached
- caches the contents of what origin would returned 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 Accelerator are deployed within Edge Locations so you send user traffic to an Edge Location instead of directly to your web-application.

AWS Direct Connect

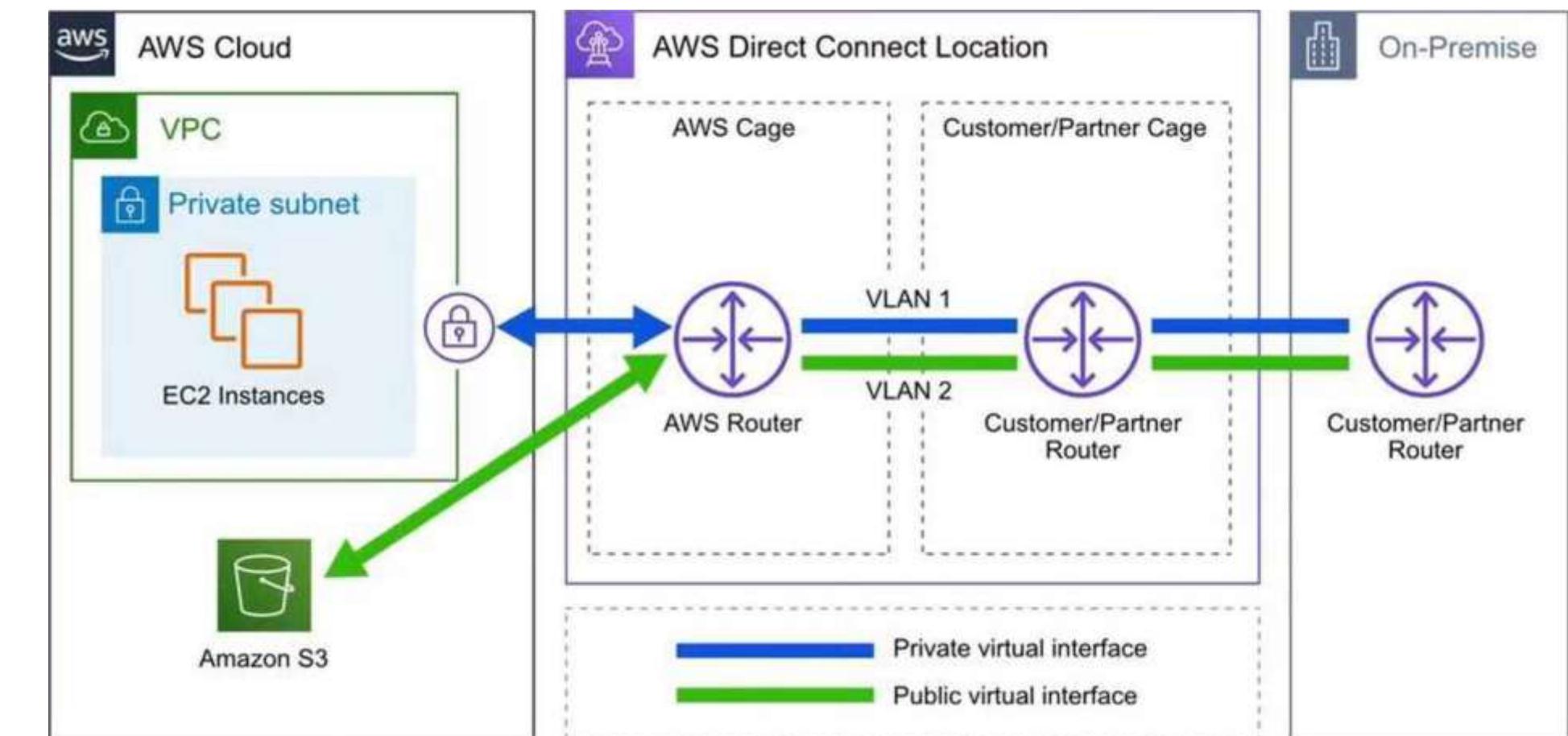


AWS Direct Connect is a **private/dedicated connection between your datacenter, office, co-location and AWS.**

Direct Connect has two **very-fast network connection options:**

1. Lower Bandwidth **50Mbps-500Mbps**
2. Higher Bandwidth **1Gbps or 10Gbps**

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



Helps **reduce network costs** and **increase bandwidth throughput**.
(great for high traffic networks)

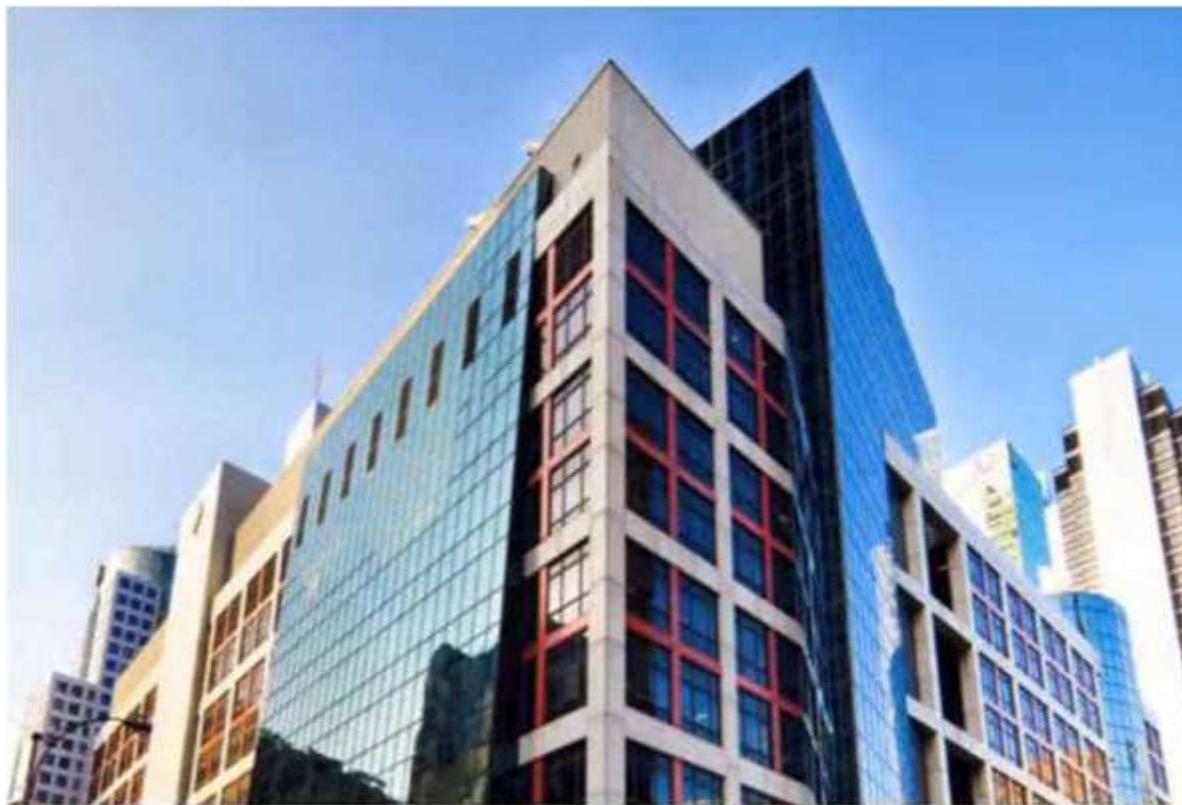
Provides a **more consistent network experience** than a typical internet-based connection,
(reliable and secure)

Global Infrastructure - Direct Connect Locations

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



A partnered datacenter in Toronto



Allied Data Centers

250 Front Street West Toronto

You would use the **AWS Direct Connect** service to order and establish a connection

Global Infrastructure - Local Zones



Local Zones are datacenters 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-1**
- Only specific AWS Services have been made available
 - EC2 Instance Types (T3, C5, R5, R5d, I3en, G4)
 - EBS (io1 and 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

To use Local Zones you need to Opt-In

Global Infrastructure - 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 to utilize their 5G networks

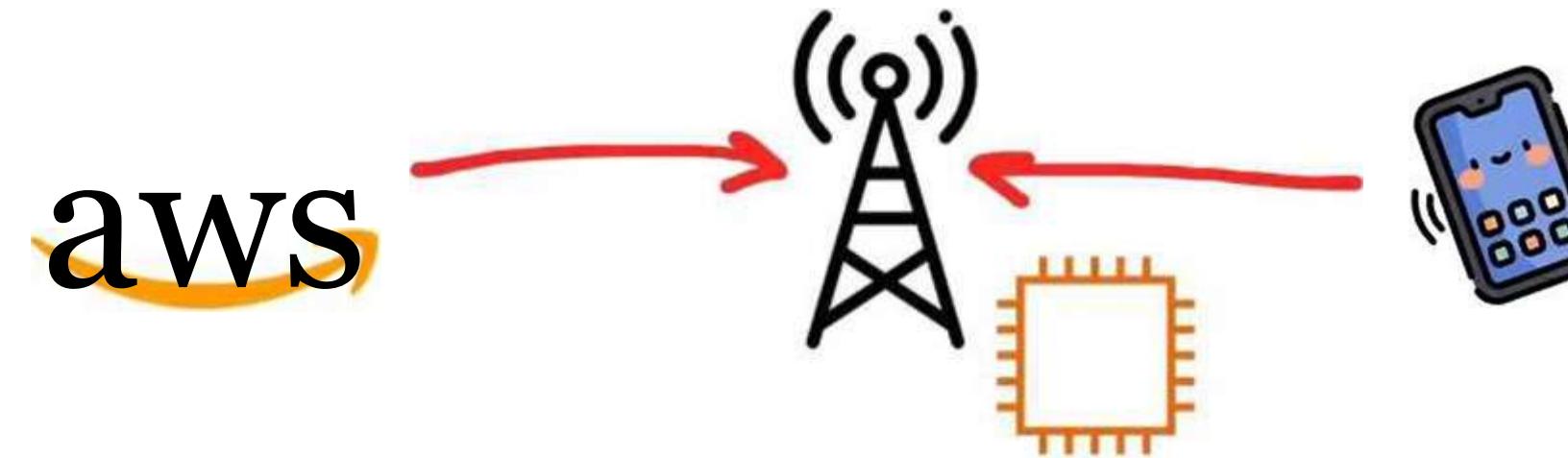
verizon

KDDI

O vodafone

*

SK telecom



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

Global Infrastructure - Data Residency

What is Data Residency?

The physical or geographic location of where an organization or cloud resources reside.

What is Compliance Boundaries?

A regulatory compliance (legal requirement) by a government or organization that describes where data and cloud resources are allowed to reside

What is Data Sovereignty?

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

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

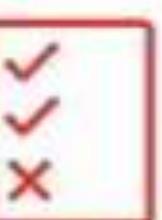


AWS Config is a Policy as Code service.

You can create rules to continuously check AWS resources configuration. If they deviate from your expectations you are alerted or AWS Config can in some cases auto-remediate.



AWS Outposts is **physical rack of servers** that you can put in your data center. Your data will reside wherever the Outpost physically resides



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

Global Infrastructure - AWS for Government



What is Public Sector?

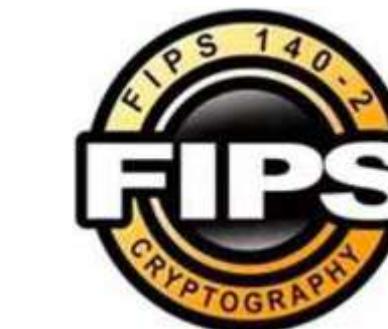
Public sectors include public goods and governmental services such as:

military
law enforcement
infrastructure
public transit

public education
health care
the government itself

AWS can be utilized by public sector or organizations developing cloud workloads for the public sector.

AWS achieves this by meeting **regulatory compliance programs** along with specific governance and security controls



AWS has special regions for US regulation called **GovCloud**

Global Infrastructure - GovCloud (US)



Federal Risk and Authorization Management Program (FedRAMP)
a US government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.

What is GovCloud?

A Cloud Service Provider (CSP) generally will offer an **isolated region** to run FedRAMP workloads.



AWS GovCloud Regions allow customers to host sensitive **Controlled Unclassified Information** and 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 and 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

Global Infrastructure - AWS in China



AWS China is the AWS cloud offerings in Mainland China.

AWS China is completely isolate *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 have a Chinese Business License (ICP license)

Not all services are available in china eg. Route53

Running in Mainland China (instead of Singapore) means you would not need to traverse the The Great Firewall.

AWS has two Regions in Mainland China:



Ningxia CN-NorthWest-1
Operated by NSWCF



Beijing CN-North-1
operated by SINNET



Global Infrastructure - Sustainability



Amazon co-founded the Climate Pledge to achieve Net-Zero Carbon Emissions by 2040 across all of Amazon's business (this includes AWS)

sustainability.aboutamazon.com

AWS Cloud's Sustainability goals are composed of three parts:

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 centers surveyed.

3. Water Stewardship

Direct evaporative technology to cool our data center

Use of non-potable water for cooling purposes (recycled water)

On-site water treatment allows us to remove scale-forming minerals and reuse water for more cycles

Water efficiency metrics to determine and monitor optimal water use for each AWS Region

AWS purchases and retires environmental attributes to cover the non-renewable energy for AWS Global Infrastructure:

- Renewable Energy Credits (RECs)
- Guarantees of Origin (GOs)

Global Infrastructure - AWS Ground Station



AWS Ground Station is a fully managed service that **lets you control satellite communications**, process data, and scale your operations without having to worry about building or managing your own ground station infrastructure

Use cases for Ground Station: To use Ground Station:

- weather forecasting
- You schedule a Contact (select satellite, start and end time, and the ground location)
- surface imaging
- communications
- video broadcasts
- use the AWS Ground Station EC2 AMI to launch EC2 instances that will uplink and downlink data during the contact or receive downlinked data in an Amazon S3 bucket.

Use Case:

A company reaches an agreement with a Satellite Imagery Provider to take satellite photos of a specific region. They use AWS Ground Station to communicate that company's Satellite and download the S3 image data.



@isidurumm on Unsplash

SUBSCRIBE

Global Infrastructure - AWS Outposts



AWS Outposts is a fully managed service that offers the same AWS infrastructure, AWS services, APIs, and tools to virtually any datacenter, co-location space, or on-premises facility for a truly consistent hybrid experience.

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

42U Rack

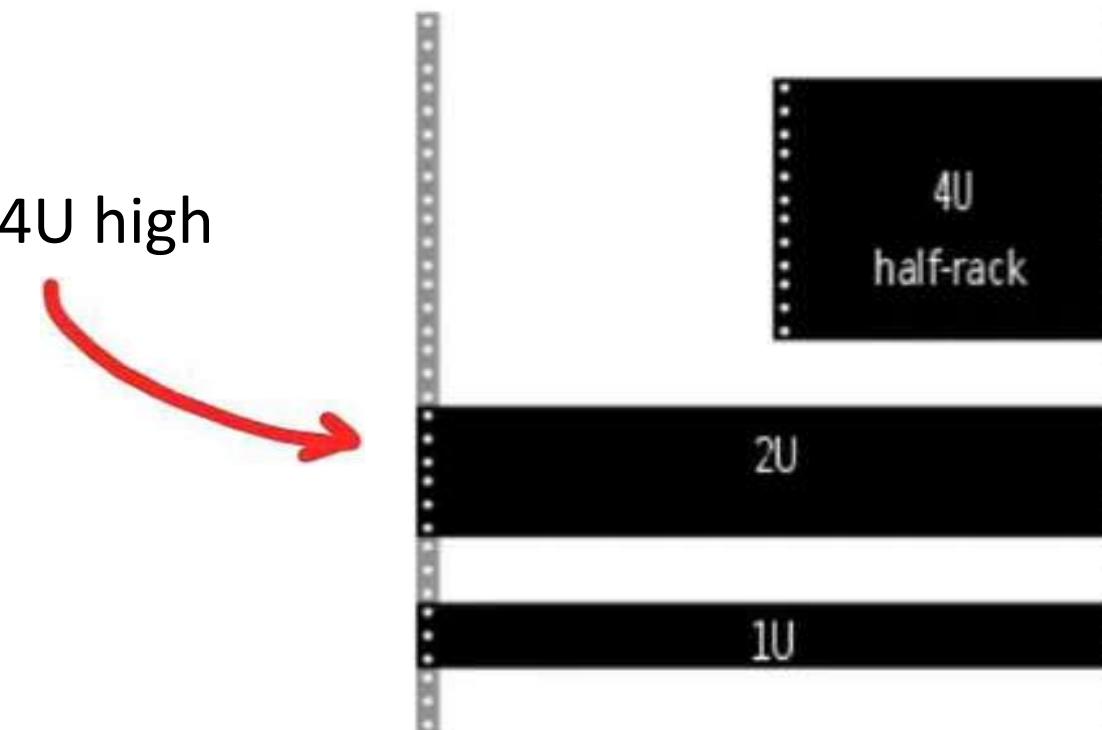


What is a Server Rack? Rack Heights

A frame design to hold and U stands for "rack units" or "U spaces" with is equal to 1.75 inches, organize IT equipment.

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



Global Infrastructure - AWS Outposts

AWS Outposts comes in 3 form factors: 42U, 1U and 2U

This a full rack of servers provided by AWS These are servers that you can place into your existing racks:

42U



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 8TB of local NVMe storage

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

AWS Application Programming Interface (API)

What is an Application Programming Interface (API)?

An API is software that allows two applications/services to talk to each other.

The most common type of API is via HTTP/S requests.

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



Each AWS Service has its own **Service Endpoint** which you send requests

```
GET / HTTP/1.1 1
host: monitoring.us-east-1.amazonaws.com
x-amz-target: GraniteServiceVersion20100801.GetMetricData x-
amz-date: 20180112T092034Z

Authorization: AWS4-HMAC-SHA256
Credential=REDACTEDREDACTED/20180411/
Content-Type: application/json Accept: application/json Content-Encoding: amz-1.0
Content-Length: 45 Connection: keep-alive
```



To authorize use you will need generate a **signed request**

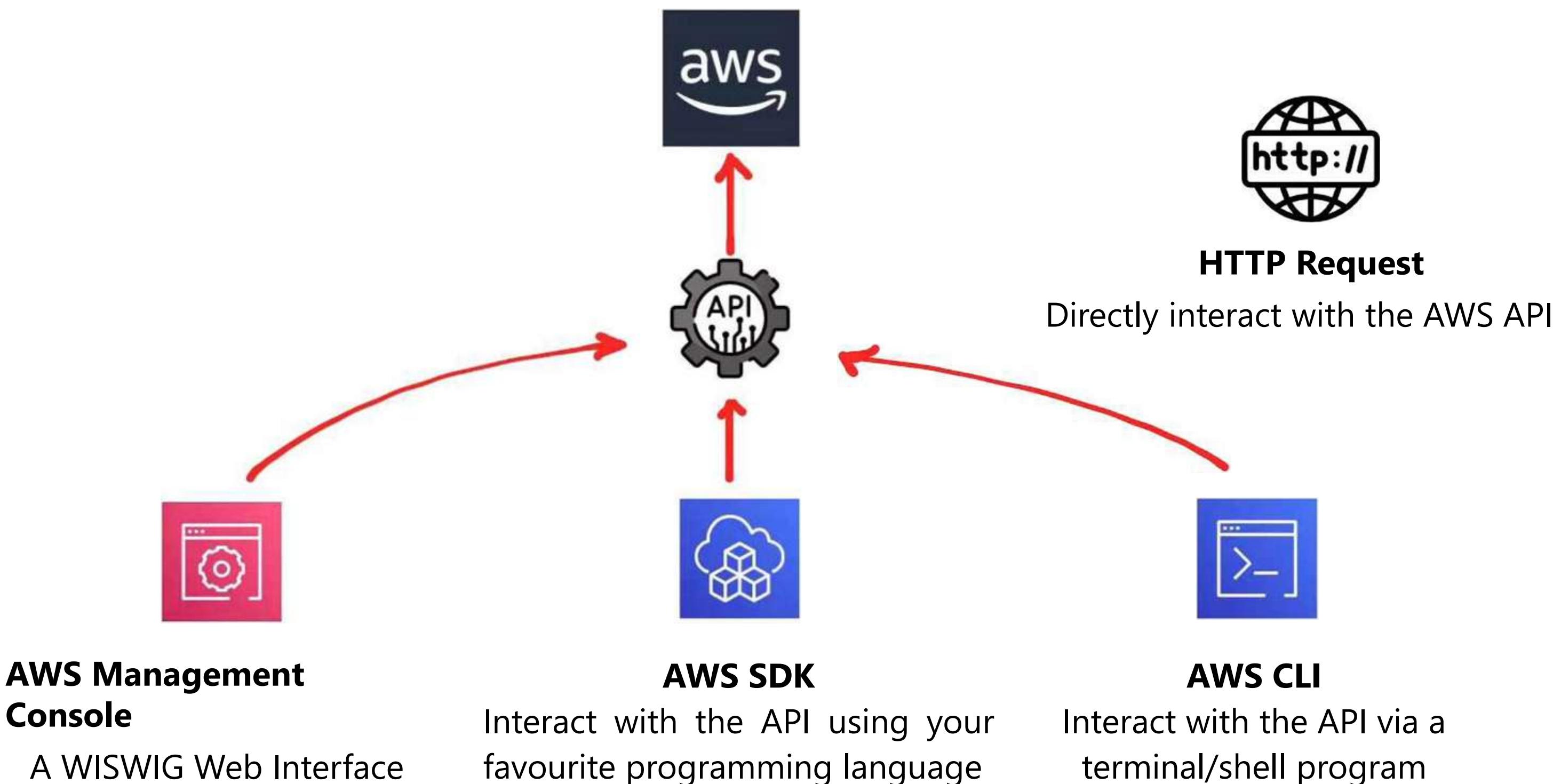
You make a separate request with your AWS credentials and get back a token.

You also need to provide an **ACTION**

And accompanying **PARAMETERS** as a **PAYOUT**

AWS Application Programming Interface (API)

Rarely do users directly send HTTP requests directly to the AWS API.
It's much easier to interact with the API via a variety of Developer Tools



AWS Management Console

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

The screenshot shows the AWS Management Console homepage. The top navigation bar includes the AWS logo, a search bar, and various icons for notifications and account settings. The main content area is titled "Console Home". On the left, there's a sidebar titled "Recently visited" with links to EC2, VPC, Billing and Cost Management, IAM, EC2 Global View, and Resource Groups & Tag Editor. The right side shows the "Applications" section, which is currently empty. It features a "Create application" button, a "Region" dropdown set to "Asia Pacific (Mumbai)", and a "Find applications" search bar. Below these are sorting options for "Name", "Description", and "Region", and a "No applications" message with a "Get started by creating an application." link.

Point and Click to manually launch and configure AWS resources with limited programming knowledge.

This is known as "**ClickOps**" since you can perform all your system operations via clicks.

AWS Management Console - Service Console

AWS Service each have their own customized console.

You can access these consoles by **searching** the service name.

The screenshot shows the AWS Management Console interface. At the top, there is a search bar with the text "ec2" and a magnifying glass icon. To the right of the search bar are several icons: a close button ("X"), a refresh button, a bell icon, a question mark icon, a gear icon, and a dropdown menu set to "Asia Pacific (Mumbai)".

Below the search bar, the page title is "Services". On the left side, there is a sidebar with the following links: "Services" (which is currently selected and highlighted in blue), "Features", "Resources New", "Documentation", "Knowledge articles", "Marketplace", "Blog posts", "Events", and "Tutorials".

The main content area displays three service cards:

- EC2**: Virtual Servers in the Cloud. It features an orange icon with a server and a gear, the text "EC2", and a description "Virtual Servers in the Cloud". There is also a star icon in a rounded box to the right.
- EC2 Image Builder**: A managed service to automate build, customize and deploy OS images. It features an orange icon with a server and a gear, the text "EC2 Image Builder", and a description "A managed service to automate build, customize and deploy OS images". There is also a star icon in a rounded box to the right.
- EC2 Global View**: EC2 Global View provides a global dashboard and search functionality that lets you f... It features an orange icon with a server and a gear, the text "EC2 Global View", and a description "EC2 Global View provides a global dashboard and search functionality that lets you f...". There is also a star icon in a rounded box to the right.

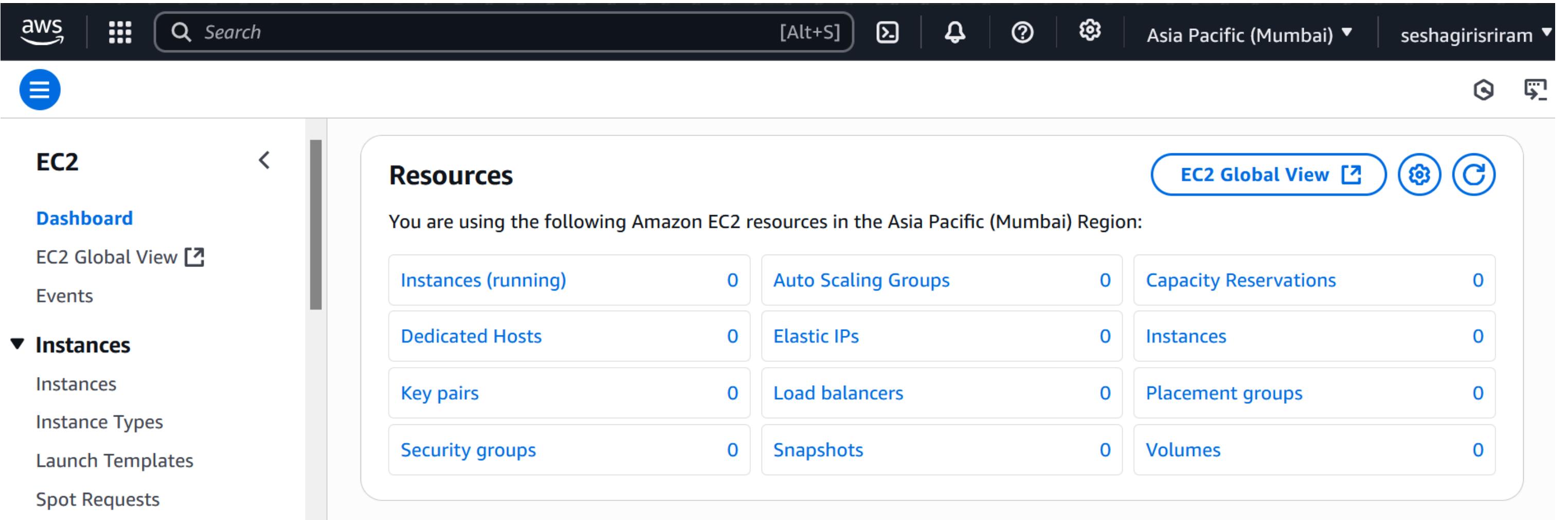
AWS Management Console - Service Console

AWS Service each have their own customized console.

Some AWS Services Console will act as an umbrella console containing many AWS Services: eg



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



The screenshot shows the AWS Management Console interface for the Amazon EC2 service. The top navigation bar includes the AWS logo, a search bar, and various global settings like region (Asia Pacific (Mumbai)) and user (seshagirisriram). On the left, a sidebar menu for 'EC2' lists options like 'Dashboard', 'EC2 Global View', 'Events', and 'Instances'. Under 'Instances', sub-options include 'Instances', 'Instance Types', 'Launch Templates', and 'Spot Requests'. The main content area is titled 'Resources' and displays a summary of EC2 resources in the Asia Pacific (Mumbai) Region. It shows 0 instances (running), 0 auto scaling groups, 0 capacity reservations, 0 dedicated hosts, 0 elastic IPs, 0 instances, 0 key pairs, 0 load balancers, 0 placement groups, 0 security groups, 0 snapshots, and 0 volumes.

Resource Type	Count
Instances (running)	0
Auto Scaling Groups	0
Capacity Reservations	0
Dedicated Hosts	0
Elastic IPs	0
Instances	0
Key pairs	0
Load balancers	0
Placement groups	0
Security groups	0
Snapshots	0
Volumes	0

AWS Account ID

Every AWS Account has a unique Account ID.

The **Account ID** can be easily found by dropping down the current user in the Global Navigation

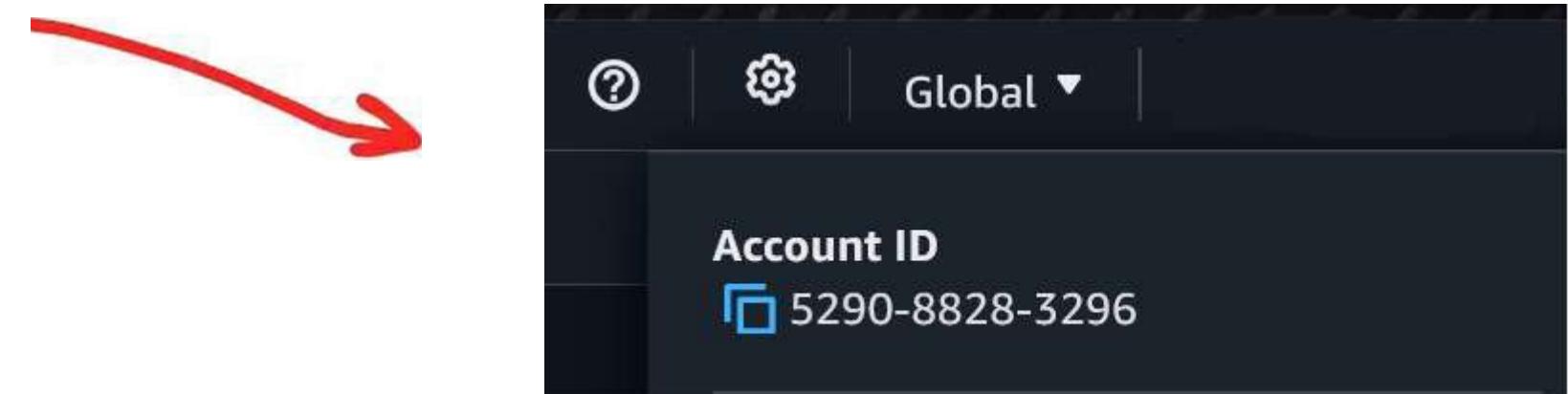
The AWS Account ID is composed of 12 digits eg:

- 123456789012
- 121212121212 • 498241098510

The AWS Account ID is used

- when logging in with a non-root user
- Cross-Account Roles
- Support Cases

It is generally good to keep your Account ID private as it is one of many components used to identify an account for attack by a malicious actor.

A screenshot of the "IAM user sign in" form. A large red arrow points from the text "The AWS Account ID is used" to the "Account ID or alias" input field. The field contains the placeholder text "(Don't have?)". Other fields include "Remember this account" (unchecked), "IAM username" (empty), "Password" (empty), "Show Password" (unchecked), and "Having trouble?" (link).

AWS Tools for PowerShell



What is PowerShell?

PowerShell is a task automation and configuration management framework.

A **command-line shell** and a **scripting language**.

Unlike most shells, which accept and return text, PowerShell is built on top of the .NET Common Language Runtime (CLR), and accepts and returns .NET objects.

A screenshot of a Windows PowerShell window titled "Windows PowerShell". It displays network adapter information for three adapters: "Link-local IPv6 Address : fe80::541f:9e31:7df6:9847%22", "Ethernet adapter vEthernet (New Virtual Switch)", and "Unknown adapter Local Area Connection". Each adapter shows its IPv4 address, subnet mask, and default gateway. The prompt at the bottom is "PS C:\Users\Andrew>".

AWS Tools for PowerShell lets you interact with the AWS API via PowerShell Cmdlets

Cmdlet is a special type of command in PowerShell in the form of capitalized verb-and-noun e.g. *New-S3Bucket*

A screenshot of a PowerShell window with a dark background. It shows the command "PS > New-S3Bucket -BucketName website-example -Region us-west-2" being run. Below it, a table displays the results: "CreationDate" (8/16/19 8:45:38 PM) and "BucketName" (website-example). The prompt at the bottom is "PS C:\Users\Andrew>".

Amazon Resource Name (ARNs)

Amazon Resource Names (ARNs) uniquely identify AWS resources.

ARNs are required to specify a resource unambiguously across all of AWS

The ARN has the following *arn:partition:service:region:account-id:resource-id*

format variations

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

Resource ID

Could be a number name or path:

- user/Bob
- instance/i-1234567890abcdef0

Service - Identifies the service

- ec2
- s3
- iam

Region - which AWS resource

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

Account ID .

121212121212 •
123456789012

In the AWS Management Console its common to be able
to copy the ARN to your clipboard



arn:aws:s3:::my-bucket

Name	my-webserver-alb
ARN	arn:aws:elasticloadbalancing:us-east-1:123456789012:loadbalancer/app/my-webserver-alb/31e9d2ce26643cd8

Copied

Paths 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)

What is a CLI?

A 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.

What is a Terminal?

A terminal is a text only interface (input/output environment)

What is a Console?

A console is a physical computer to physically input information into a terminal

What is a Shell?

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

- **Bash**
- Zsh
- PowerShell



The screenshot shows a Windows 95-style MS-DOS Prompt window titled "MS-DOS Prompt". The window contains the following text:

```
Microsoft(R) Windows 95  
(C)Copyright Microsoft Corp 1981-1996.  
C:\WINDOWS>command.com /?  
Starts a new copy of the Windows Command Interpreter.  
COMMAND [[drive:]path] [device] [/E:nnnnn] [/L:nnnn] [/U:nnn] [/P] [/MSG]  
[drive:]path Specifies the directory containing COMMAND.COM.  
device Specifies the device to use for command input and output.  
/E:nnnnn Sets the initial environment size to nnnnn bytes.  
(nnnnn should be between 256 and 32,768).  
/L:nnnn Specifies internal buffers length (requires /P as well).  
(nnnn should be between 128 and 1,024).  
/U:nnn Specifies the input buffer length (requires /P as well).  
(nnn should be between 128 and 255).  
/P Makes the new Command Interpreter permanent (can't exit).  
/MSG Stores all error messages in memory (requires /P as well).  
/LOW Forces COMMAND to keep its resident data in low memory.  
/Y Steps through the batch program specified by /C or /K.  
/C command Executes the specified command and returns.  
/K command Executes the specified command and continues running.  
C:\WINDOWS>
```

People commonly (erroneously) use **Terminal, Shell or Console** to generally describe interacting with a Shell.

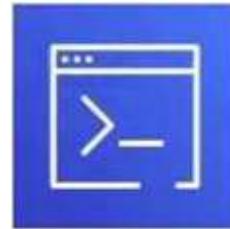
AWS Command Line Interface (CLI)

```
sesha ➤ 📂 ~ ➤ ✘ aws --v ➤ aws default@ap-south-1
aws-cli/2.26.6 Python/3.13.2 Windows/11 exe/AMD64
in cmd at 14:57:23
```

```
PowerShell 7.5.1
Loading personal and system profiles took 2673ms.
sesha ➤ 📂 System32 ➤ 0ms ➤ 🌐 aws --version ➤ pwsh 86% 15:05:47
aws-cli/2.26.6 Python/3.13.2 Windows/11 exe/AMD64
sesha ➤ 📂 System32 ➤ 1.11s ➤ 🌐 pwsh 86% 15:05:56
```

```
sidkalpop@SRIRAM-LENOVO:~$ aws --version
aws-cli/2.27.5 Python/3.13.2 Linux/5.15.167.4-microsoft-standard-WSL2 exe/x86_64.ubuntu.24
sidkalpop@SRIRAM-LENOVO:~$
```

AWS Command Line Interface (CLI)



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 diagram illustrates the AWS CLI process. On the left, a terminal window shows a complex multi-line AWS CLI command for filtering EC2 instances based on tags. A red arrow points from this command to the right, where the output is displayed as a tabular format titled 'DescribeInstances'. The output lists three instances: one in 'us-east-2b' with ID i-057750d42936e468a and name 'my-prod-server', and two in 'us-east-2a' with IDs i-001efd250faaa6ffa and i-027552a73f021f3bd and names 'test-server-1' and 'test-server-2' respectively.

```
aws ec2 describe-instances \
--filters Name=tag-key,Values=Name \
--query 'Reservations[*].Instances[*].
{Instance:InstanceId,AZ:Placement.AvailabilityZone,Name:Tags[?
Key==`Name`][0].Value}' \
--output table
```

AZ	Instance	Name
us-east-2b	i-057750d42936e468a	my-prod-server
us-east-2a	i-001efd250faaa6ffa	test-server-1
us-east-2a	i-027552a73f021f3bd	test-server-2



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**.



You can use the **AWS SDK** 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++

```
s3 = Aws::S3::Resource.new({ region:  
aws_default_region, credentials:  
Aws::Credentials.new( aws_access_key_id,  
aws_secret_access_key  
)  
>)  
bucket = s3.bucket s3_bucket  
file = File.open file_path  
md5 = Digest::MD5.hexdigest file.read  
md5 = Base64.encode64([md5].pack("H*")).strip  
attrs = {  
key: data["path"], body:  
10.read(ftie), content_md5:  
md5  
>  
resp = bucket.put_object(attrs)
```

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!

Preinstalled Tools

AWS CLI, Python, Node.js git, make, pip, sudo, tar, tmux, vim, wget, and zip and more

Storage included

1 GB of storage free per AWS region

Saved files and settings

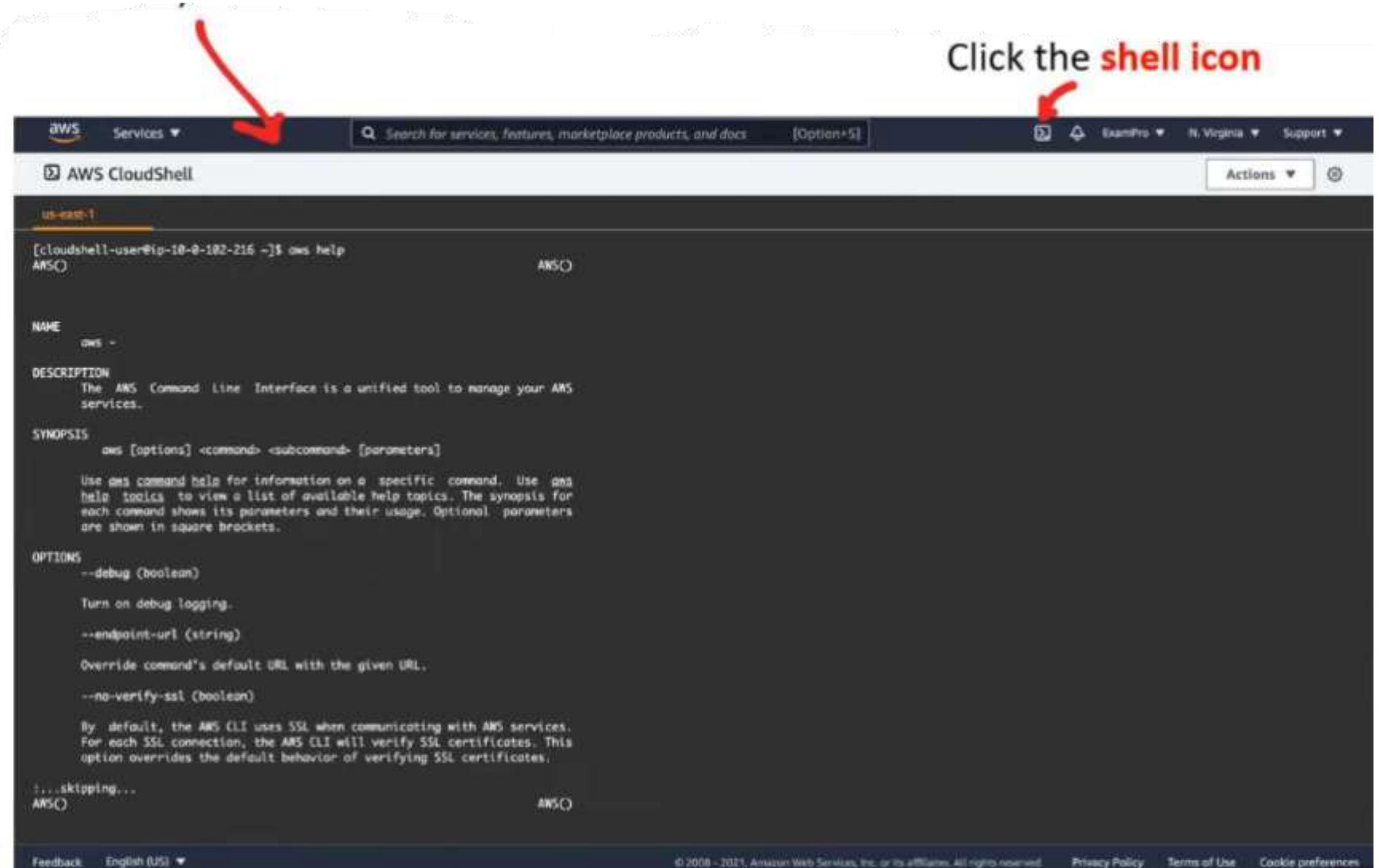
Files saved in your home directory are available in future sessions for the same AWS region

Shell Environments

Seamlessly switch between

- Bash
- PowerShell
- Zsh

AWS CloudShell is ovoilable in select regions



Infrastructure as Code (IaC)

Infrastructure as Code (IaC)

You write a configuration script to **automate creating, updating or destroying** cloud infrastructure.

- IaC is a **blueprint** of your infrastructure.
- IaC allows you to easily **share, version or inventory** your cloud infrastructure.

AWS has two offerings for writing Infrastructure as Code.



AWS CloudFormation (CFN)

CFN is a Declarative IaC tool



AWS Cloud Development Kit (CDK)

CDK is an Imperative 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

Imperative

- You say what you want, and 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

CloudFormation

AWS CloudFormation 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.

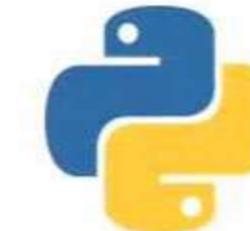
CloudFormation can be easier for DevOps Engineers who do not have a background in web programming languages.

Since CDK generates out CloudFormation its still important to be able to read and understand CloudFormation in order to debug IaC stacks.

```
Ec2Instance
  Type AWS EC2 Instance
  Properties:
    ImageId:
      Fn::FindInMap:
        - "RegionMap"
        Ref: "AWS::Region"
        - "AMI"
    KeyName:
      Ref: "KeyName"
    NetworkInterfaces:
      AssociatePublicIpAddress "true"
      DeviceIndex: "0"
      GroupSet:
        Ref "myVPCEC2SecurityGroup"
      SubnetId:
        Ref "PublicSubnet"
```

Cloud Development Kit

AWS CDK allows you to use your favorite programming language to write Infrastructure as Code (IaC)



TypeScript v NodeJS Python Java ASP.NET

```
const bucket = new Bucket(this, 'MyBucket');
const result = bucket.addToResourcePolicy(new iam.PolicyStatement({ actions:
  ['s3:GetObject'],
  resources: [bucket.arnForObjects('file.txt')], principals:
  [new iam.AccountRootPrincipal()],
> }):
```

- 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 and Integration Testing

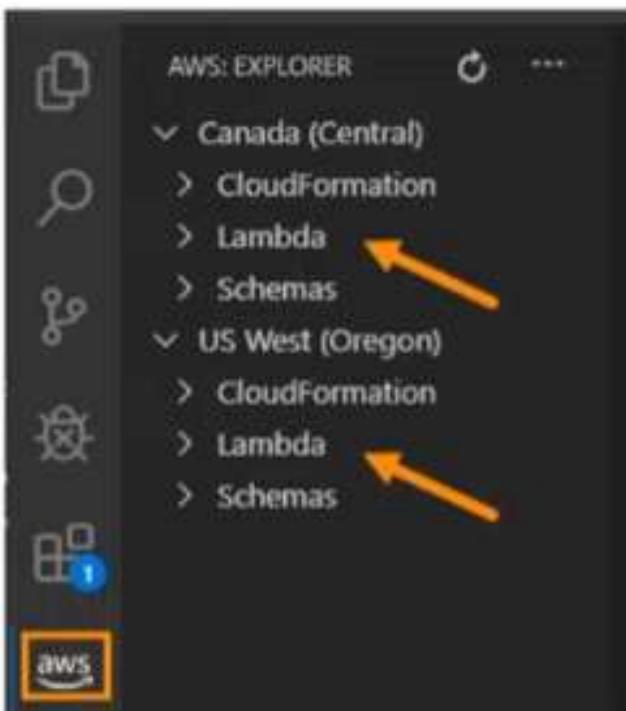
AWS SDK looks similar, but the key difference is CDK ensures Idempotent of your Infrastructure

AWS Toolkit for VSCode

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

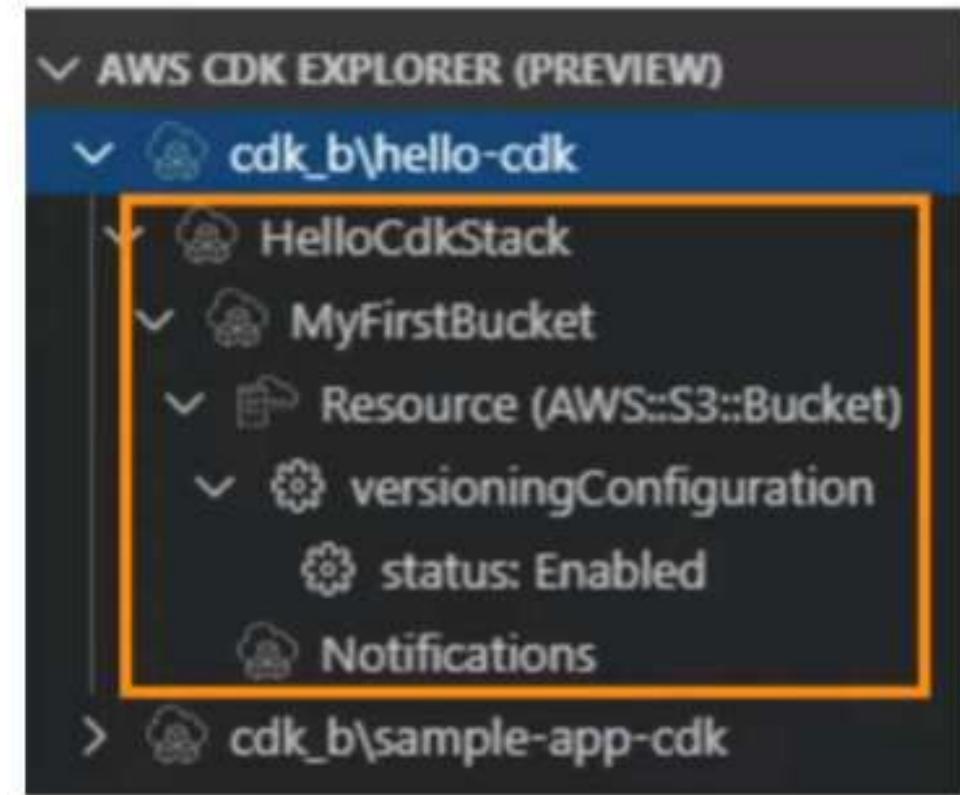
1. AWS Explorer

Explore a wide range of AWS resources to your linked AWS Account



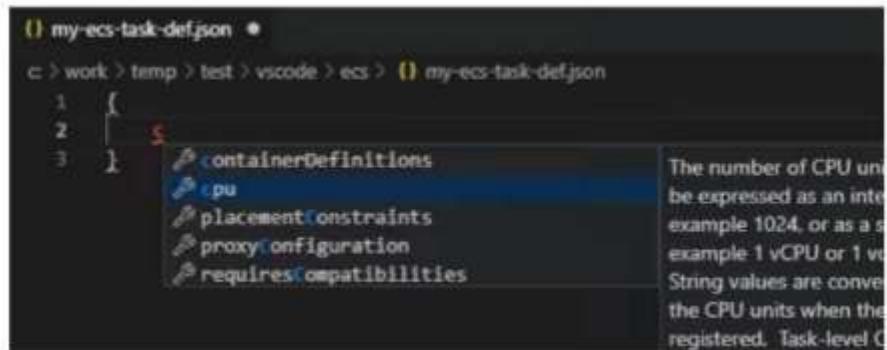
2. AWS CDK Explorer (PREVIEW)

Allows you to explore your stacks defined by CDK.



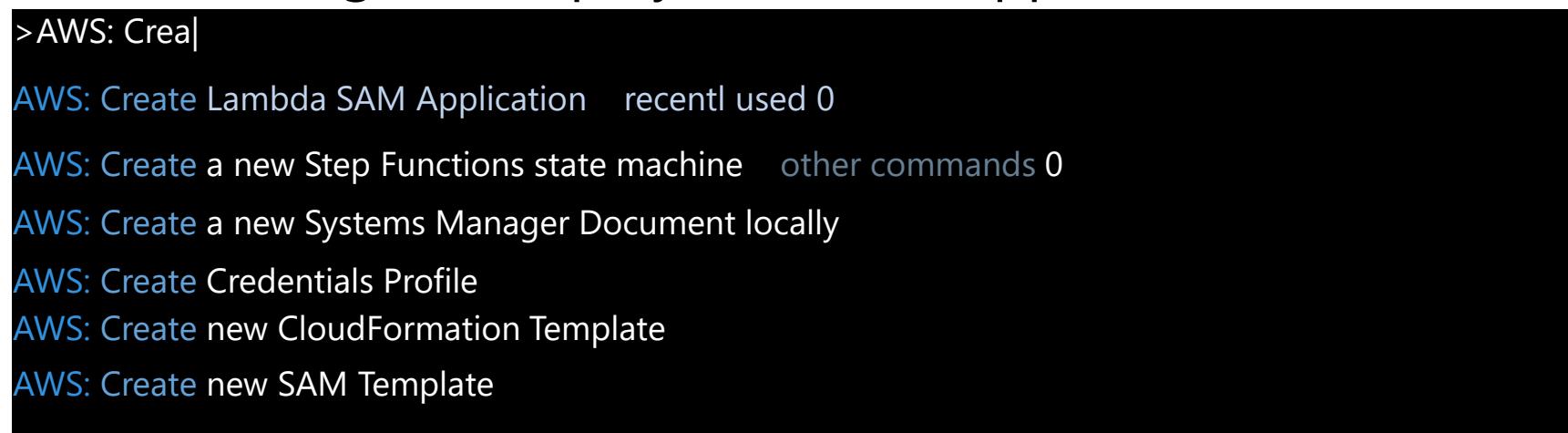
3. Amazon Elastic Container Service

Provides IntelliSense for ECS task-definitions files



4. Serverless Applications

Create, debug and deploy serverless applications via SAM and CFN



Access Keys

Access Keys is a **key and 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

Generate an Access Key and Secret

A screenshot of the AWS IAM 'Access keys' page. The page has a dark header with the title 'Access key'. Below the header, there is a warning message: 'If you lose or forget your secret access key, you cannot retrieve it. Instead, create a new access key and make the old key inactive.' There are two tabs at the top: 'Access key' and 'Secret access key'. Under the 'Access key' tab, there is a table with one row. The row contains a blue square icon, the access key ID 'AKIAWXMA6QKQ053BDQ5G', and a blue square icon with a copy symbol. To the right of the table, there is a 'Create New' button. The rest of the page is mostly blank.

Access Keys

Access keys (2)					
	Access key ID	Created on	Access key last used	Region last used	Service
<input type="radio"/>	AKIAWXMA6QKQKXC56HVR	67 days ago	3 days ago	ap-south-1	sts
<input type="radio"/>	AKIAWXMA6QKQ053BDQ5G	2 minutes ago	None	N/A	N/A

- 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.

Access Keys

Access Keys are to be stored in `~/.aws/credentials` and follow a **TOML** file format

Default will be the access key used when no profile is specified.

You can store multiple access keys by giving the **profile** names.

```
[default]
aws_access_key_id=AKIAI0SF0DNN7EXAMPLE
aws_secret_access_key=wJalrXUtnFEMI/K7MDENG/bPxRfICYEXAMPLEKEY
[exampro]
aws_access_key_id=AKIAI0SF0DNN7EXAMPLE
aws_secret_access_key=wJalrXUtnFEMI/K7MDENG/bPxRfICYEXAMPLEKEY
region=ca-central-1
```

You can use the **aws configure** CLI command to populate the credential file.

```
$ aws configure
AWS Access Key ID [None]: AKIAI0SF0DNN7EXAMPLE
AWS Secret Access Key [None]: wJalrXUtnFEMI/K7MDENG/bPxRfICYEXAMPLEKEY Default region name
[None]: us-west-2 Default output format [None]: json
```

The AWS SDK will automatically read from these environment variables.

This is the safe way of using an Access Key within your code.

```
$ export AWS_ACCESS_KEY_ID=AKIAI0SF0DNN7EXAMPLE
$ export AWS_SECRET_ACCESS_KEY=wJalrXUtnFEMI/K7MDENG/bPxRfICYEXAMPLEKEY $
export AWS_DEFAULT_REGION=us-west-2
```

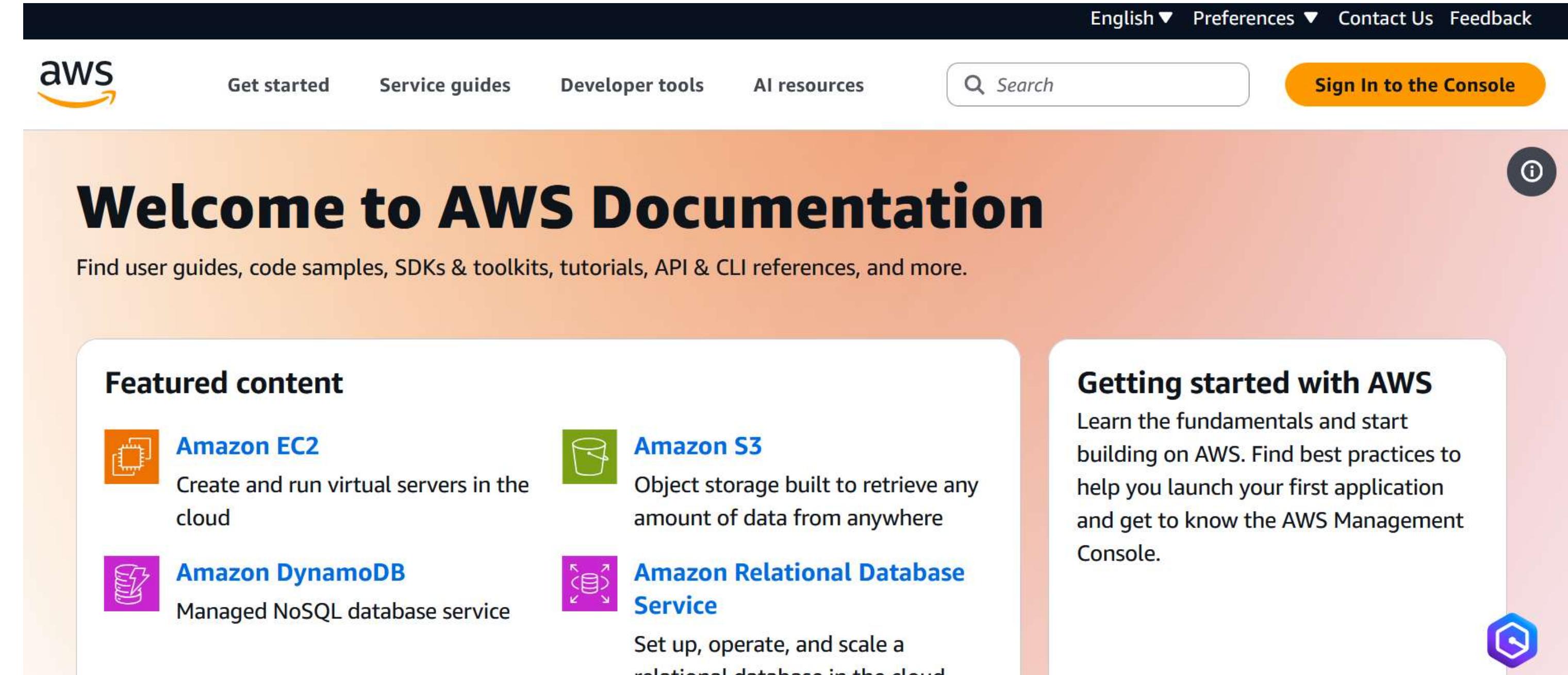
AWS Documentation

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

docs.aws.amazon.com

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

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



The screenshot shows the AWS Documentation homepage. At the top, there's a navigation bar with links for English, Preferences, Contact Us, and Feedback. Below the navigation is the AWS logo and a search bar. The main heading is "Welcome to AWS Documentation". Below it, a sub-headline says "Find user guides, code samples, SDKs & toolkits, tutorials, API & CLI references, and more." On the left, there's a "Featured content" section with four items: Amazon EC2 (orange icon), Amazon S3 (green icon), Amazon DynamoDB (purple icon), and Amazon Relational Database Service (pink icon). Each item has a brief description. To the right, there's a "Getting started with AWS" section with a description and a "Sign In to the Console" button. A small purple hexagonal icon is in the bottom right corner of the page.

English ▾ Preferences ▾ Contact Us Feedback

aws

Get started Service guides Developer tools AI resources

Search

Sign In to the Console

Welcome to AWS Documentation

Find user guides, code samples, SDKs & toolkits, tutorials, API & CLI references, and more.

Featured content

- Amazon EC2**
Create and run virtual servers in the cloud
- Amazon S3**
Object storage built to retrieve any amount of data from anywhere
- Amazon DynamoDB**
Managed NoSQL database service
- Amazon Relational Database Service**
Set up, operate, and scale a relational database in the cloud

Getting started with AWS

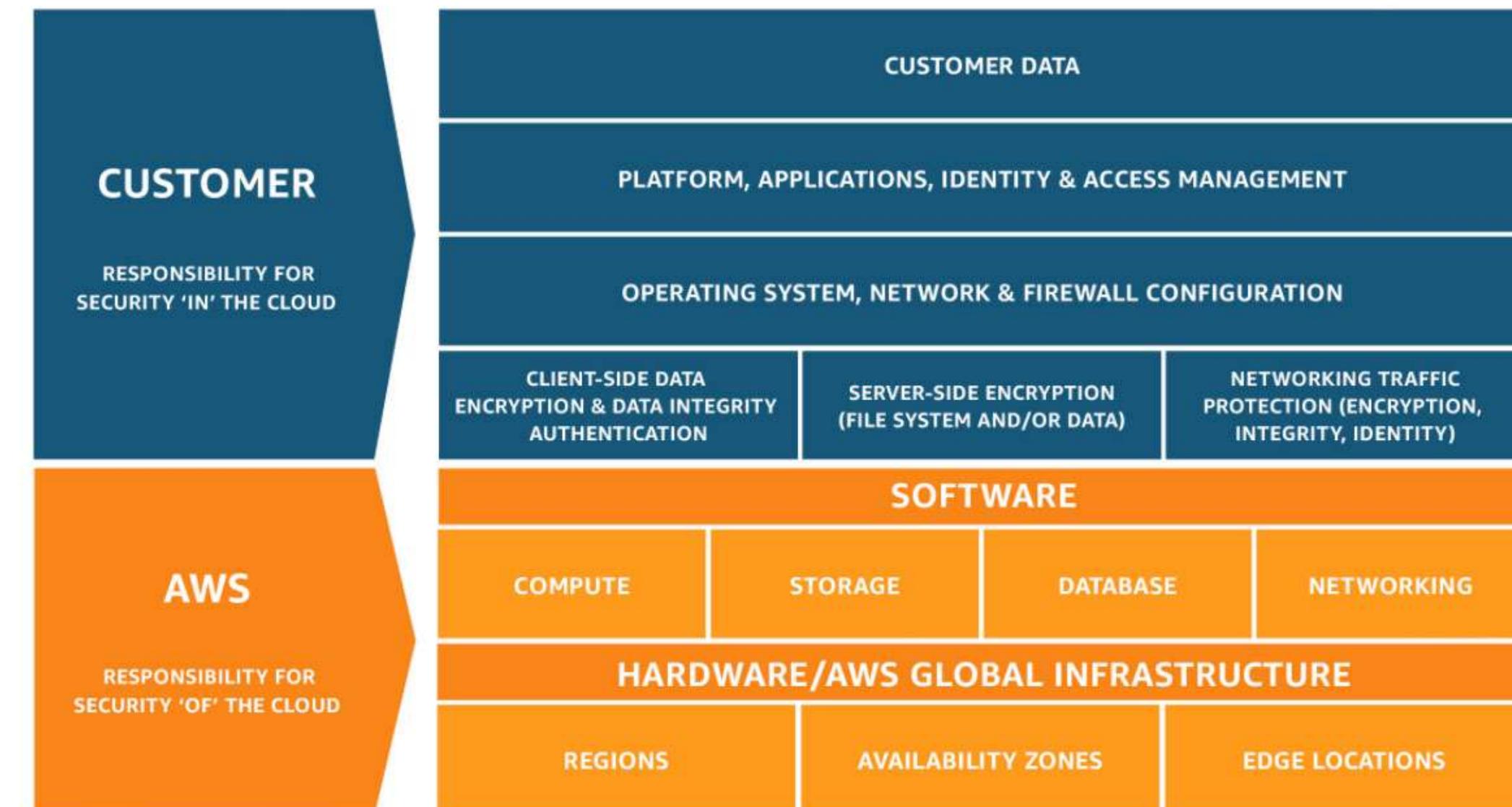
Learn the fundamentals and start building on AWS. Find best practices to help you launch your first application and get to know the AWS Management Console.

Shared Responsibility Model

The **Shared Responsibility Model** is a **cloud security framework** that defines the security obligations of the customer versa the Cloud Service Provider (CSP) e.g. AWS.

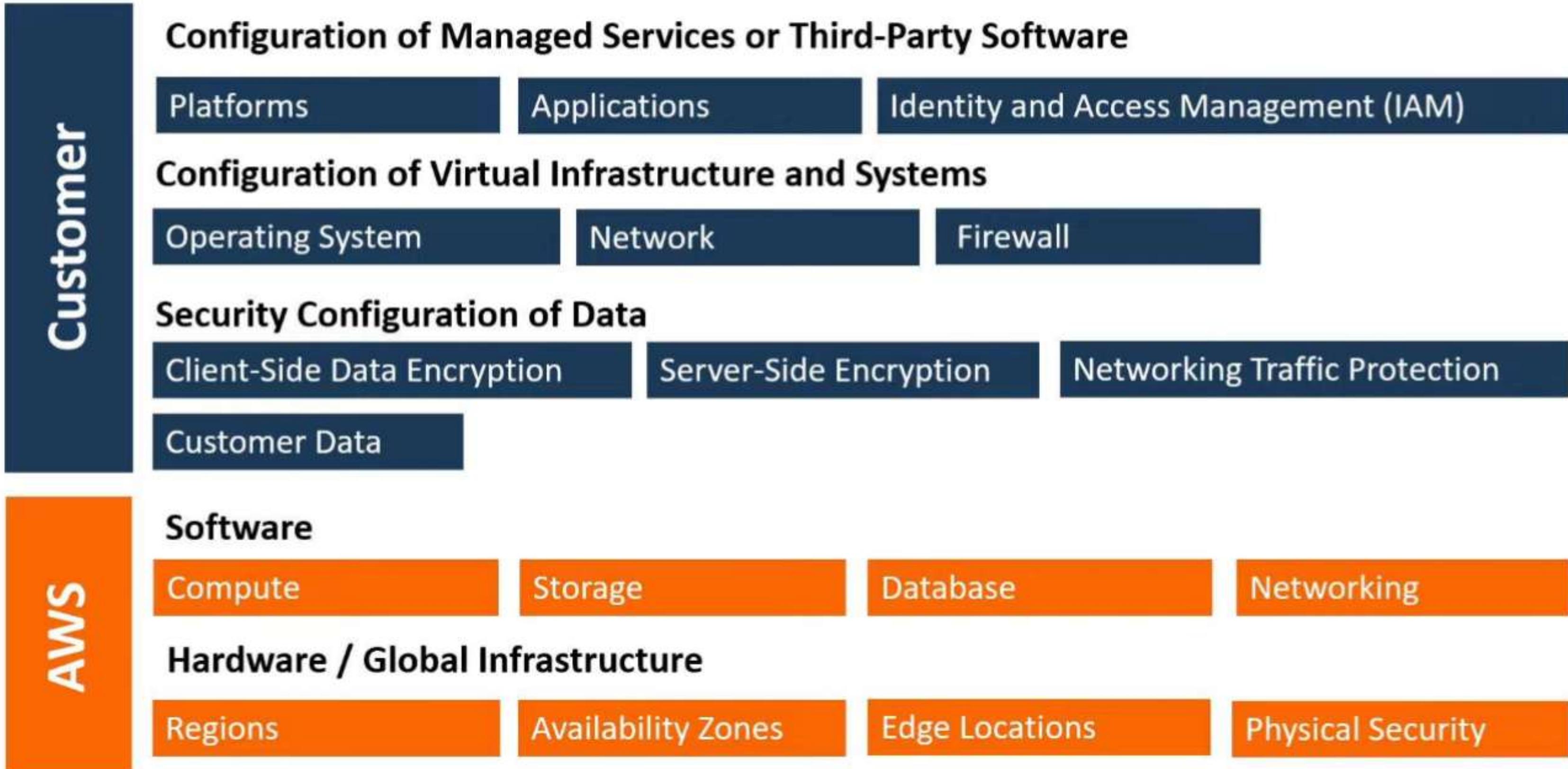
Each CSP has their own variant of the Shared Responsibility Model but they are all generally the same.

AWS Shared Responsibility Model



The **type of cloud deployment model** and/or **the scope of cloud service category** can result in specialized Shared Responsibility Models.

AWS Shared Responsibility Model



AWS Shared Responsibility Model



Customers are responsible for Security **in** the Cloud

Data
Configuration



Hardware
Operation of Managed Services
Global Infrastructure

AWS is responsible for Security **of** the Cloud

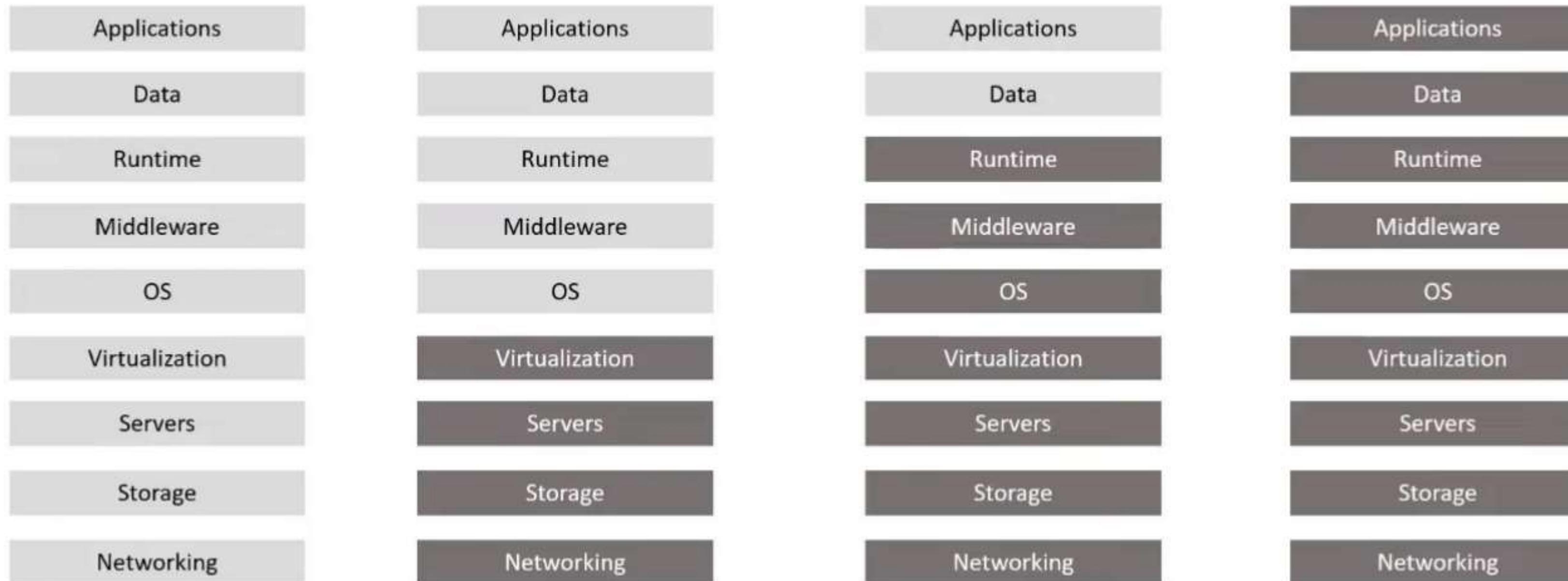
Types of Cloud Computing Responsibility

On-Premise

Infrastructure as a Service

Platform as a Service

Software as a Service



Legend: Customer is Responsible

CSP is Responsible

Shared Responsibility Model - Compute

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

Infrastructure as a Service (IAAS)



Bare Metal

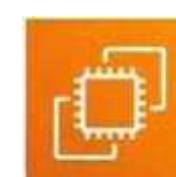
EC2 Bare Metal Instance

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

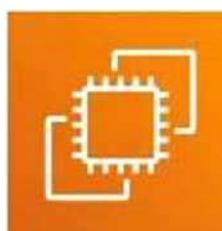
- Deployment, Container Runtime, Networking, Storage, Security, Physical Machine, (basically everything)

Shared Responsibility Model - Compute

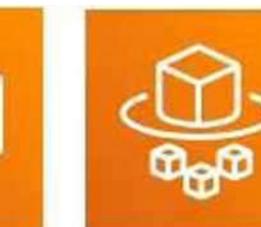


Bare Metal

EC2 Bare Metal Instance EC2 Dedicated Instance/Host Elastic Compute Cloud (EC2)



Dedicated Virtual Machines (VM)



Containers

Elastic Container Service(ECS)
Elastic Kubernetes Service (EKS)
AWS Fargate*



Functions

AWS Lambda



Customer
Responsibility

Level of Control

AWS
Responsibility

Shared Responsibility Model

The **Shared Responsibility Model** is a simple visualization that helps determine what the customer is responsible for and what the CSP is responsible for related to AWS.

The customer is responsible for the data and the **configuration** of access controls that resides in AWS.

The customer is responsible for the **configuration** of cloud services and granting access to users via permissions.

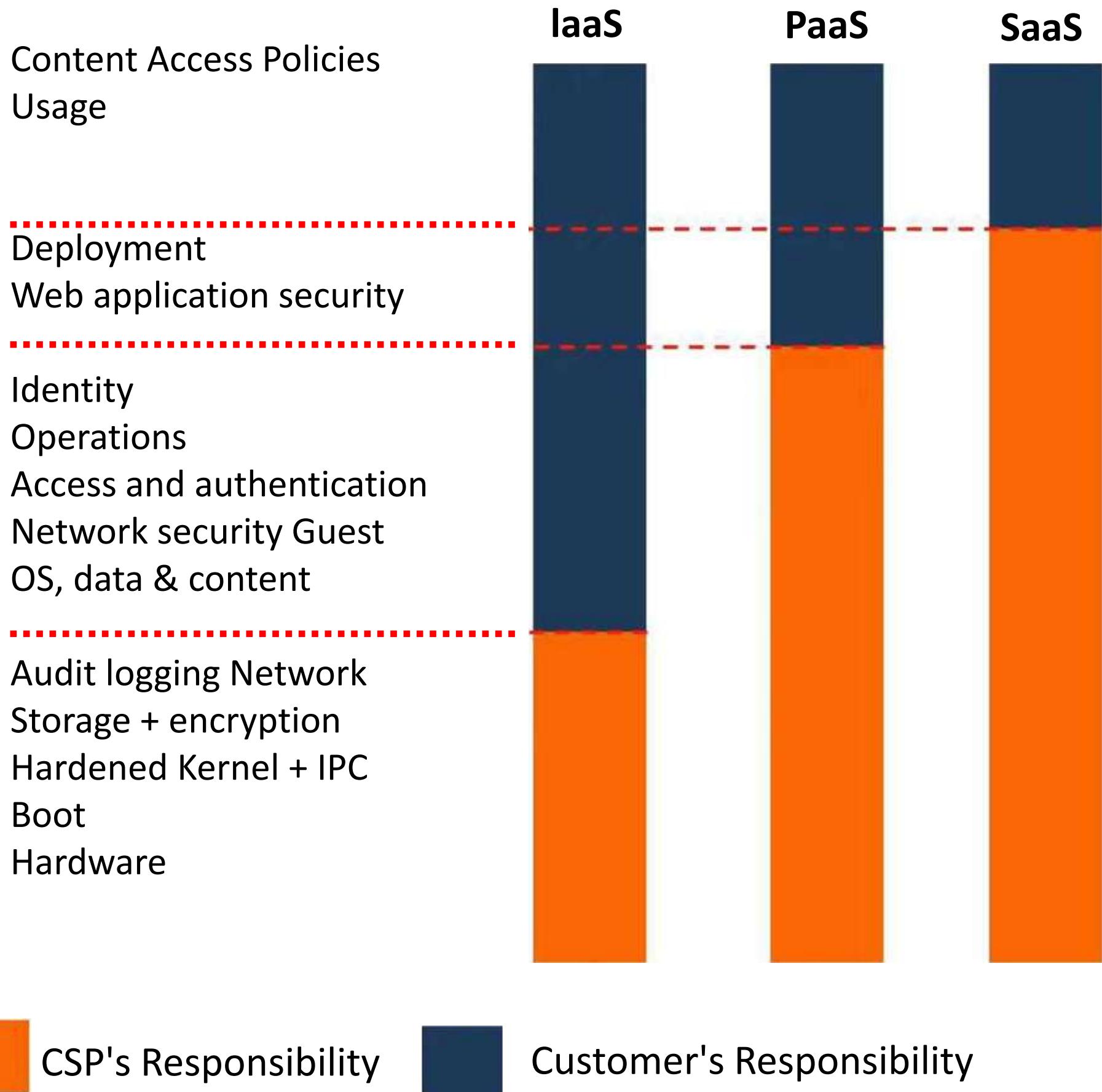
CSP is generally responsible for the underlying Infrastructure.

Responsibility of in the cloud

If you can configure or store it then you (the customer) are responsible for it.

Responsibility of the cloud

If you can not configure it then CSP is responsible for it



Shared Responsibility Model - Architecture

Less Responsibility



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.

More Responsibility

Computing Services



Elastic Compute Cloud (EC2) allows you to launch **Virtual Machines (VM)**

What is a Virtual Machine?

A Virtual Machine (VM) is an emulation of a physical computer using software.

Server Virtualization allows you to easily **create, copy, resize or migrate** your server.

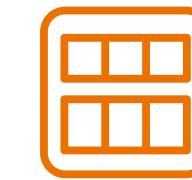
Multiple VMs can run **on the same physical server** so you can share the cost with other customers.

Imagine if your server or computer was an executable file on your computer

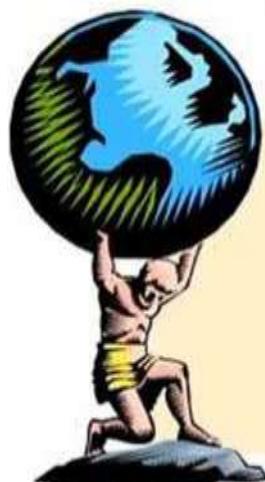
When we launch a Virtual Machine we call it an "**instance**"

EC2 is **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 eg. 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, eg. S3, RDS, DynamoDB, Lambdas

Computing Services

Virtual Machines — an emulation of a physical computer using software

 **Amazon LightSail** is the **managed virtual server service**. It is the "friendly" version of EC2 Virtual Machines
When you need to launch a Linux or Windows server but don't have much AWS knowledge, eg. Launch a Wordpress

Containers — virtualizing an Operation 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)

 **Elastic Container Service (ECS)** is a **container orchestration service** that support **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*



 **Elastic Container Registry (ECR)** is **repository for container images**. In order to launch a containers you need an image. An image just means a saved copy. A repository just means a storage that has version control.

ECS Fargate is **serverless orchestration container service**. It is the same as ECS expect you pay-on-demand per running container (With ECS you have to keep a 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.

 **Elastic Kubernetes Service (EKS)** is a **fully managed Kubernetes service**. Kubernetes (K8) is an open-source orchestration software that was created by Google and 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.

 **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 and how long function is allowed to run before timing out. You are charged based on the runtime of the serverless function rounded to the nearest 100ms.

Higher Performance Computing Services

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

- Nitro Cards — specialized cards for VPC, EBS and Instance Storage and controller card
- Nitro Security Chips — Integrated into motherboard. Protects hardware resources.
- Nitro Hypervisor — lightweight hypervisor Memory and CPU allocation Bare Metal-like performance

Bare Metal Instance You can launch EC2 instance that have no hypervisor so you can run workloads directly on the hardware for maximum performance and control. The **M5 and R5** EC2 instances run are bare metal.



Bottlerocket is a Linux-based open-source operation system that is purpose-built by AWS for running containers on Virtual Machines or bare metal hosts

What is High Performance Computing (HPC)?

A cluster of hundreds of thousands of 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 computers or would take too long.



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

Edge and Hybrid Computing Services

What is Edge Computing?

When you push your computing workloads outside of your networks to run close to the destination location, eg.
Pushing computing to run on phones, IoT Devices, or external servers not within your cloud network.

What is Hybrid Computing?

When you're able to run workloads on both your on-premise datacenter and AWS Virtual Private Cloud (VPC)



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



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

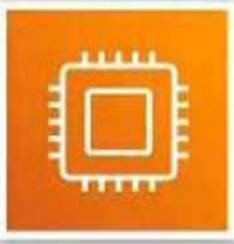


VMWare Cloud on AWS allows you to **manage on-premise virtual machines using VMWare** as EC2 instances.  The data-center must be using VMWare for Virtualization.



AWS Local Zones are **edge datacenters located outside of an AWS region** so you can use AWS closer to end destination.

When you need faster computing, storage and databases in populated areas that are outside of an AWS Region



Cost and Capacity Management Computing Services

Cost Management How do we save money?

Capacity Management How do we meet the demand of traffic and usages though adding or upgrading servers?



EC2 Spot Instances, Reserved Instances and Savings Plan

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



AWS Batch plans, schedules, and executes **your batch computing workloads** across the full range of AWS compute services, can utilize Spot Instance to save money.



AWS Compute Optimizer suggests how to **reduce costs and improve performance** by using machine learning to analyze your previous usage history



EC2 Autoscaling Groups (ASGs)

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



Elastic Load Balancer (ELB)

Distributes traffic to multiple instance, can re-route traffic from unhealthy instance to healthy instances, can route traffic to EC2 instances running in different Availability Zones



AWS Elastic Beanstalk (EB) is for easily deploying web-applications without developers having to worry about setting up and understanding the underlying AWS Services. Similar to **Heroku**.



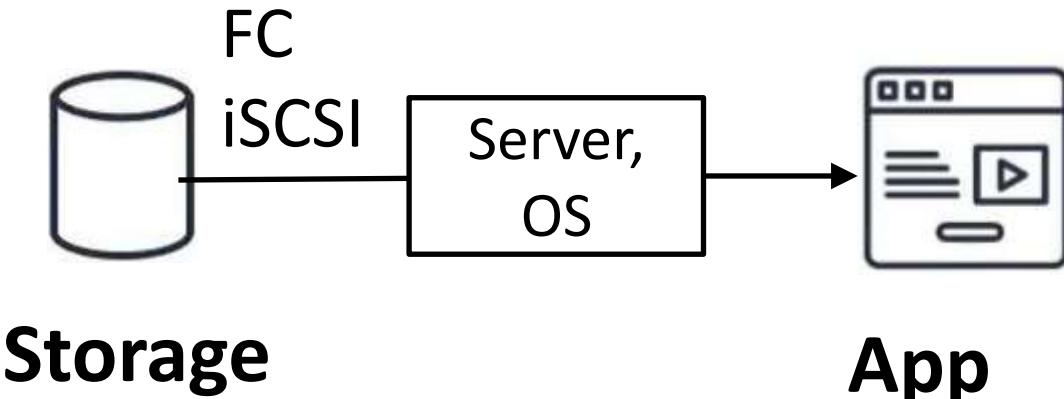
Types of Storage Services



Elastic Block Storage (EBS) - Block

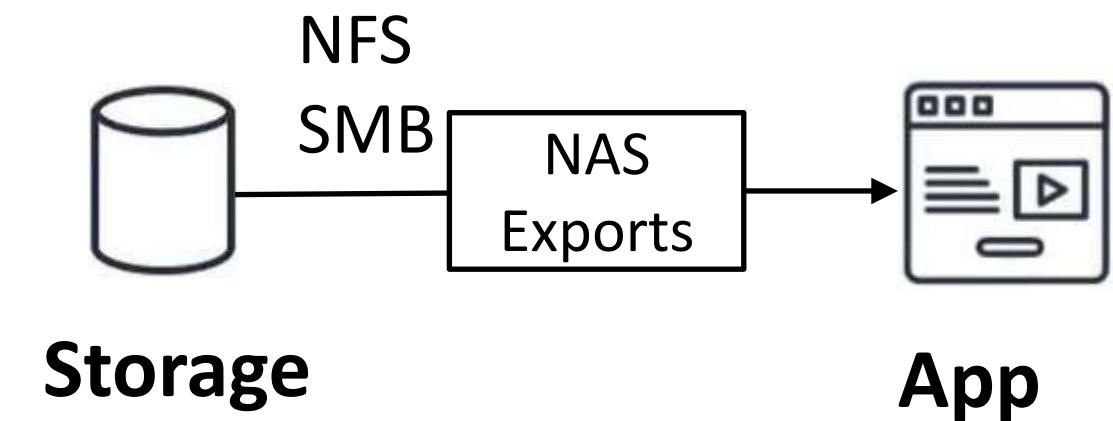
Data is split into evenly split blocks. Directly accessed by the Operating System. Supports only a single write volume.

When you need a virtual hard disk/drive attached to a VM.



AWS Elastic File Storage (EFS) - File

File is stored with data and metadata. Multiple connections via a network share. Supports multiple reads, writing locks the file.

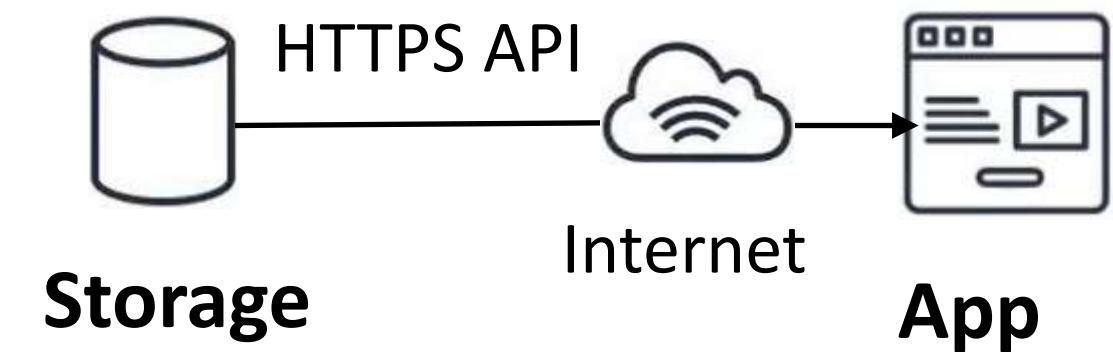


When you need a file-share where multiple users or VMs need to access the same drive



Amazon Simple Storage Service (S3) - Object

Object is stored with data, metadata and Unique ID. Scales with limited no file limit or storage limit. Supports multiple reads and writes (no locks)



When you just want to upload files, and not have to worry about underlying infrastructure. Not intended for high IOPs

Introduction to S3

What is Object Storage (Object-based Storage)?

data storage architecture that manages data as objects, **as opposed** to other storage architectures:

- **file systems** which manages data as a files and file hierarchy, and
- **block storage** which manages data as blocks within sectors and tracks.



S3 provides you with **unlimited storage**.

You don't need to think about the underlying infrastructure

The S3 Console provides an interface for you to upload and access your data



S3 Object

Objects contain your data. They are like files.

Object may consist of:

- **Key** this is 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
(think like having a domain name)

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 and Durability for Cheaper Storage*

S3 Standard (default)

Fast! 99.99% Availability, 11 9's Durability. Replicated across at least three AZs **S3**

Intelligent Tiering

Uses ML to analyze object usage and determine the appropriate storage class.

Data is moved to the most cost-effective access tier, without any performance impact or added overhead.

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)

S3 One-Zone-IA

Still Fast! Objects only exist in one AZ. Availability (is 99.5%). but cheaper than Standard IA by 20% less
(Reduce durability) Data could get destroyed. A retrieval fee is applied.

S3 Glacier

For long-term cold storage. Retrieval of data can take minutes to hours but the off is very cheap storage **S3**

Glacier Deep Archive

The lowest cost storage class. Data retrieval time is 12 hours.

S3 Outposts has its own storage class

Cheaper

AWS Snow Family

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



Snowcone

Comes in two sizes:

- 8 TB of Storage (HHD)
- 14 TB of Storage (SSD)



Snowball Edge

Comes generally in two type:

- Storage Optimized
 - 80 TB
- Compute Optimized
 - 39.5 TB



Snowmobile

100 PB of storage



Data is delivered to Amazon S3

Storage Services



Simple Storage Service (S3) is a **serverless object storage service**. You can upload very large files and 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.



S3 Glacier is a **cold storage service**. It design as a low cost storage solution for **archiving and long-term backup**.

It uses previous generation HDD drives to get that low cost. Its highly secure and durable.



Elastic Block Store (EBS) is a **persistent block storage service**. It is a virtual hard drive in the cloud you attach to EC2 instances. You can choose different kinds of hard drives: **SSD, IOPS SSD, Throughput HHD, Cold HHD**



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**



Storage Gateway is a **hybrid cloud storage** service that extends your on-premise storage to cloud



File Gateway extends your local storage to AWS S3



Volume Gateway caches your local drives to S3 so you have a continuous backup of local files in the cloud



Tape Gateway stores files onto virtual tapes for backing up your files on very cost effective long term storage.

Storage Services



AWS Snow Family are **storage devices used to physically migrate large amounts of data to the cloud.**

- Snew&aH-afld **Snowball Edge** are briefcase size data storage devices. **50-80 Terabytes**
- **Snowmobile** is a cargo container filled with racks of storage and compute that is transported via semi-trailer tractor truck to transfer up to **100PB** of data per trailer.
- **Snowcone** is a very small version of Snowball that can transfer **8TB** of data.



AWS Backup a fully **managed backup service** that makes it easy to centralize and automate the backup of data across multiple AWS services eg. EC2, EBS, RDS, DynamoDB, EFS, Storage Gateway. You create backup plans.



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



Amazon FSx is a **feature rich and highly-performant file system**. That can be used for Windows (SMB) or Linux (Lustre)



Amazon FSx for Window File Server uses the SMB protocol and allows you to mount FSx to Windows servers



Amazon FSx for Lustre uses Linux's Lustre file system and allows you to mount FSx to Linux servers

What is a Database?

A database is a **data-store that stores semi-structured and structured data**.

A database is more **complex data stores** because it **requires using formal design and modeling techniques**

Databases can be generally categorized as either:

- **Relational databases**
 - Structured data that strongly represents tabular data (tables, rows and 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:

- specialized language to query (retrieve data)
- specialized modeling strategies to optimize retrieval for different use cases
- more fine tune control over the transformation of the data into useful data structures or reports



Normally a databases infers someone is using a **relational row-oriented data store**

What is Data Warehouse?

A relational datastore designed for **analytic workloads**, which is generally **column-oriented data-store**

Companies will have **terabytes and millions of rows of data**,
and they need a fast way to be able to produce analytics reports

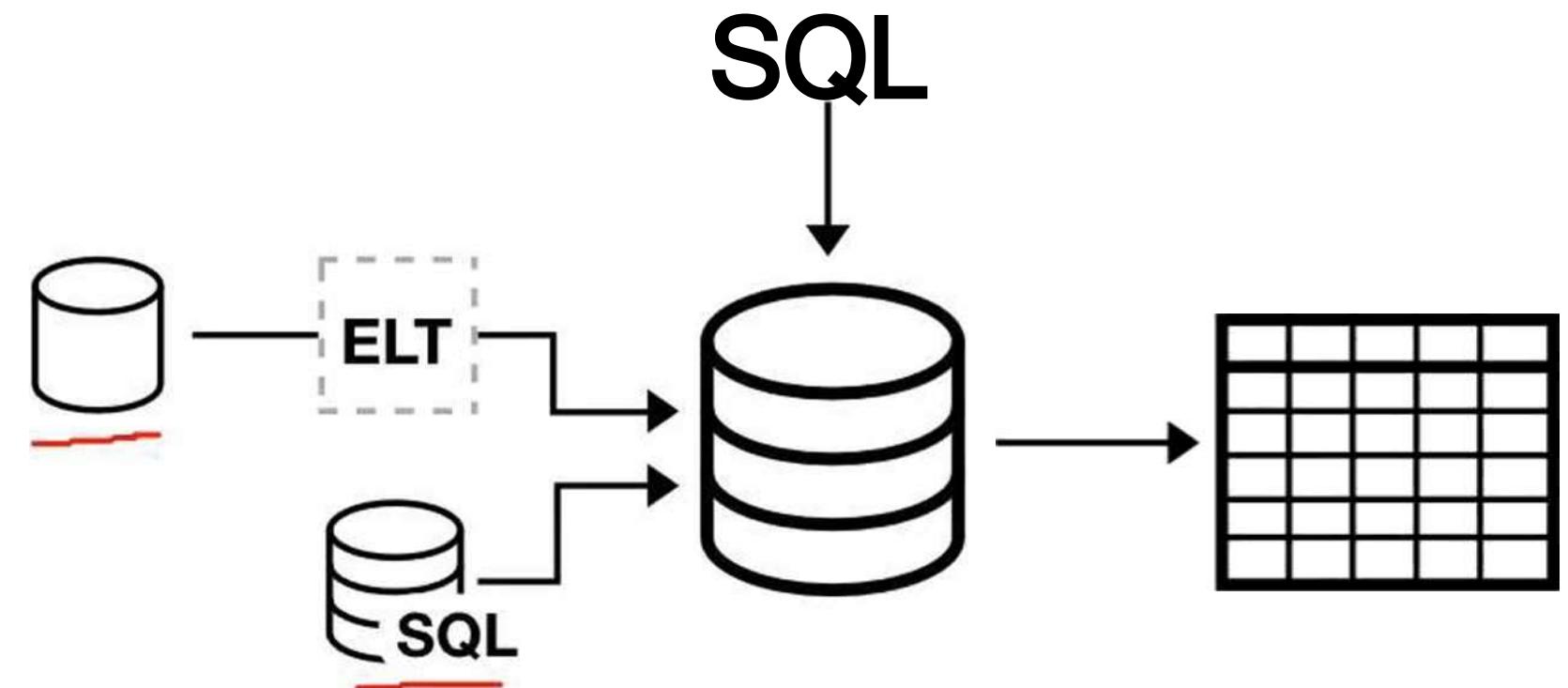
Data warehouses generally perform **aggregation**

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

Data warehouses are generally designed be HOT • Hot means they can return queries very fast even though they 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 and user reports.

A data warehouse needs to consume data from a 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.

A key/value stores a **unique key** alongside a value



Key	Value
Data	1010101000101011001010010101001
Worf	0110101100010101010101011100010
Ro Laren	0010101001010110010101010101010

Key values stores are **dumb and fast**.

They generally lack features like:

- Relationships
- Indexes
- Aggregation

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

A simple key/value store will interpret this data resembling a dictionary (aka Associative arrays or hash)

A key/value store can resemble tabular data, it does not have to have the consistent columns per row (hence its schema less)

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

Due to their simple design they can scale well beyond a relational database

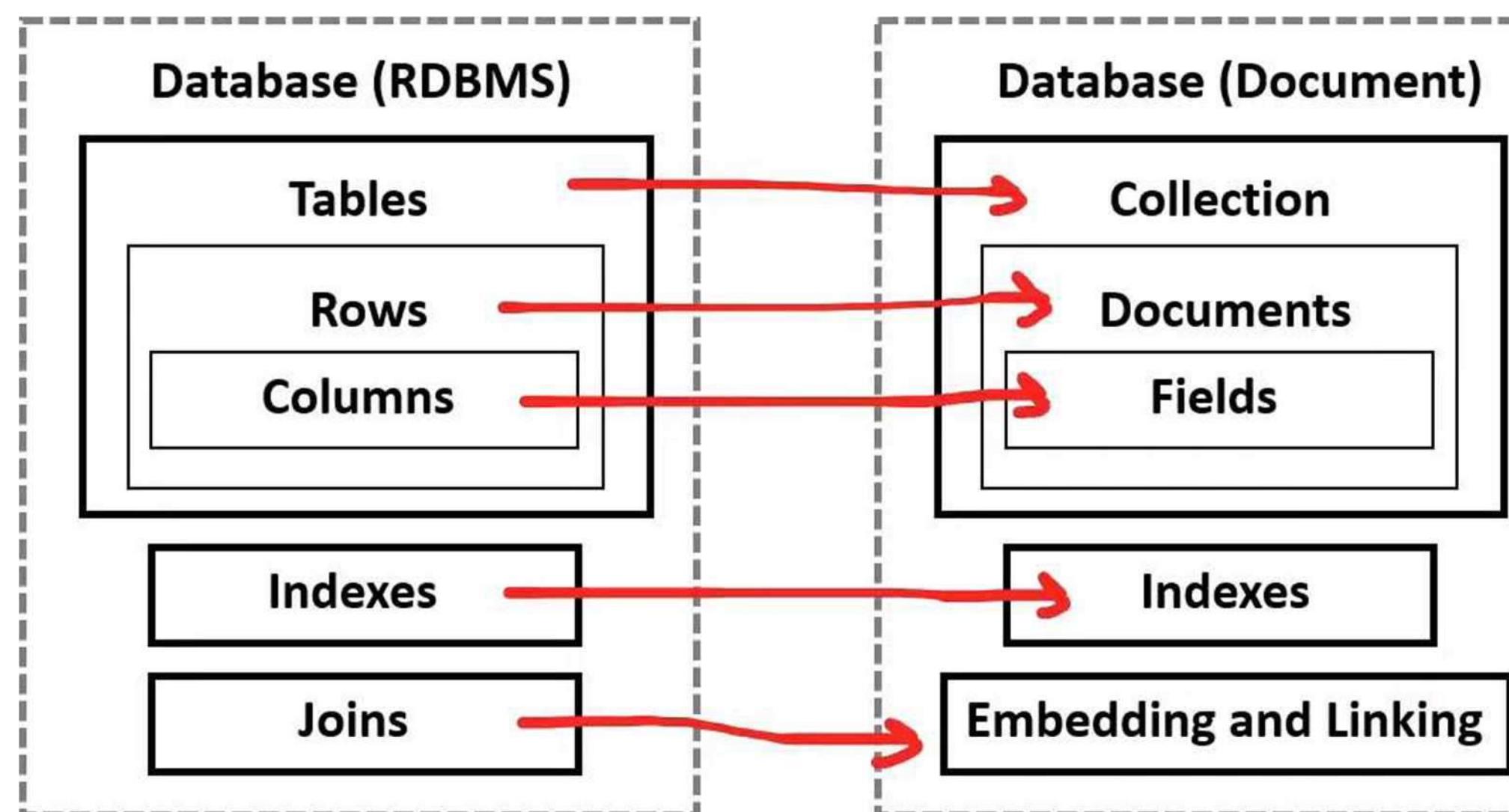
What is a Document store?

A **document store** is a NOSQL database that stores **documents** as its primary data structure.

A document could be an XML but more commonly is JSON or JSON-Like

Document stores are sub-class of Key/Value stores

The components of a document store compared to Relational database



NoSQL Database Service



DynamoDB is a serverless **NoSQL key/value and 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 and very fast we should think DynamoDB



In **2019, Amazon** the online shopping retail shutdown their last Oracle database and completed their migration to DynamoDB. They had 7,500 Oracle Database and 75 petabytes of data. With DynamoDB they reduce costs by 60% and reduce latency by 40%

When we want a massively scalable database



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

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



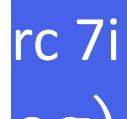
When you want a MongoDB database.



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



Relational Database Services



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



MySQL



MariaDB



MySQL - The most popular open-source SQL database that was purchased and now owned by Oracle.

MariaDB - When Oracle bought MySQL. MariaDB made a fork (copy) of MySQL was made under a different open-source license.

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 license to use it.

Microsoft SQL Server - Microsoft's proprietary SQL database. You have to buy a license to use it.

Aurora - Fully managed database.



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

When you want a highly available, durable, scalable and 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 demand*



RDS on VMware allows you to deploy RDS supported engines to on an-premise data-center. The datacenter must be using VMware for server virtualization. *When you want databases managed by RDS on your own datacenter*



vmware
vSphere

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 are keeping data "hot". Meaning that we can run a very complex query and a large amount of data and get that data back very fast.

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



ElastiCache is a managed database of the **in-memory** and **caching** open-source databases **Redis** or **Memcached**. ***When you need to improve the performance of application by adding a caching layer in-front of 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 and cryptographically variable transaction logs.

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



Database Migration Service (DMS) is database migration service. You can migrate from:

- on-premise database to AWS
- from two database in different or same AWS accounts using different SQL engines
- from an SQL to NoSQL database



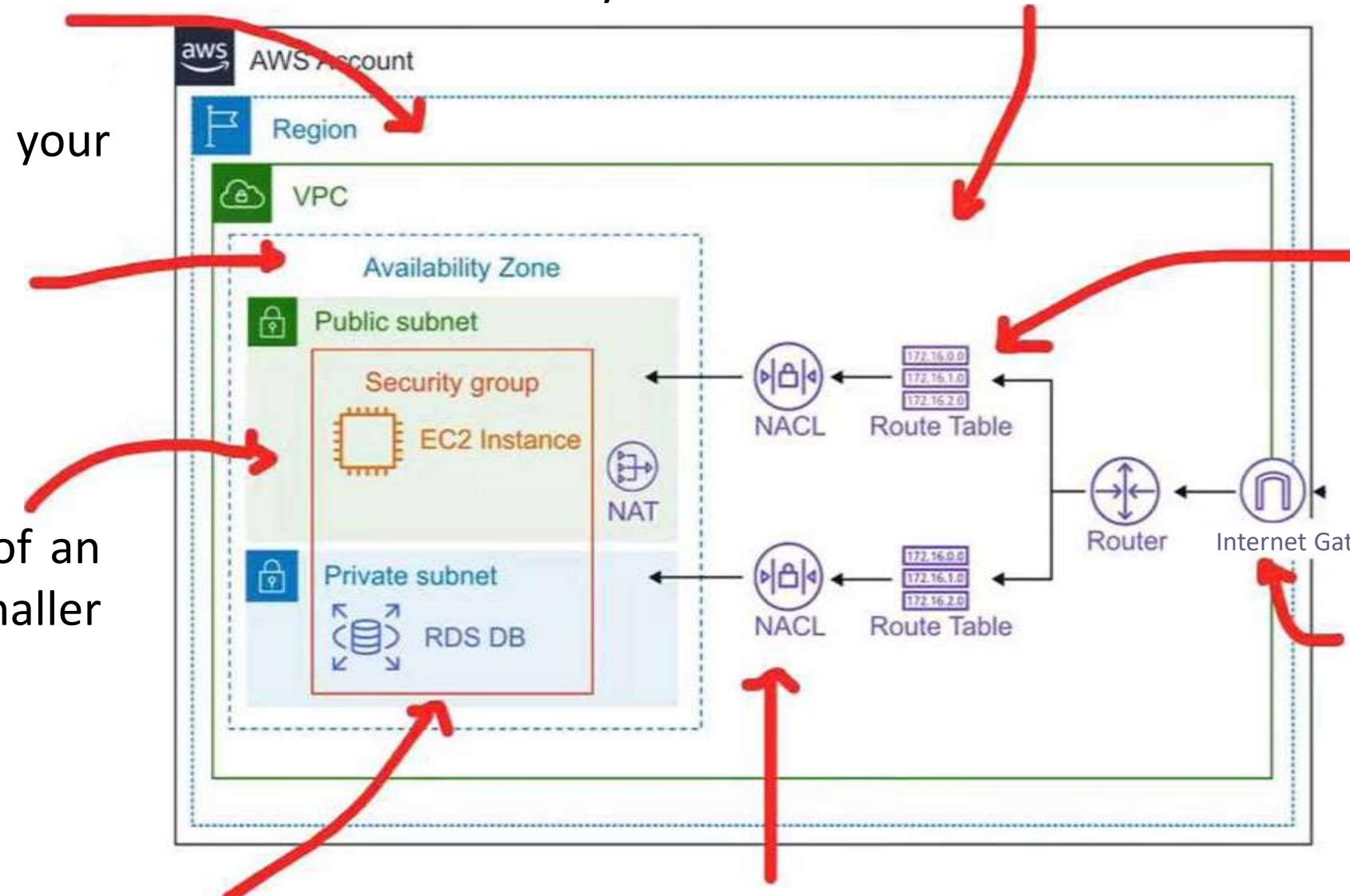
Cloud-Native Networking Services

Region the geographical location of your network

AZ the data center of your AWS resources

Subnets a logical partition of an IP network into multiple, smaller network segments

VPC a logically isolated section of the AWS Cloud where you can launch AWS resources



Security Groups Acts as firewall at the instance level

NACLs Acts as a firewalls at the subnet level

Route Tables determine where network traffic from your subnets are directed

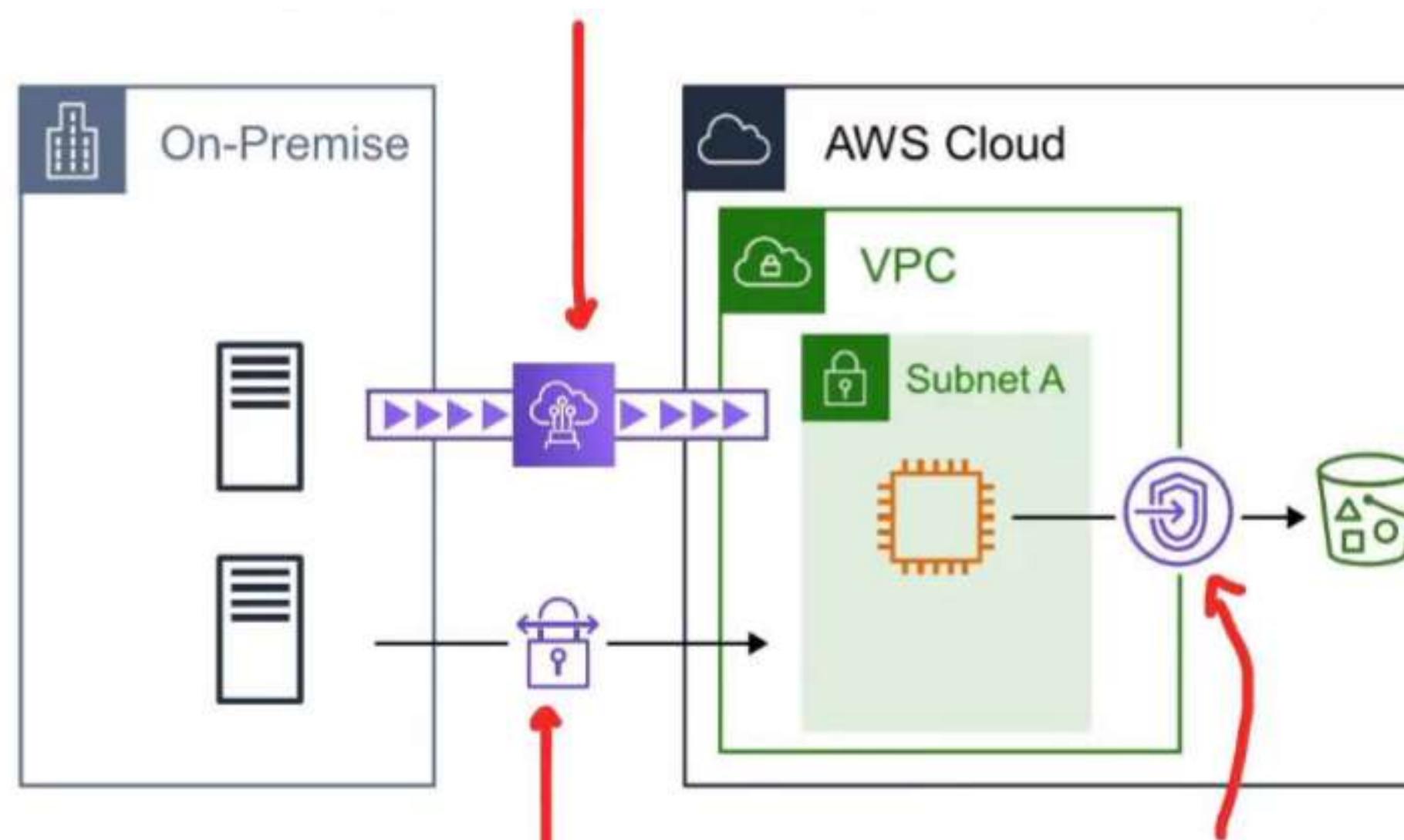
Internet Gateway (IGW) Enable access to the Internet



Enterprise/Hybrid Networking



DirectConnect dedicated gigabit connection from onpremise data-center to AWS (a very fast connection)



AWS Virtual Private Network (VPN) a secure connection between on-premise, remote offices, mobile employees.



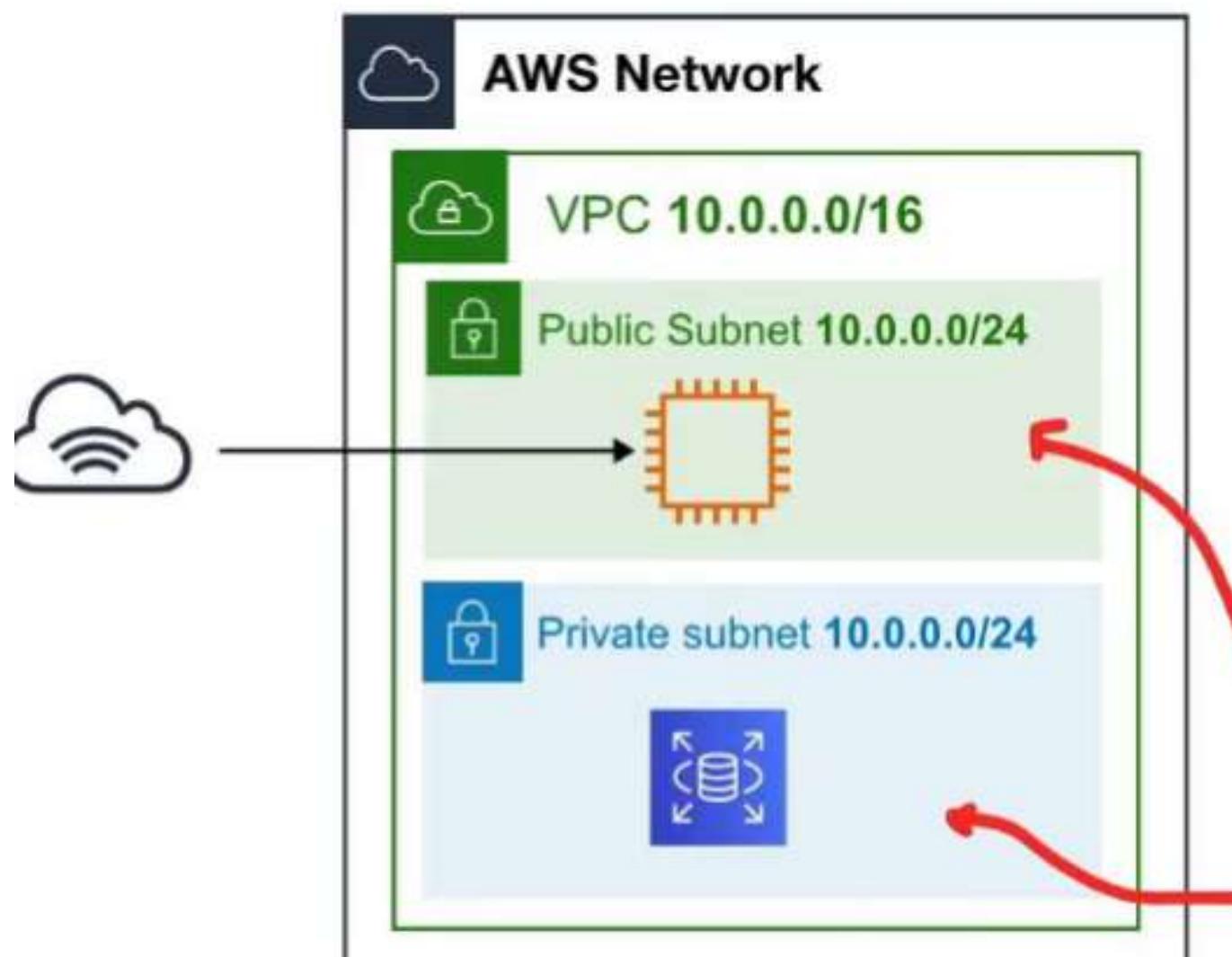
Private Links (VPC Interface Endpoints) keeps traffic within the AWS network and not traverse the internet to keep traffic is secure.



Virtual Private Cloud (VPC) and 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 Range of $10.0.0.0/16 = 65,536$ IP Addresses



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

Subnets **need to have a smaller CIDR range than to the VPC** represent their portion.
eg Subnet CIDR Range $10.0.0.0/24 = 256$ IP Addresses

A Public Subnet is one that can reach the internet

A Private Subnet is one that cannot reach the internet

Security Groups vs NACLs

Network Access Control Lists (NACLs)

Acts as a virtual **firewall** at the **subnet level**

You create **Allow and Deny rules**.

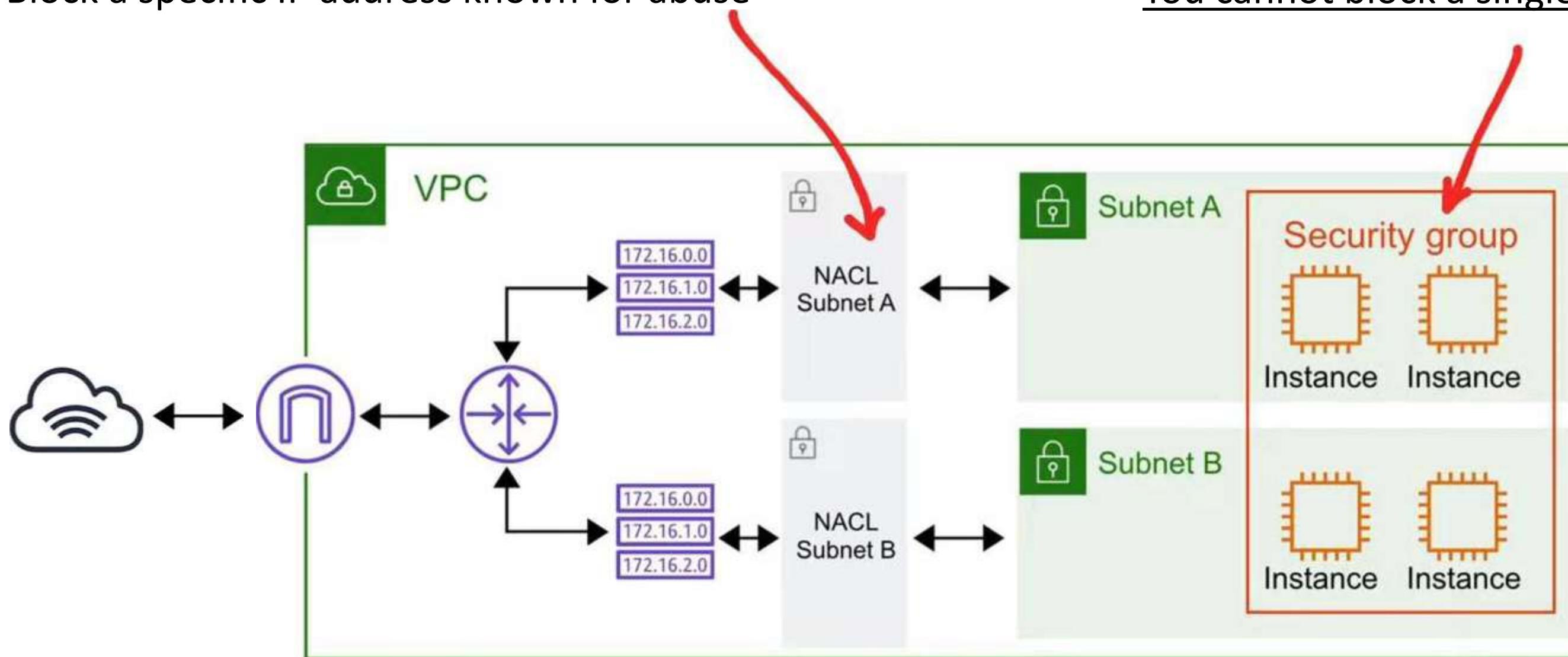
eg. Block a specific IP address known for abuse

Security Groups

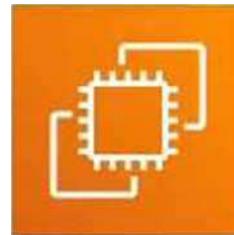
Acts as a virtual **firewall** at the **instance level**

Implicitly denies all traffic. You create only **Allow rules**,
eg. Allow an EC2 instance access on port 22 for SSH eg.

You cannot block a single IP address.



Introduction to EC2



Elastic Compute Cloud (EC2) is a **highly configurable virtual server**.

EC2 is resizable **compute capacity**. It takes **minutes** to launch new instances.

Anything and everything on AWS uses EC2 Instance underneath.

Choose OS via
Amazon Machine Image (AMI)



Choose **Instance Type**

t2.nano
\$0.0065/hour (\$4.75/month)
1 vCPU 0.5GB Mem

C4.8xlarge
\$1,591/hour (\$1161.43/month)
36 vCPU 60GB Mem 10 Gigabit performance

Add Storage (**EBS, EFS**)

SSD

HDD

Virtual Magnetic

Tape Multiple Volumes

Configure Instance

Security Groups, Key Pairs, User Data, IAM Roles, Placement Groups

EC2 Instance Families

What are Instance Families?

Instance families are different combinations of CPU, Memory, Storage and 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.

Commonly instance families are called "Instance Types" but an instance type is a combination of size and family

General Purpose

A1 T2_ T3 T3a T4g M4 M5 M5a M5n M6zn M6g M6i Mac

balance of compute, memory and networking resources **Use-cases** web servers and code repositories

Compute Optimized

C5 C4 Cba C5n C6g C6gn

Ideal for compute bound applications that benefit from high performance processor **Use-cases** scientific modeling, dedicated gaming servers and ad server engines

Memory Optimized

R4 R5 R5a R5b R5n XI Xle High Memory zld

fast performance for workloads that process large data sets in memory.

Use-cases in-memory caches, in-memory databases, real time big data analytics

Accelerated Optimized

P2 P3 P4 G3 G4ad G4dn F1 Infl VT1

hardware accelerators, or co-processors

Use-cases Machine learning, computational finance, seismic analysis, speech recognition

Storage Optimized

I3 I3en D2 D3 D3en HI

high, sequential read and write access to very large data sets on local storage **Use-cases** NoSQL, in-memory or transactional databases, data warehousing

EC2 Instance Types

An instance type is a particular **instance size and instance family**:

A common pattern for instance sizes:

- nano
- micro
- small
- medium
- large
- xlarge
- 2xlarge
- 4xlarge
- 8xlarge



Currently selected: t2.micro (- ECUs, 1 vCPUs, 2.5 GHz, -, 1 GiB memory, EBS only)				
	Family	Type	vCPUs	Memory (GiB)
	t2	t2.nano	1	0.5
<input checked="" type="checkbox"/>	t2	t2.micro <small>Free tier eligible</small>	1	1
	t2	t2.small	1	2
	t2	t2.medium	2	4
	t2	t2.large	2	8
	t2	t2.xlarge	4	16

There are many exceptions to this pattern for sizes e.g.

- c6g.metal - is a bare metal machine.
- C5.9xlarge - Is not a power of 2 or even number size

EC2 Instance Sizes

EC2 Instance Sizes **generally double** in price and key attributes

Name	vCPU	RAM (GiB)	On-Demand per hour	On-Demand per month
t2.small t2.	1	12	\$0,023	\$16.79
medium	2	24	\$0.0464	\$33.87
t2. large	2	36	\$0.0928	\$67.74
t2.xlarge	4	54	\$0.1856	\$135.48

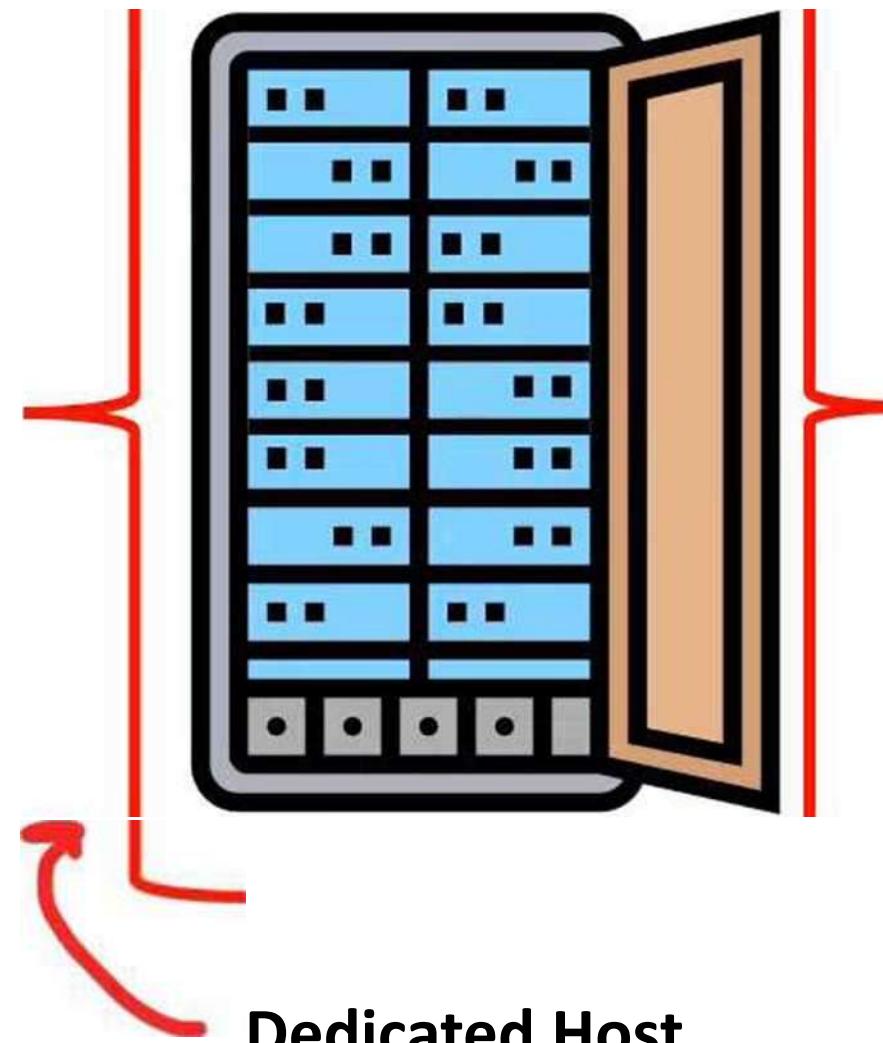
EC2 - Dedicated Host

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

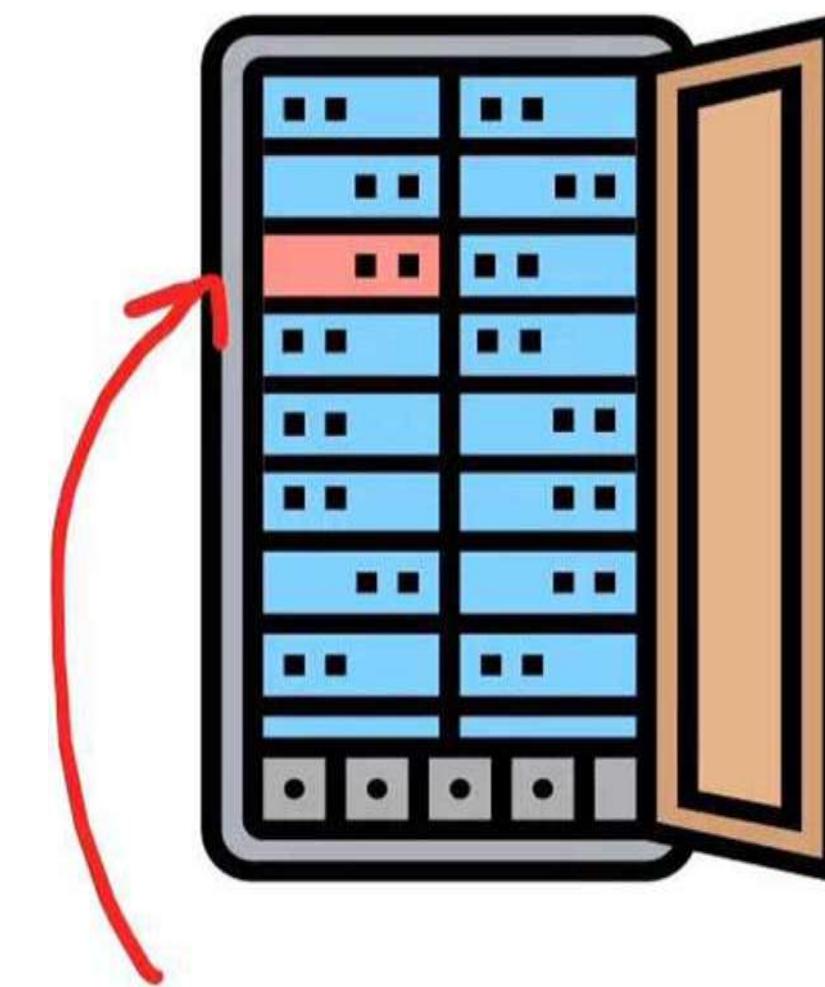
	Dedicated Instance	Dedicated Hosts
Isolation	Instance Isolation	Physical Server Isolation
Billing	Per instance billing (+\$2 per region fee)	Per host billing
Visibility of Physical characteristics	No Visibilities	Sockets, cores, host ID
Affinity between a host and instance	No Affinity	Consistency deploy to the same instances to the same physical server
Targeted instance placement	No control	Additional control over instance placement on physical server
Automatic instance placement	Yes	Yes
Add capacity using an allocation request	No	Yes

EC2 Tenancy

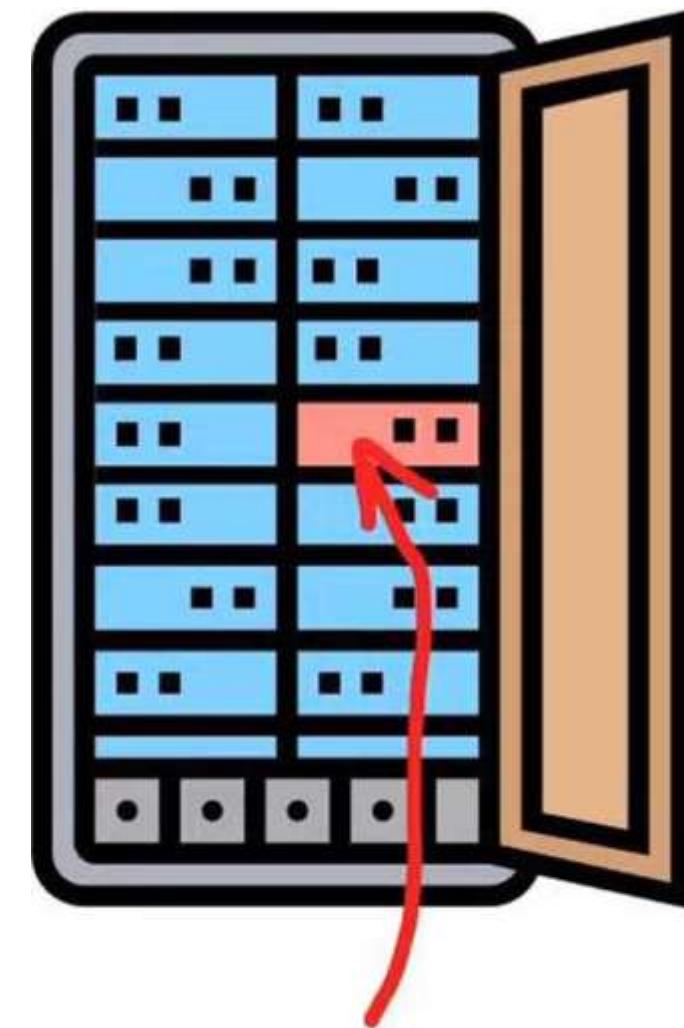
EC2 has three levels of tenancy:



Dedicated Host
Your server lives here and you have control of the physical attributes



Dedicated Instance
Your server always lives here



Default
Your instance live here *until reboot*

EC2 Pricing Models

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

On-Demand

Least Commitment

- low cost and flexible
- only pay per hour or the *second
- short-term, spiky, unpredictable workloads
- cannot be interrupted
- For first time apps

Reserved up to 75% off

Best Long-term

- steady state or predictable usage
- commit to EC2 over a 1 or 3 year term
- Can resell unused reserved instances

Spot up to 90%

Biggest Savings

- request spare computing capacity
- flexible start and end times
- Can handle interruptions (server randomly stopping and starting)
- For non-critical background jobs

Dedicated

Most Expensive

Dedicated servers
Can be on-demand or reserved or spot When you need a guarantee of isolate hardware (enterprise requirements)

AWS Savings Plan is another way to save but can be used for more than just EC2.

On-Demand

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

When you **launch** an EC2 instance it is
by default using On-Demand Pricing



On-demand has **no up-front payment** and **no long-term commitment**

You are charged by the **second (minimum of 60 seconds)** or the **hour**

per-second for:

Linux, Windows, Windows with SQL Enterprise, Windows with SQL Standard, and Windows with SQL Web Instances that do not have a separate hourly charge

per-hour:

full hour for all other instance types.

When looking up pricing it will always show
EC2 pricing is the **hourly rate**

Viewing 363 of 363 available instances						
Instance name	On-Demand hourly rate	vCPU	Memory	Storage	Network performance	
t2.nano	\$0.0058	1	0.5 GiB	EBS Only	Low	
t2.micro	\$0.0116	1	1 GiB	EBS Only	Low to Moderate	

On-Demand is for applications where the workload is for **short-term, spiky or unpredictable**.

When you have a **new app** for development or you want to run 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 savings.

You commit to a **1 Year** or **3 Year** contract.

Reserved Instances do not renew automatically

When they expire your instance will use On-Demand with no interruption to service

Class — The less flexible the greater the savings

Standard Up to **75%** reduced pricing compared to on-demand. You can modify **RI Attributes**.

Convertible Up to **54%** reduced pricing compared to on demand. You can exchange **RI** based on **RI Attributes** if greater or equal in value.

Scheduled AWS no longer offers Scheduled RI

Payment Options — The greater upfront the greater the savings

All Upfront

Full payment is made at the start of the term **Partial Upfront**
A portion of the cost must be paid upfront and the remaining hours in the term are billed at a discounted hourly rate

No Upfront

You are offered a discounted hourly rate for every hour within the term, regardless of whether the Reserved Instance is being used

RIs can be **shared between multiple accounts within an AWS Organization**
Unused RIs can be sold in the **Reserved Instance Marketplace**

Reserved Instances (RI)

Filter base on your requirements

Purchase Reserved Instances X

Only show offerings that reserve capacity

Platform	Tenancy	Offering class
Linux/UNIX	Default	Convertible

Instance type	Term	Payment option
t3.micro	1 month to 12 months	Any

Seller	Term	Effective rate	Upfront price	Hourly rate	Payment option	Offering class	Quantity available	Desired quantity	Normalized units per hour	Add to cart
AWS	12 months	\$0.008	\$69.00	\$0.000	All upfront	Convertible	Unlimited	1	0.5	<input type="button" value="Add to cart"/>
AWS	12 months	\$0.008	\$0.00	\$0.008	No upfront	Convertible	Unlimited	1	0.5	<input type="button" value="Add to cart"/>
AWS	12 months	\$0.008	\$35.00	\$0.004	Partial upfront	Convertible	Unlimited	1	0.5	<input type="button" value="Add to cart"/>

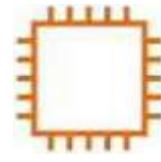
Your cart: 1 Reserved Instance, total due now: **\$69.00**

Monthly recurring cost: **\$0.00**

Additional taxes may apply.

Reserved Instances (RI) — RI Attributes

RI Attributes (aka Instance Attributes) are limited based on Class Offering and can affect the final price of an RI instance. There are 4 RI Attributes:



1. Instance type: For example, m4.large. This is composed of the instance family (for example, m4) and the instance size (for example, large).



2. Region: The Region in which the Reserved Instance is purchased.



3. Tenancy: Whether your instance runs on shared (default) or single-tenant (dedicated) hardware.



4. Platform: The operating system eg. Windows or Linux/Unix.

Regional and Zonal RI

When you purchase a RI, you determine **the scope** of the Reserved Instance.

The scope **does not affect the price**.

Regional RI: purchase for a Region

does *not* reserve capacity.

RI discount applies to instance usage in any AZ in the Region.

RI discount applies to instance usage within the instance family regardless of size. Only supported on Amazon Linux/Unix Reserved Instances with default tenancy.

You can queue purchases for regional RI

Zonal RI: purchase for an Availability Zone

reserves capacity in the specified Availability Zone.

RI discount applies to instance in the selected AZ (No AZ Flexibility)

No instance size flexibility
RI discounts apply to instance usage for the specified instance type and size only.

You can't queue purchases for zonal RI

RI Limits

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

Per month you can purchase

- **20 Regional Reserved Instances *per Region***
- **20 Zonal Reserved Instances *per AZ***

Regional Limits

You cannot exceed your running On-Demand Instance limit by purchasing regional Reserved Instances.

The default On-Demand Instance limit is 20.

Before purchasing RI ensure your On-Demand limit is equal to or greater than your RI you intend to purchase

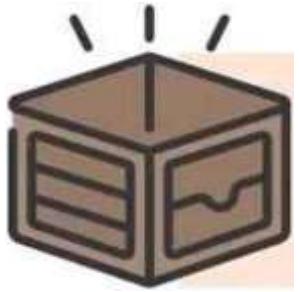
Zonal Limits

You can exceed your running On-Demand Instance limit by purchasing zonal Reserved Instances

If you already have 20 running On-Demand Instances, and you purchase 20 zonal Reserved Instances, you can launch a further 20 On-Demand Instances that match the specifications of your zonal Reserved Instances

Capacity Reservations

EC2 instances are backed by different kind of hardware, and so there is a **finite amount of servers** available within an Availability Zone per instance type or family.



You go to launch a specific type of EC2 instance but AWS has ran out of that server!

Capacity Reservation is a service of EC2 that allows you to **request a reserve of EC2 instance type** for a specific Region and 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

Instance Type
c4.2xlarge

EBS-optimized
If the selected instance type is EBS-optimized by default, you will not be able to uncheck this option.

Instance store
Temporary block-level storage. Data persists only during the life of the instance.

Platform
Linux/UNIX

Availability Zone
ca-central-1a

Tenancy
Default - run a shared hardware instance

Quantity
1

Reservation ends
Ending your reservation releases held capacity and thus prevents additional instances from being launched against it. Any launched instances continue to run and accrue applicable charges. You can view and manage those instances, if any, from the Launch Request page.

Manually
I will cancel my reservation when I am finished.

Specific time
Prevent launching instances against this reservation.

2021/10/10 10:06

Instance eligibility
Indicate the criteria for instances that can fulfill this reservation.

Any instance with matching details
Instance type, platform, and Availability Zone must match what is specified in the reservation.

Only instances that specify this reservation
When you launch instances, you must specify the reservation ID or the ARN associated with this reservation.

Standard vs Convertible RI

There are some key difference between Standard and Convertible

Standard RI

RI attributes can be modified

- Change the AZ within same Region
- Change the scope of the Zonal RI to Regional RI or visa versa
- Change the instance size (Linux/Unix only, default tenancy)
- Change network from Ec2-Classic to VPC and visa-versa

Can't be exchanged

Can be bought or sold in the RI Marketplace

Convertible RI

RI attributes can't be modified (you perform an exchange)

Can be exchanged during the term for another Convertible RI with new RI attributes, including:

- instance family
- instance type
- platform
- scope
- tenancy

Can't be bought or sold in the RI Marketplace

RI Marketplace



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

- Reserved Instances can be sold after they have been active for at least 30 days and 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 and capacity benefit of your reservation until it's sold and the transaction is complete.
- Your company name (and 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 and 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 and 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 maximize the utility of their idle servers.



It's like when a hotel offers booking discounts to fill vacant suites or planes offer discount to fill vacant seats

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 and end times or applications that are only feasible at **very low** compute costs.

O Load balancing workloads
Launch instances of the same size, in any Availability Zone. Good for running web services.

Flexible workloads
Launch instances of any size, in any Availability Zone. Good for running batch and CI/CD jobs.

Big data workloads
Launch instances of any size, in a single Availability Zone. Good for MapReduce jobs.



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

Termination Conditions

Instances can be terminated by **AWS at anytime**

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 ran.

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

think of everyone living in an apartment



Multi-Tenant

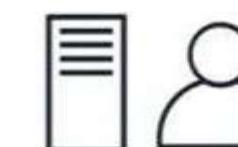


Single Tenant

think of everyone having their own house



Single-Tenant



Single-Tenant



Single-Tenant

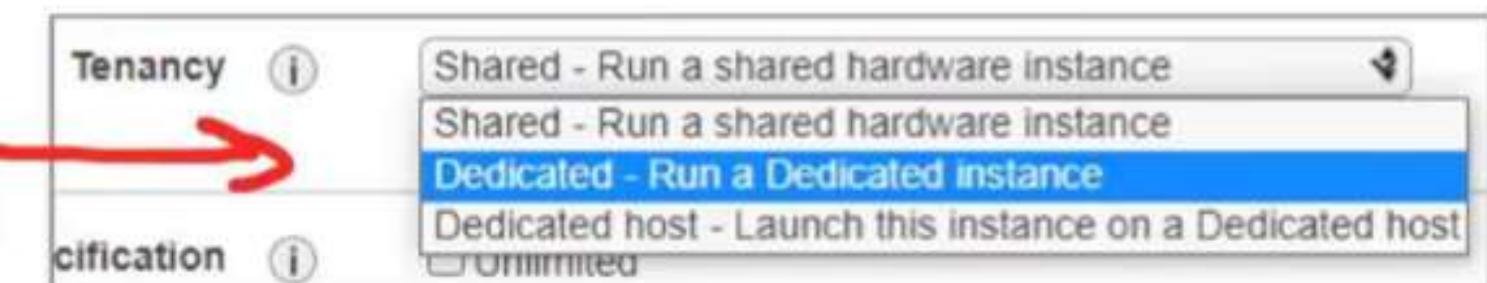
When multiple customers are running workloads on the same hardware. **Virtual Isolation** is what separates customers

When a single customer has dedicated hardware. **Physical Isolation** is what separates customers

Dedicated can be offered for:

- **On-demand**
- **Reserved (up to 60% savings)**
- **Spot (up to 90% savings)**

You choose tenancy when you **launch** your EC2
(Notice there is a Dedicated Host)



Enterprises and Large Organizations may have security concerns or obligations about against sharing the same hardware with other AWS Customers.

AWS Savings Plan

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

There are 3 types of Savings Plans:

- **Compute Savings Plans**
- **EC2 Instance Savings Plans**
- **SageMaker Savings Plan**



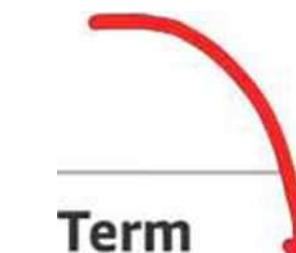
You choose the following Payment Options:

- All Upfront
- Partial Upfront
- No Upfront

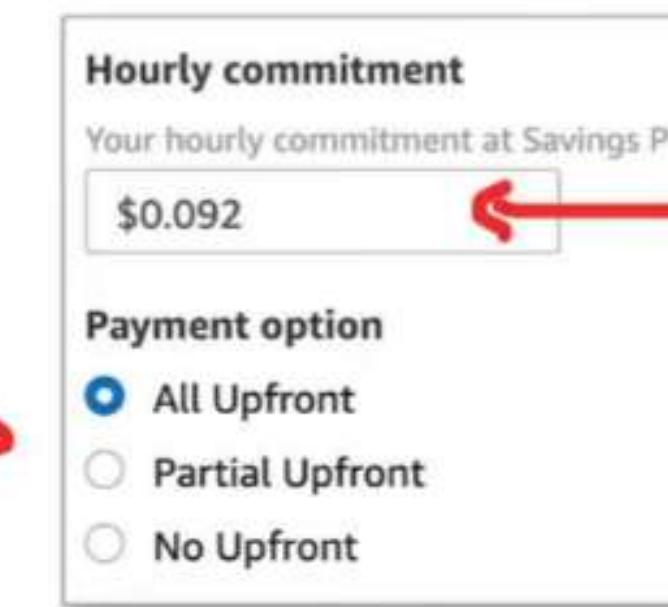
You can choose two different terms

- 1 Year
- 3 Year

•



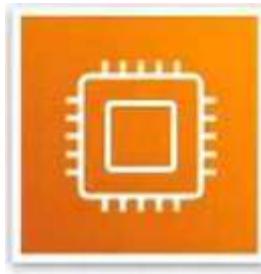
- 1-year
- 3-year



You choose an hourly commitment

AWS Savings Plan

AWS Savings Plan has 3 different savings types:



Compute

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



EC2 Instances

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

Helps you reduce SageMaker costs by up to 64%.
automatically apply to SageMaker usage regardless of instance family, size, component, or AWS region.

Zero Trust Model

The 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 defense and its security controls that protect a company's cloud resources and assets

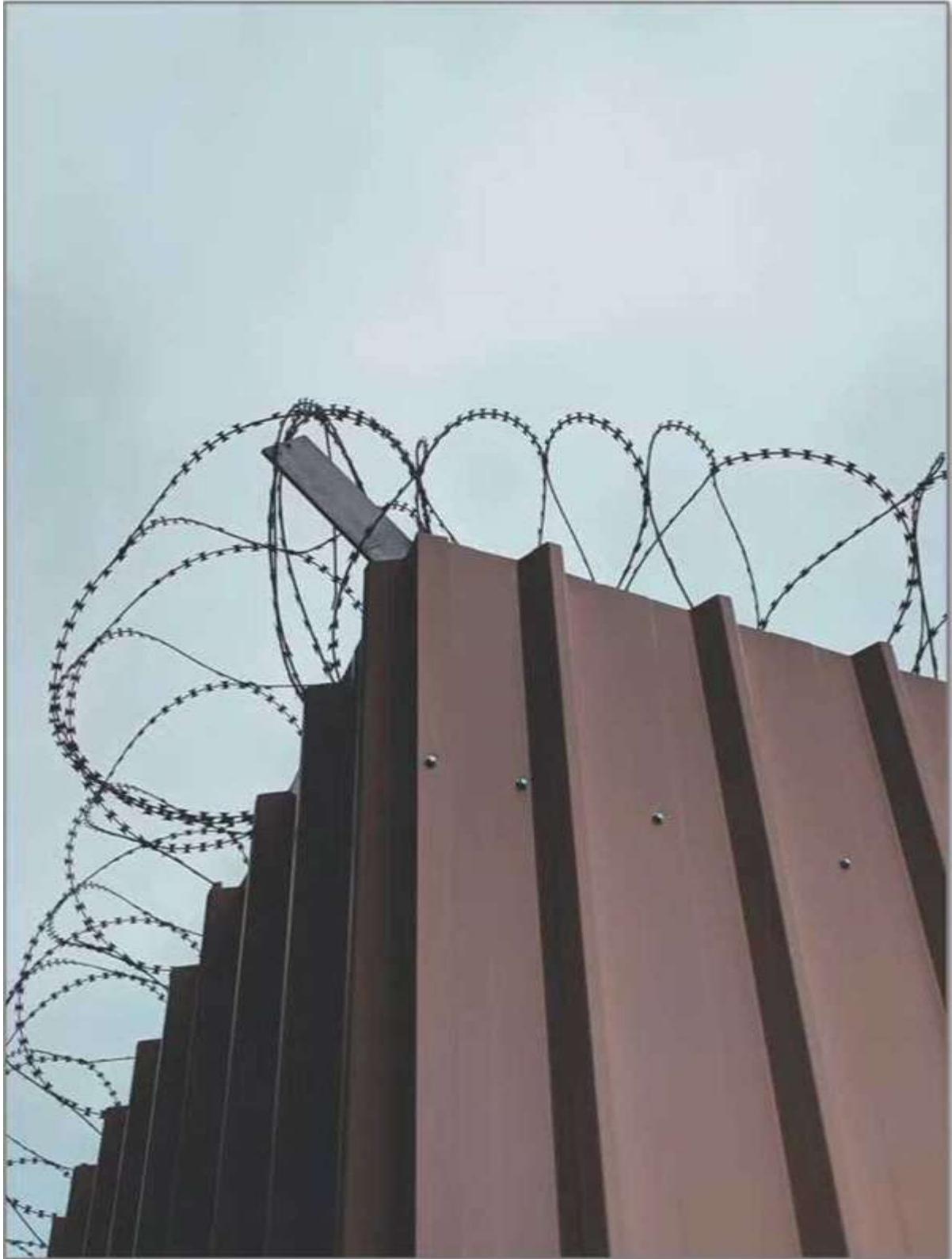
Network-Centric: (Old-Way)

traditional security focused on firewalls and VPNs since there were few employees or workstations outside the office or they were in 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 and 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

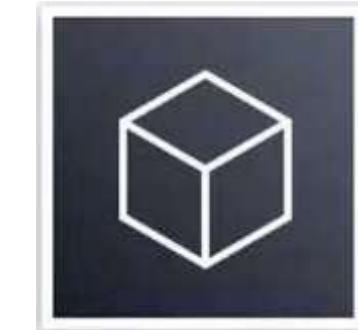


AWS Identity and Access Management (IAM)



IAM Policies
Permission Boundaries
Service Control Policies (Organization-wide Policies)
IAM Policy Conditions

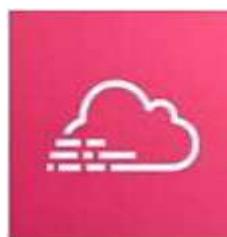
- aws:SourceIp - Restrict on IP Address
- aws:RequestedRegion - Restrict on Region
- aws:MultiFactorAuthPresent - Restrict if MFA is turned off
- aws:CurrentTime - Restrict access based on time of day



Your AWS Resources

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

A collection of AWS Services can be setup to detect identity concerns “intelligently” but requires expert knowledge



AWS CloudTrail

Tracks all API calls



Amazon GuardDuty

Detects suspicious or malicious activity based on CloudTrail and other logs



Amazon Detective

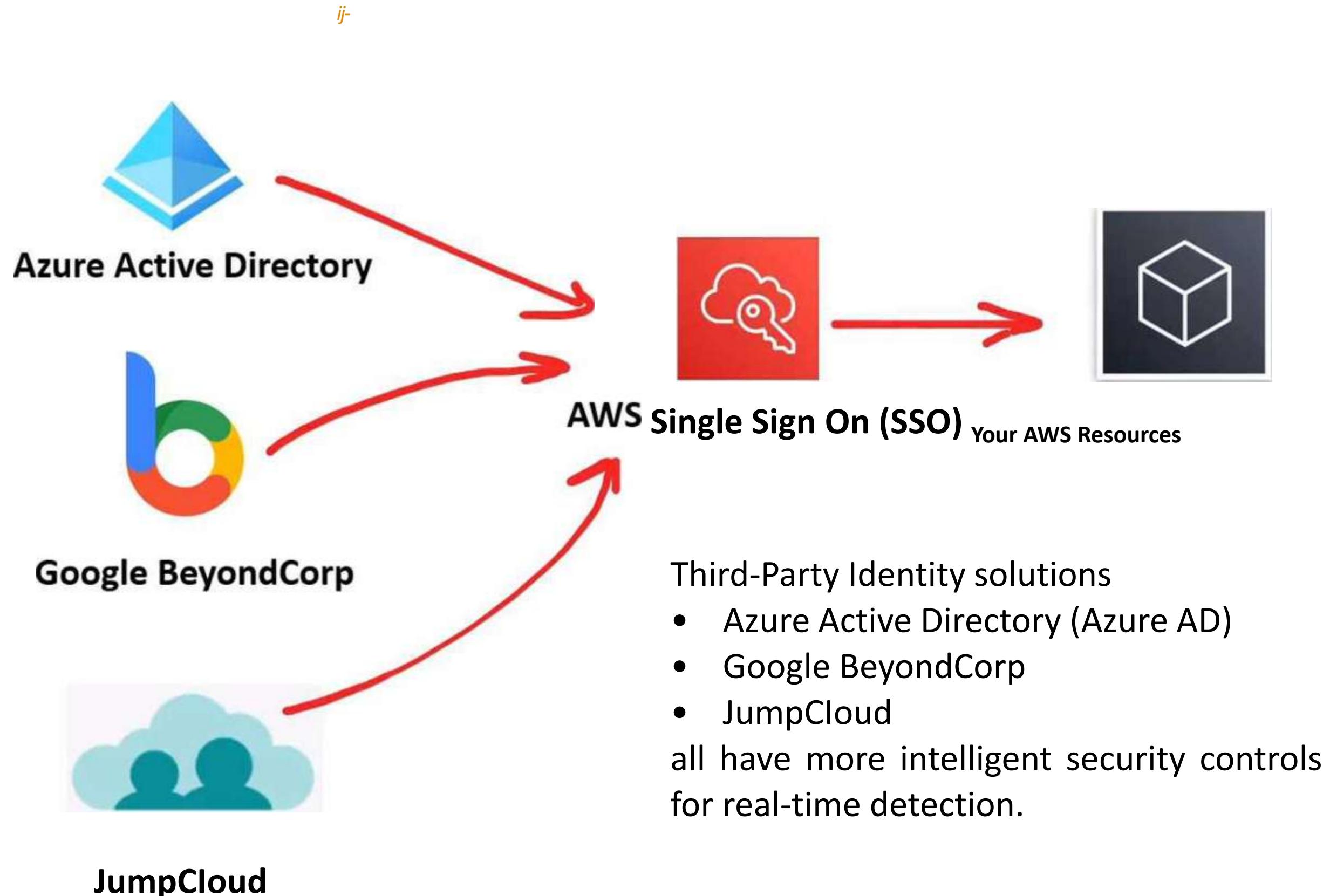
Used to analyze, investigate and quickly identify security issues (can ingest findings from Guard Duty)

Zero Trust on AWS with Third Parties

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

For example:
Azure Active Directory has Real-time and calculated risk detection based more data points than AWS eg:

- Device and Application
- Time of Day
- Location
- MFA turned on
- What is being accessed And the security controls, verifications or logic restriction is much more robust.



Third-Party Identity solutions

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

all have more intelligent security controls for real-time detection.

Directory Service

What is a directory service?

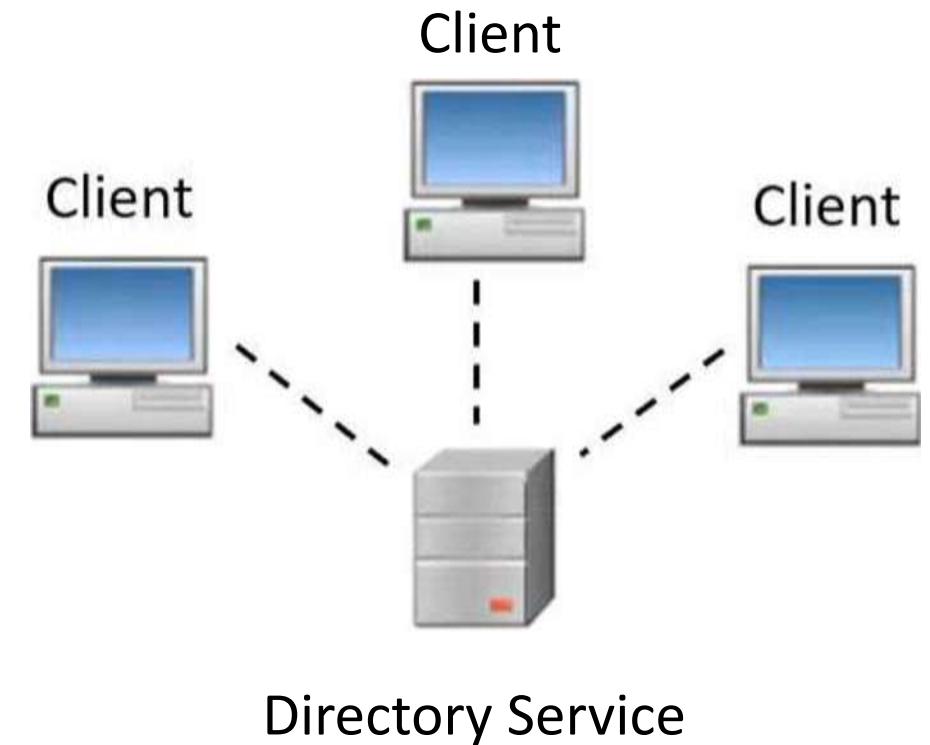
A directory service maps the **names of network resources to their network addresses.**

A directory service is shared information infrastructure for **locating, managing, administering and organizing** resources:

- 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:

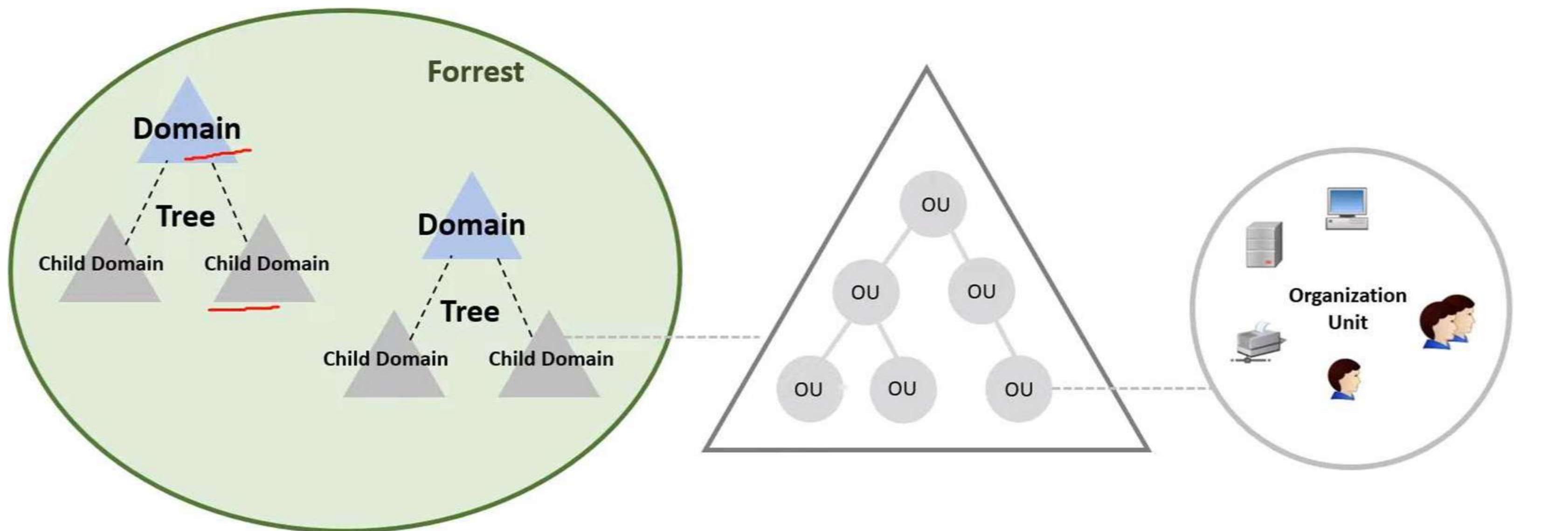
- Domain Name Service (DNS)
 - the directory service for **the internet**
- **Microsoft Active Directory**
 - Azure Active Directory
- Apache Directory Server
- Oracle Internet Directory (OID)
- OpenLDAP
- Cloud Identity
- JumpCloud

Active Directory



Windows 2000

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



Identity Providers (IdPs)

Identity Provider (IdP) a system entity that creates, maintains, and manages identity information for principals and also provides authentication services to applications within a **federation** or distributed network.

A trusted provider of your user identity that lets you use 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 and decentralized authentication protocol. Eg be able to login into a different social media platform using a Google or Facebook account *OpenID is about providing who are you*



OAuth2.0

industry-standard protocol for authorization OAuth doesn't share password data but instead uses authorization tokens to prove an identity between consumers and service providers.

Oauth is about granting access to functionality

SAML

Security Assertion Markup Language is an open standard for exchanging authentication and authorization between an identity provider and a service provider.

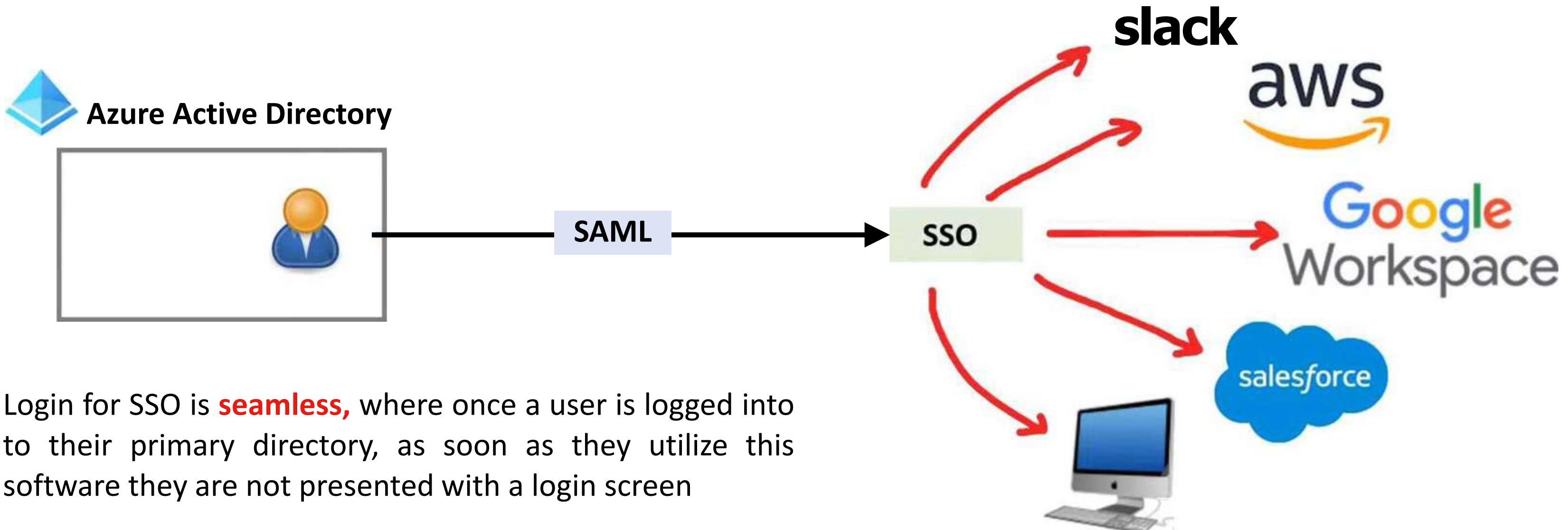
An important use case for SAML is **Single-Sign-On via web browser**.

Single—Sign—On

^

Single sign-on (SSO) is an authentication scheme that **allows a user to log in with a single ID and password to different systems and software.**

SSO allows IT departments to administrator a single identity that can access many machines and cloud services.



LDAP

Lightweight Directory Access Protocol (LDAP) is an open, vendor-neutral, industry standard application protocol for accessing and 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 and passwords

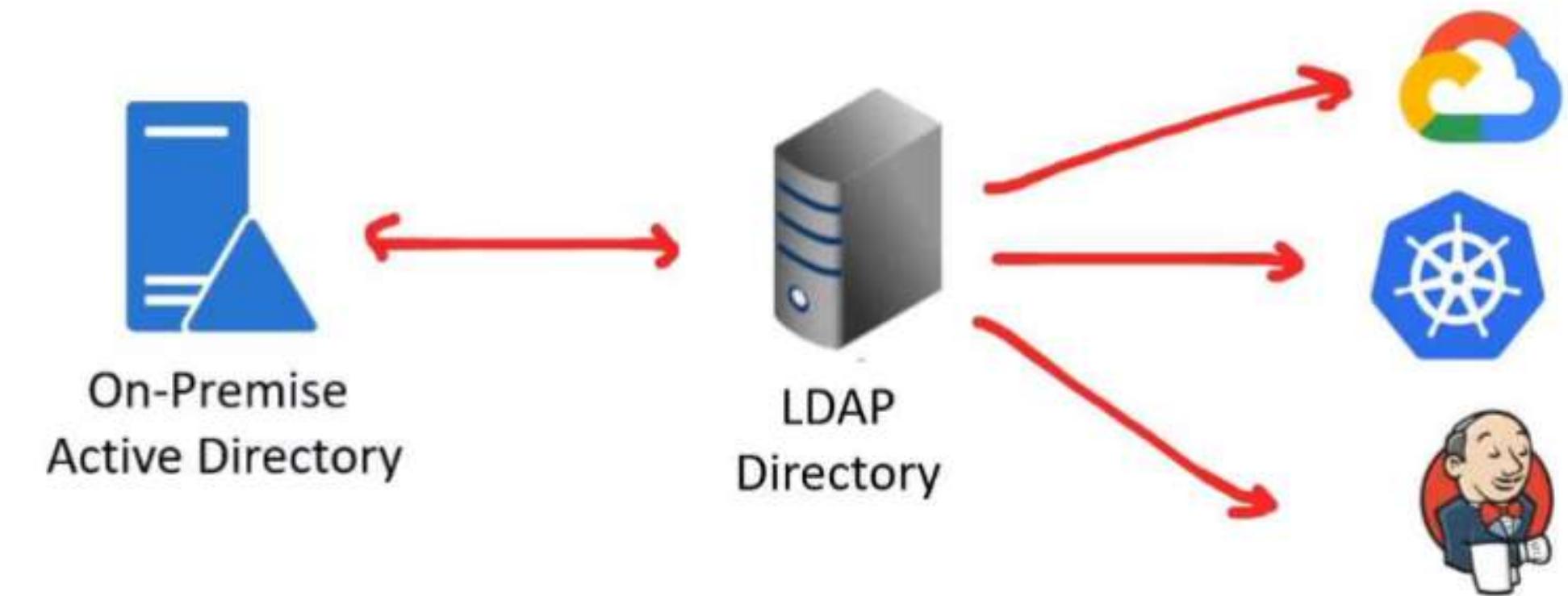
LDAP enables for **same-sign on**. Same sign-on allows users to single ID and password, but they have to enter it in every time they want to login.

Why use LDAP when SSO is more convenient?

Most SSO systems are using LDAP.

LDAP was not designed natively to work with web-applications.

Some systems only support integration with LDAP and not SSO



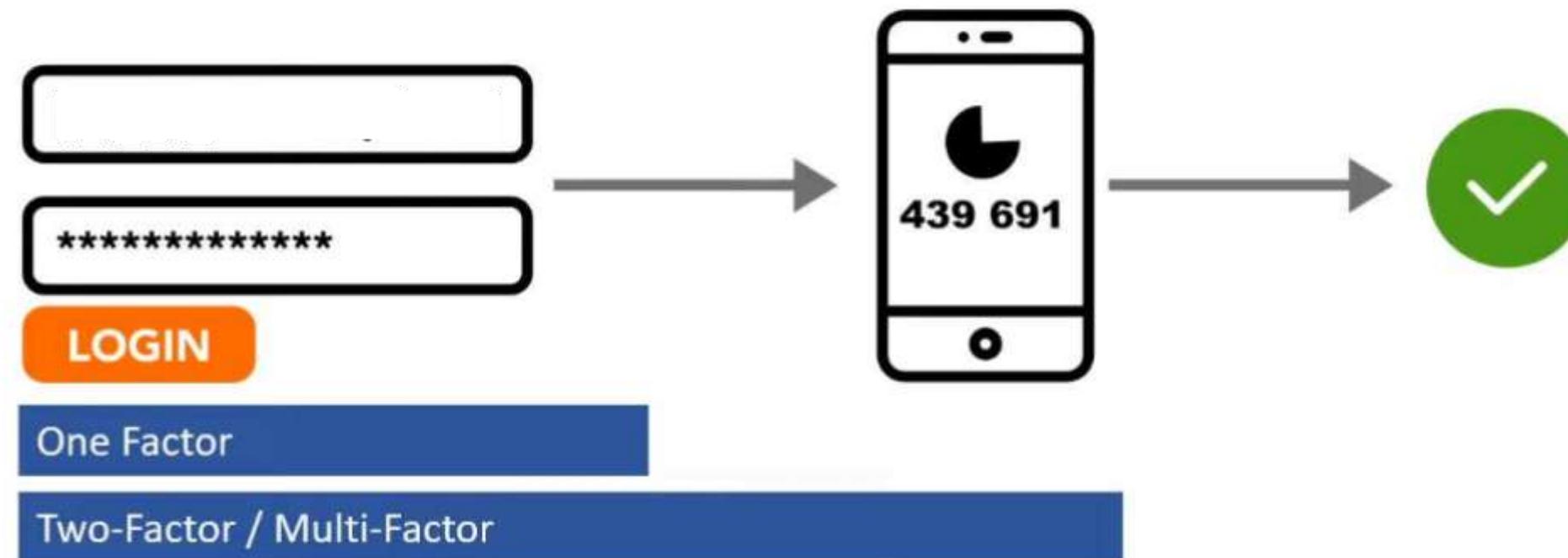
Multi-Factor Authentication

What is Multi-Factor Authentication (MFA)?

A security control where after you fill in your username/email and password **you have to use a second device** such as a phone to confirm that its you logging in.

MFA **protects** against people who have stolen your password.

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



Security Keys

What is a Security Key?

A secondary device used as second step in 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 will generate And autofill a security token.

Manage MFA device

Choose the type of MFA device to assign:

Virtual MFA device

Authenticator app installed on your mobile devi<

• U2F security key

YubiKey or any other compliant U2F device

Other hardware MFA device

Gemalto token

A popular brand of security key is an Yubikey



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

AIajs Identity and Access Management (IAM)



AWS Identity and Access Management (IAM) you can create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources.

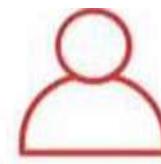


IAM Policies JSON documents which grant permissions for a specific user, group, or role to access services. Policies are attached to **IAM Identities**

IAM Permission

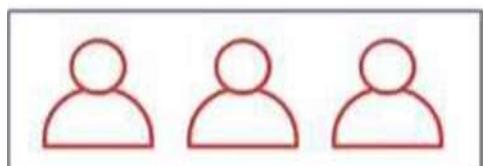
The API actions that can or cannot be performed.
They are represented in the IAM Policy document

IAM Identities



IAM Users

End users who log into the 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 and then assign it to an AWS resource

Anatomy of an IAM Policy

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

Version policy language version.

2012-10-17 is the latest version.

Statement container for the policy element you are allowed to have multiples

Sid (optional) a way of labeling your statements.

Effect Set whether the policy will Allow or Deny

Action list of actions that the policy allows or denies

Principal account, user, role, or federated user to which you would like to allow or deny access

Resource the resource to which the action(s) applies

Condition (optional) circumstances under which the policy grants permission

Anatomy of an IAM Policy

```
{ "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowViewAccountInfo",
      "Effect": "Allow",
      "Action": "iam>ListVirtualMFADevices",
      "Resource": "*"
    },
    {
      "Sid": "AllowManageOwnVirtualMFADevice",
      "Effect": "Allow",
      "Action": [
        "iam>CreateVirtualMFADevice",
        "iam>EnableMFADevice",
        "iam GetUser",
        "iam>ListMFADevices",
        "iam>ListVirtualMFADevices",
        "iam>ResyncMFADevice",
        "sts>GetSessionToken"
      ],
      "Resource": "aws:MultiFactorAuthPresent":
        "arn:aws:iam::*:user/${aws:username}false"
    }
  ],
  "DenyAllExceptListedIfNoMFA": {
    "Effect": "Deny",
    "NotAction": [
      "iam>DeactivateMFADevice",
      "iam>EnableMFADevice",
      "iam GetUser",
      "iam>GetMFADevice",
      "iam>ListMFADevices",
      "iam>ResyncMFADevice"
    ],
    "Sid": "DenyAllExceptListedIfNoMFA"
  }
},
```

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 a 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

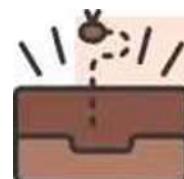


ConsoleMe is an open-source Netflix project to self-serve short-lived IAM policies so an end user can access AWS resources while enforcing JEA and JIT

<https://github.com/Netflix/consoleme>

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 e.g. device, user location, IP address what service is being accessed and when.



AWS at the time of this recording does not have Risk-based adaptative policies built into IAM

AWS Account Root User

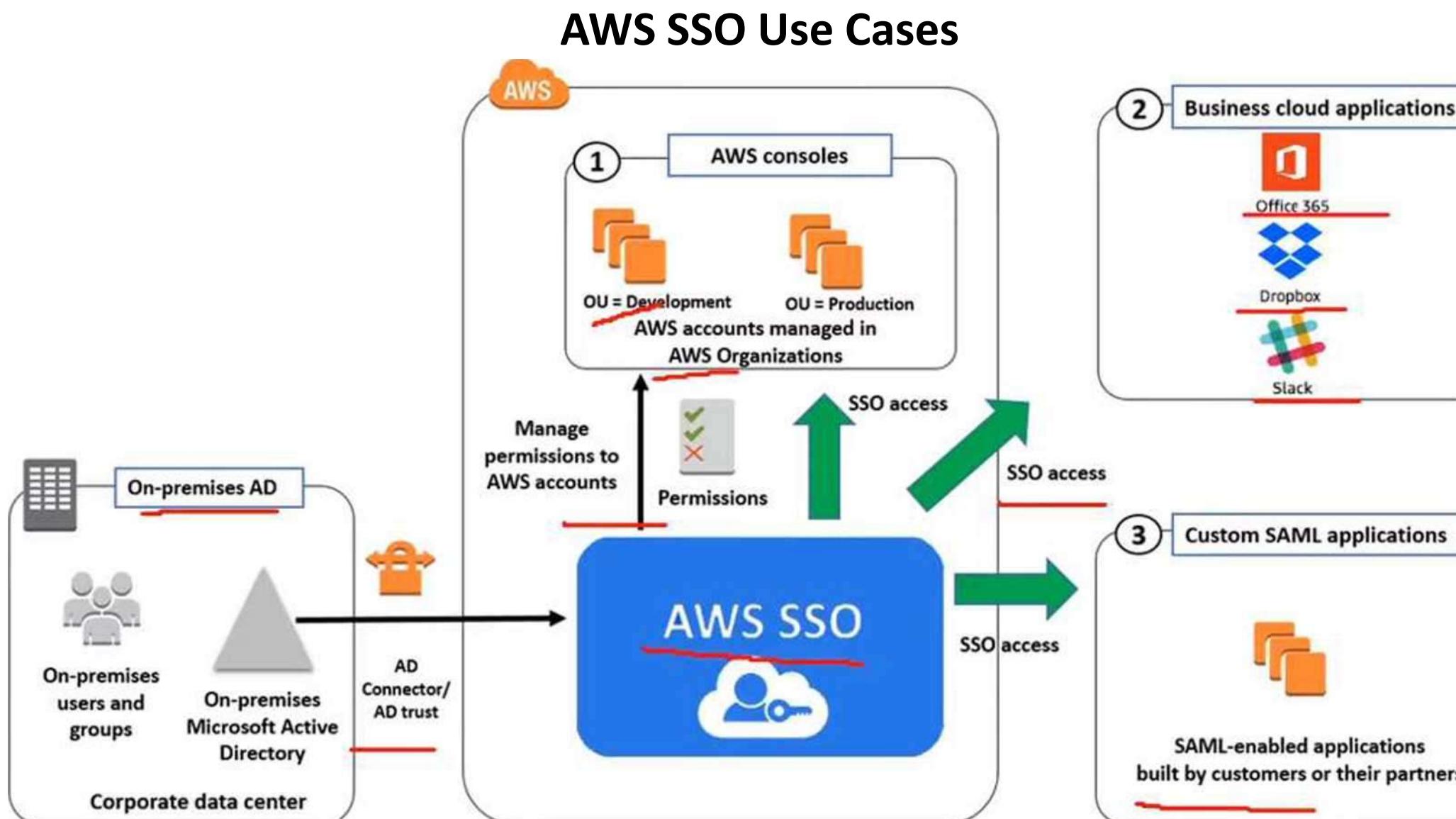
Administrative Tasks **that only the Root User can perform:**

- **Change your account settings.**
 - includes the account name, email address, root user password, and root user access keys.
 - Other account settings, such as contact information, payment currency preference, and Regions, do not 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 and restore those permissions.
- Activate IAM access to the Billing and 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** and manage access centrally across your AWS organization.



Choose your Identity Source

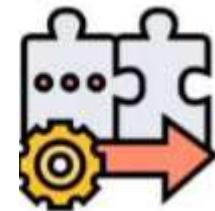
- AWS SSO
- Active Directory
- SAML 2.0 IdP

Managed User Permissions Centrally

- AWS Account
- AWS Applications
- SAML Applications

Uses get Single Click Access

Application Integration



What is Application Integration?

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



Cloud workloads encourage systems and services to be loosely coupled and so AWS has many services for the specific purpose of application integration.

The common systems or design patterns utilized for Application Integration generally are:

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

Queueing

What is a Messaging System?

Used to provide asynchronous communication and decouple processes via messages / events From a sender and receiver (producer and consumer)

What is a Queueing System?

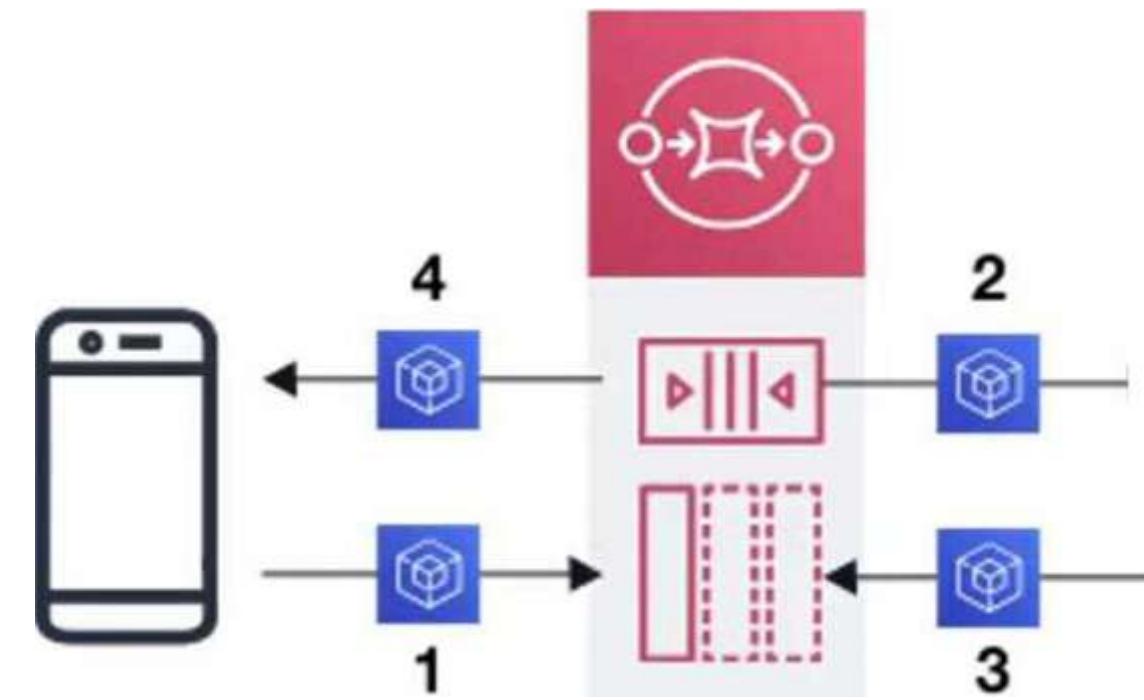
A Queueing system is a messaging system that generally will delete messages once they are consumed. Simple communication. **Not Real-time**. Have to pull. Not reactive.



Simple Queueing Service (SQS)

Fully managed **queueing service** that enables you to decouple and scale microservices, distributed systems, and serverless applications

Use Case: You need to queue up transaction emails to be sent e.g. Signup, Reset Password.



Streaming

What is streaming?

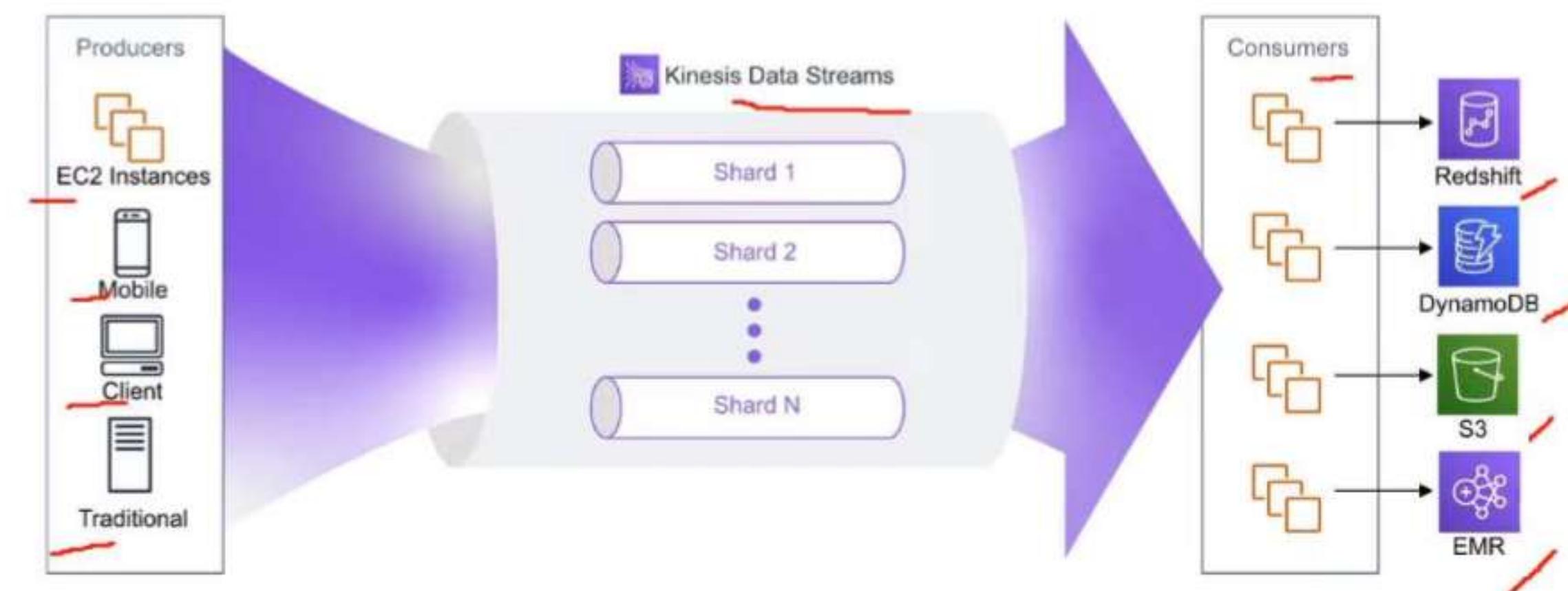
Multiple consumers can **react** to events (messages)

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, and analyzing streaming data in the cloud.



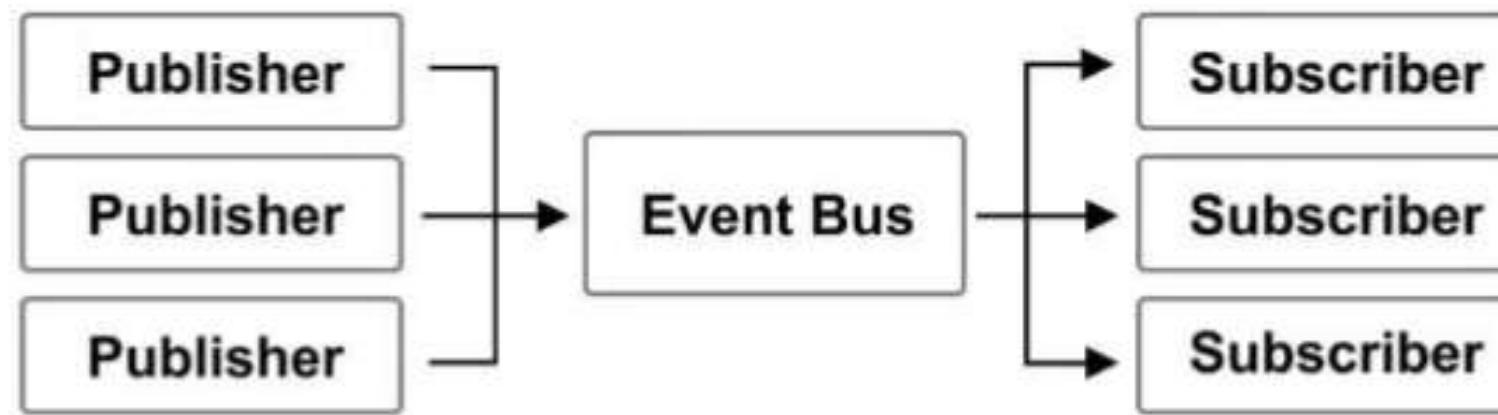
Pub/Sub

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 categorizes 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.



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

Publisher have no knowledge of who their subscribers are.

Subscribers do **not pull** for messages.

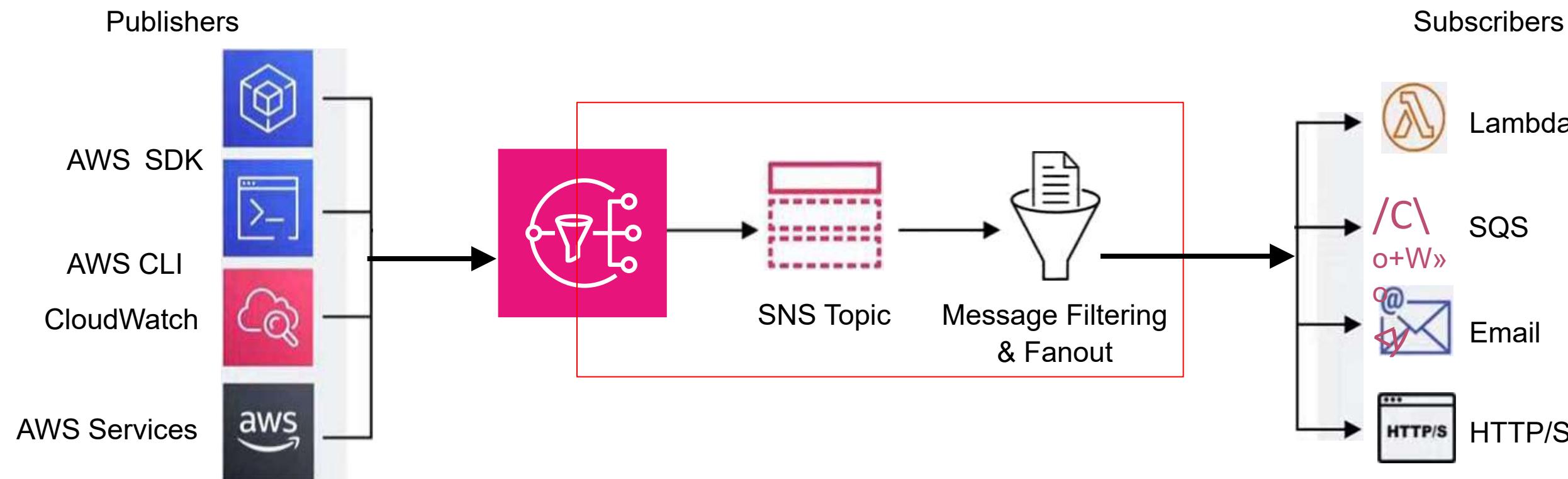
Messages are instead automatically and immediately **pushed** to subscribers.

Messages and events are interchangeable terms in pub/sub

Pub/Sub



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

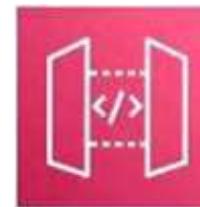


API Gateway

What is an API Gateway?

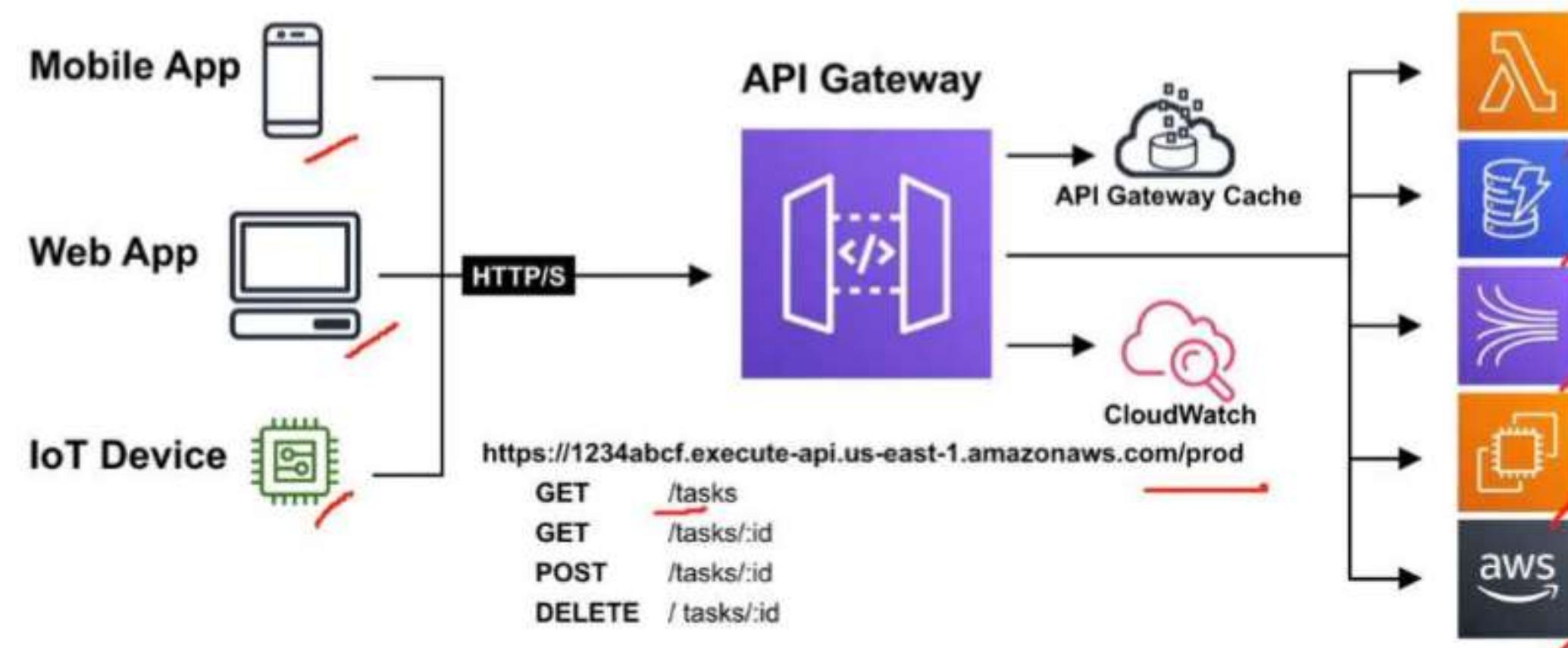
An API Gateway is a program that sits between a single-entry point and multiple backends.

API Gateway allows for throttling, logging, routing logic or formatting of the request and 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?

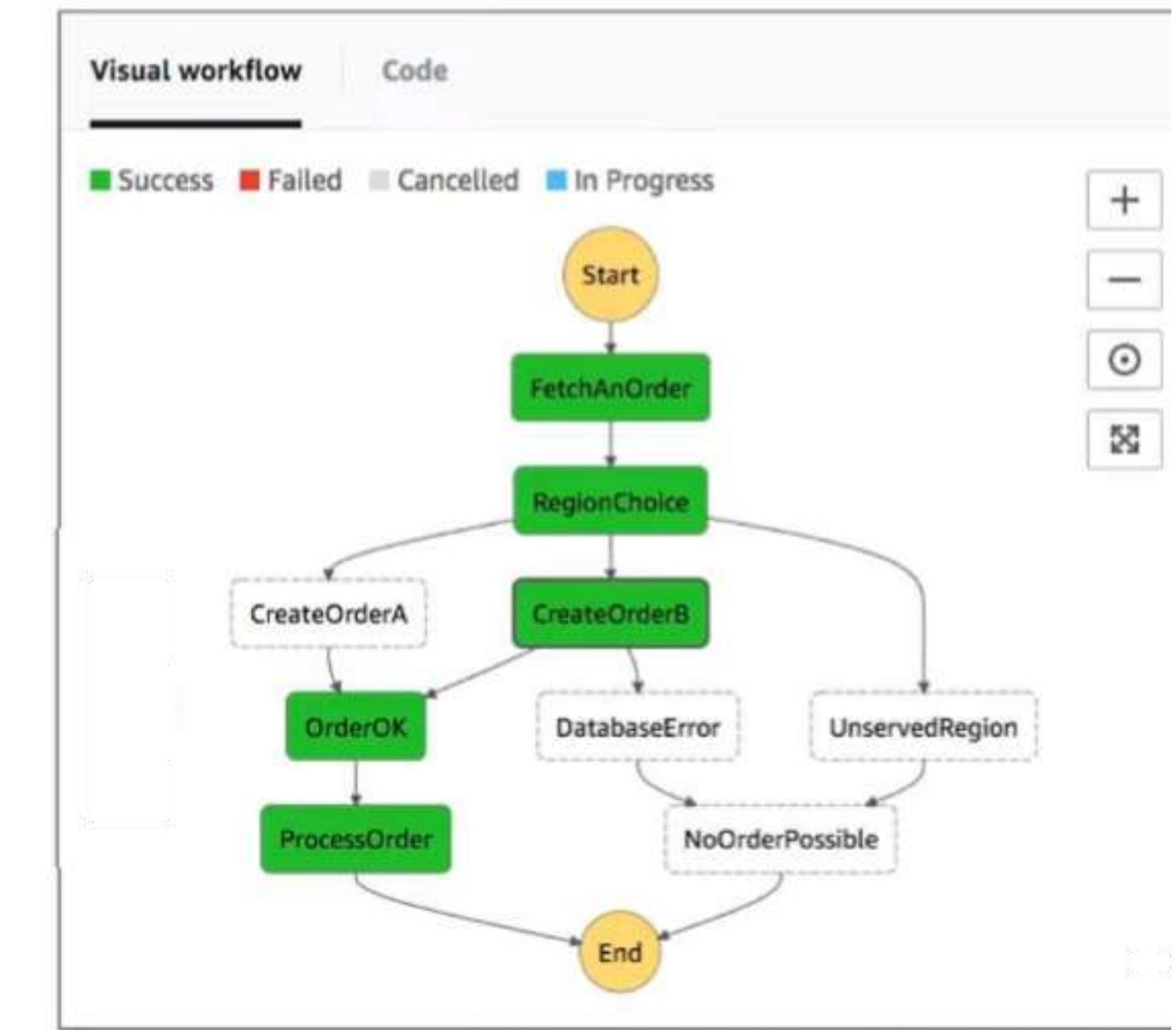
A state machine is an abstract model which decides how one state moves to another based on a series of conditions. **Think of a state machine like a flow chart.**



What is AWS Step Functions?

- Coordinate multiple AWS Services into a serverless workflow
- A graphical console to visualize the components of your application as a series of steps.
- Automatically triggers and tracks each step, and retries when there are errors, so your application executes in order and as expected, every time
- logs the state of each step, so when things go wrong, you can diagnose and debug problems quickly

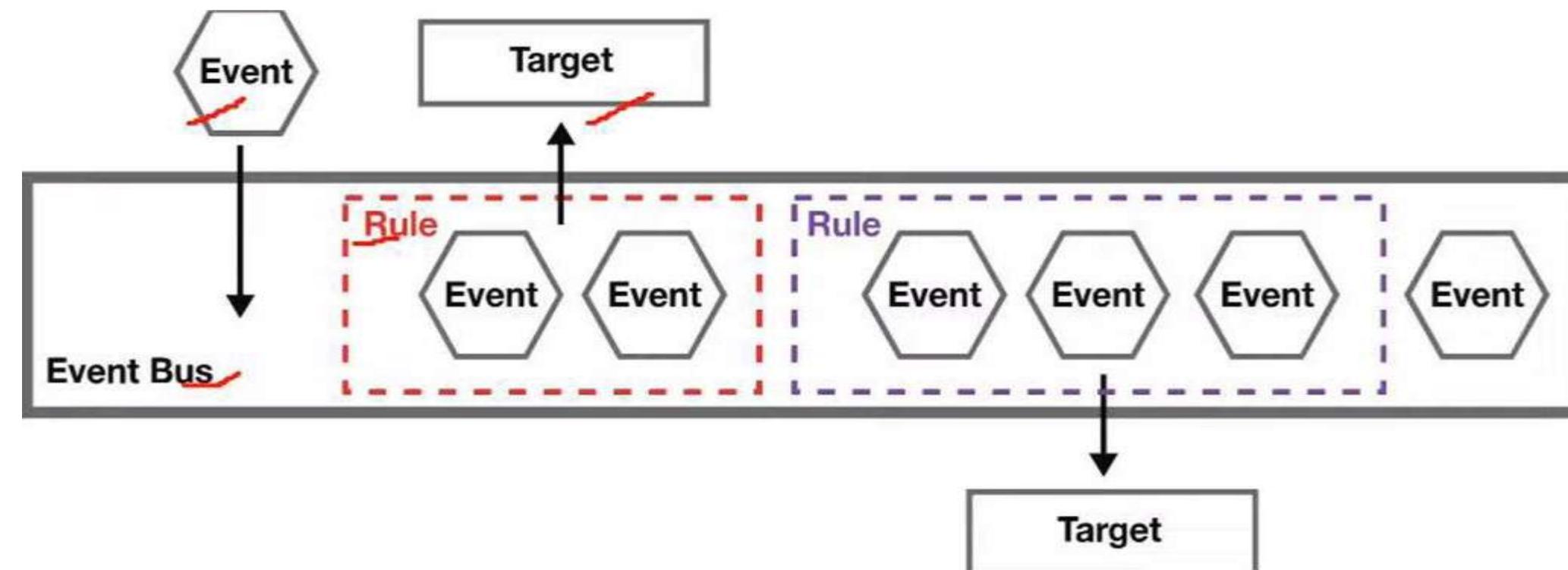
Any one of these steps could be using an AWS Service



Event Bus

What is an Event Bus?

An event bus **receives events** from a **source** and **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**.

Amazon Event Bridge

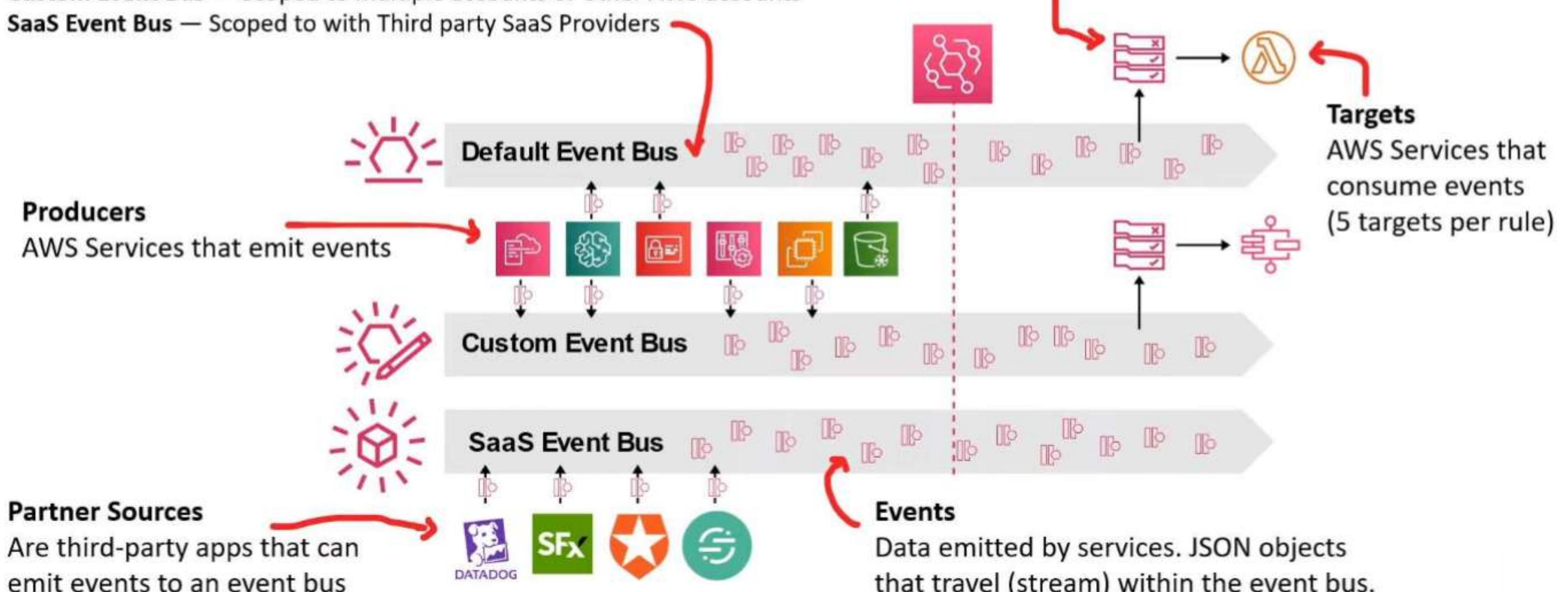
Event Bus

Holds 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 to with Third party SaaS Providers



Application Integration Services



Simple Notification Service (SNS) - a pub-sub messaging system. Sends notifications via various formats such as Plain text **Email**, HTTP/s (**webhooks**) SMS (**text messages**), **SQS** and **Lambda**. Push messages which then are sent to subscribers



Simple Queue Service (SQS) is a queueing messaging service. Send events to a queue. Other applications pull the queue for messages. Commonly used for background jobs.



Step Functions is a state machine service. It coordinate 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) is a serverless event bus that makes it easy to connect applications together from your own application, third-party services and AWS services.



Kinesis 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, click streams, ingesting data from a fleet of IOT Devices



Amazon MQ is 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 and applications. Similar to Kinesis but more robust



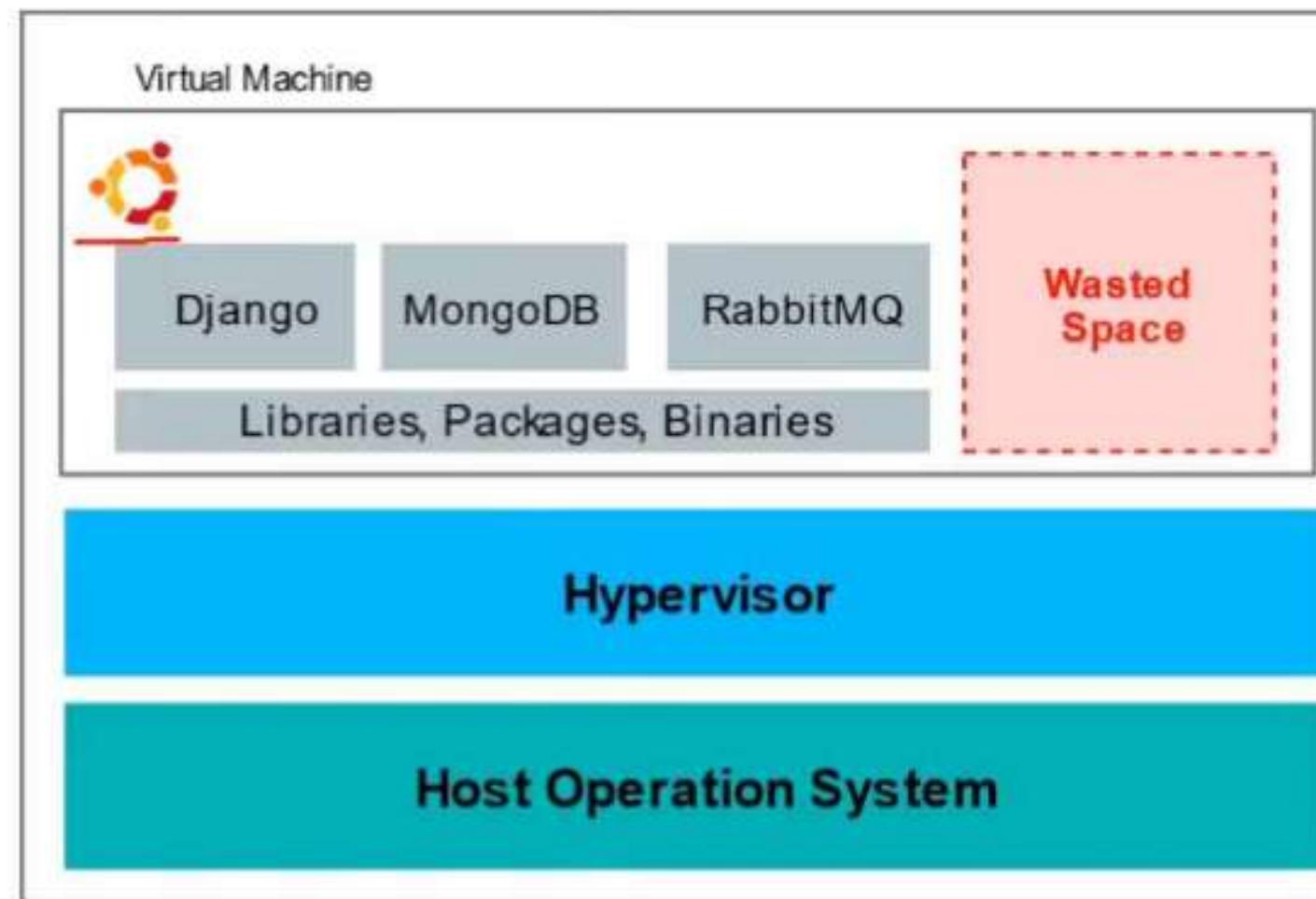
API Gateway is a fully-managed service for developers to create, publish, maintain, monitor, and secure You can create API endpoints and route them to AWS services.



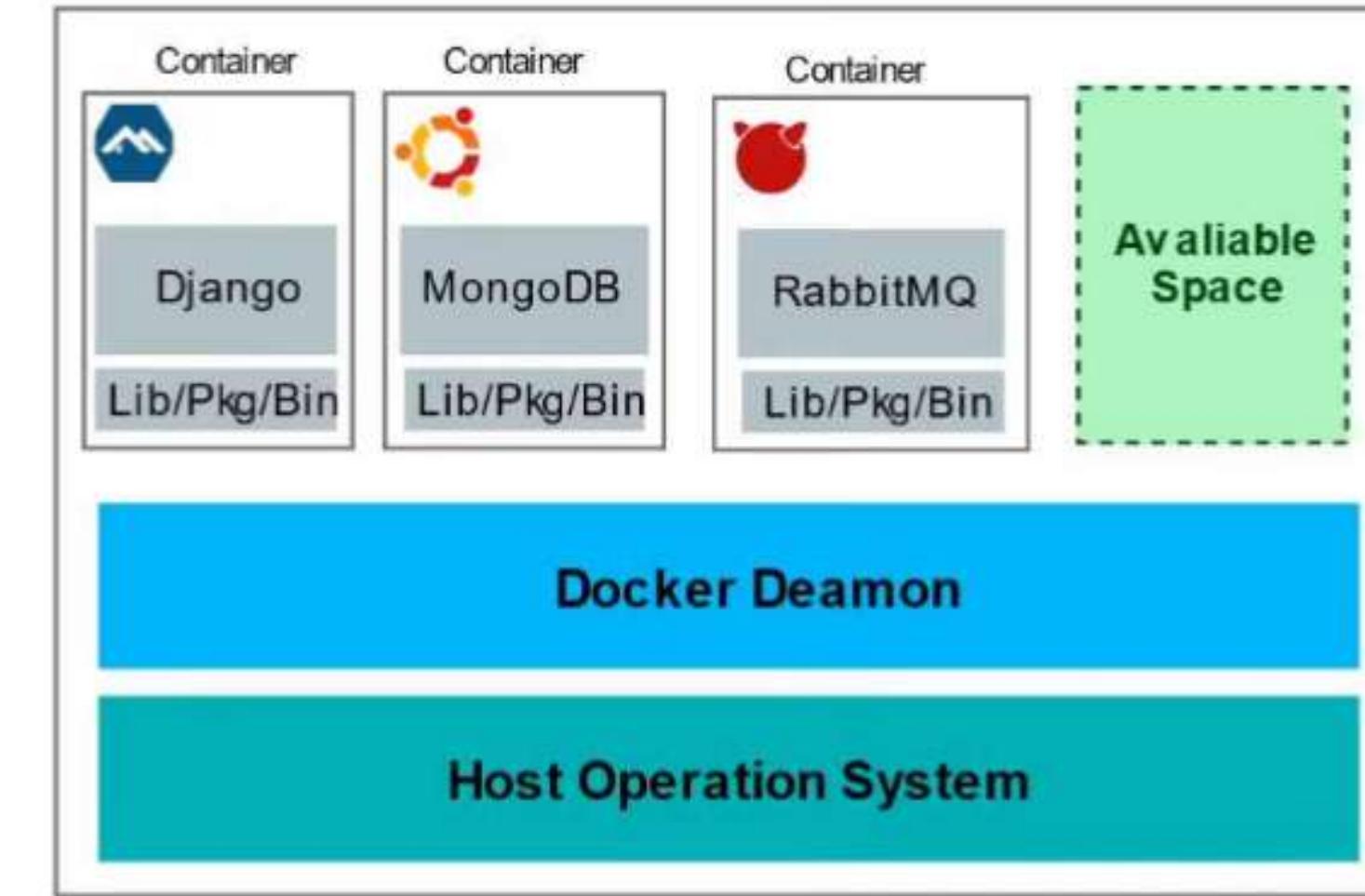
AppSync is a fully managed GraphQL service. GraphQL is an open-source agnostic query adaptor that allows you to query data from many different data sources.

VMs vs Containers

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



EC2 Instance

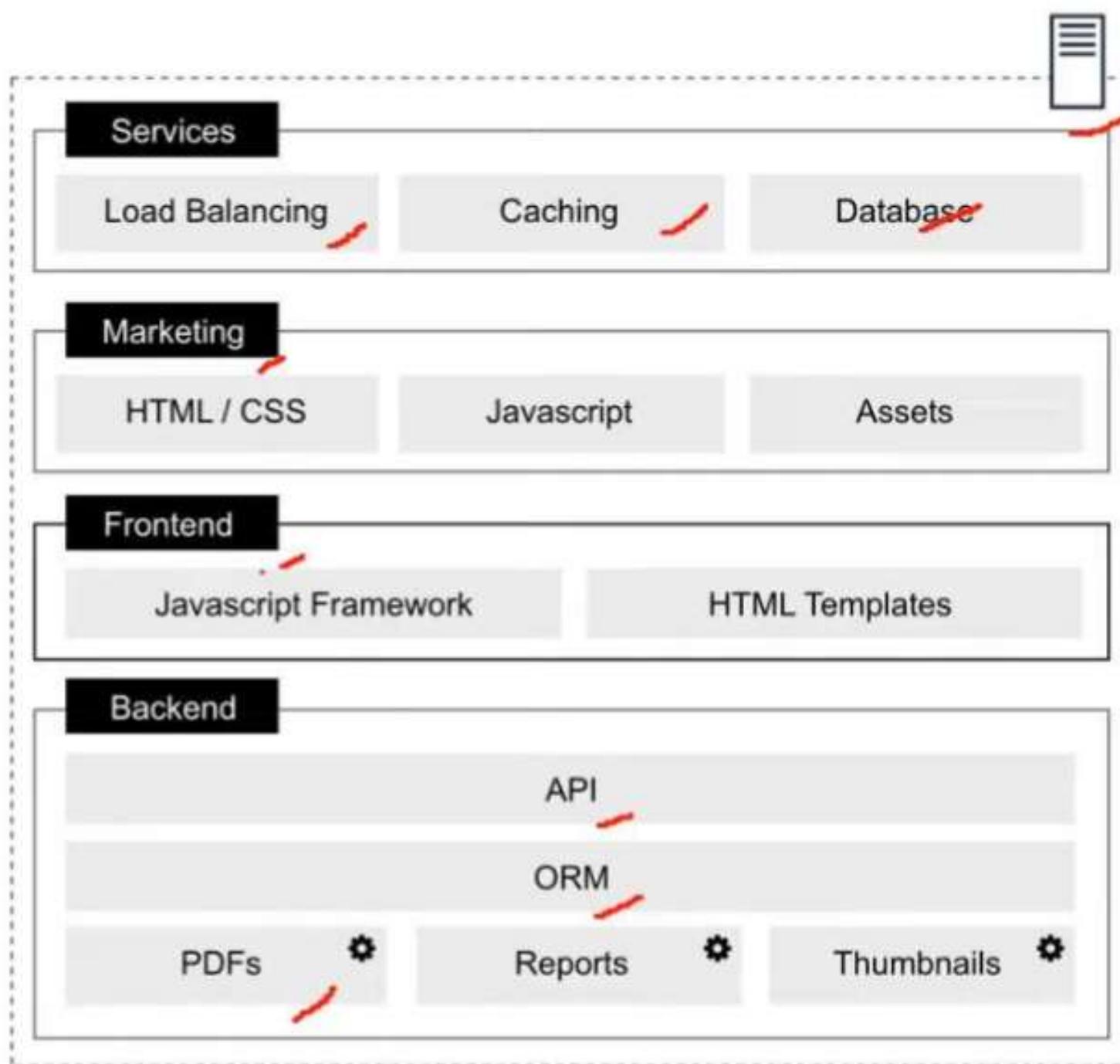


EC2 Instance

What are Microservices

Monolithic Architecture

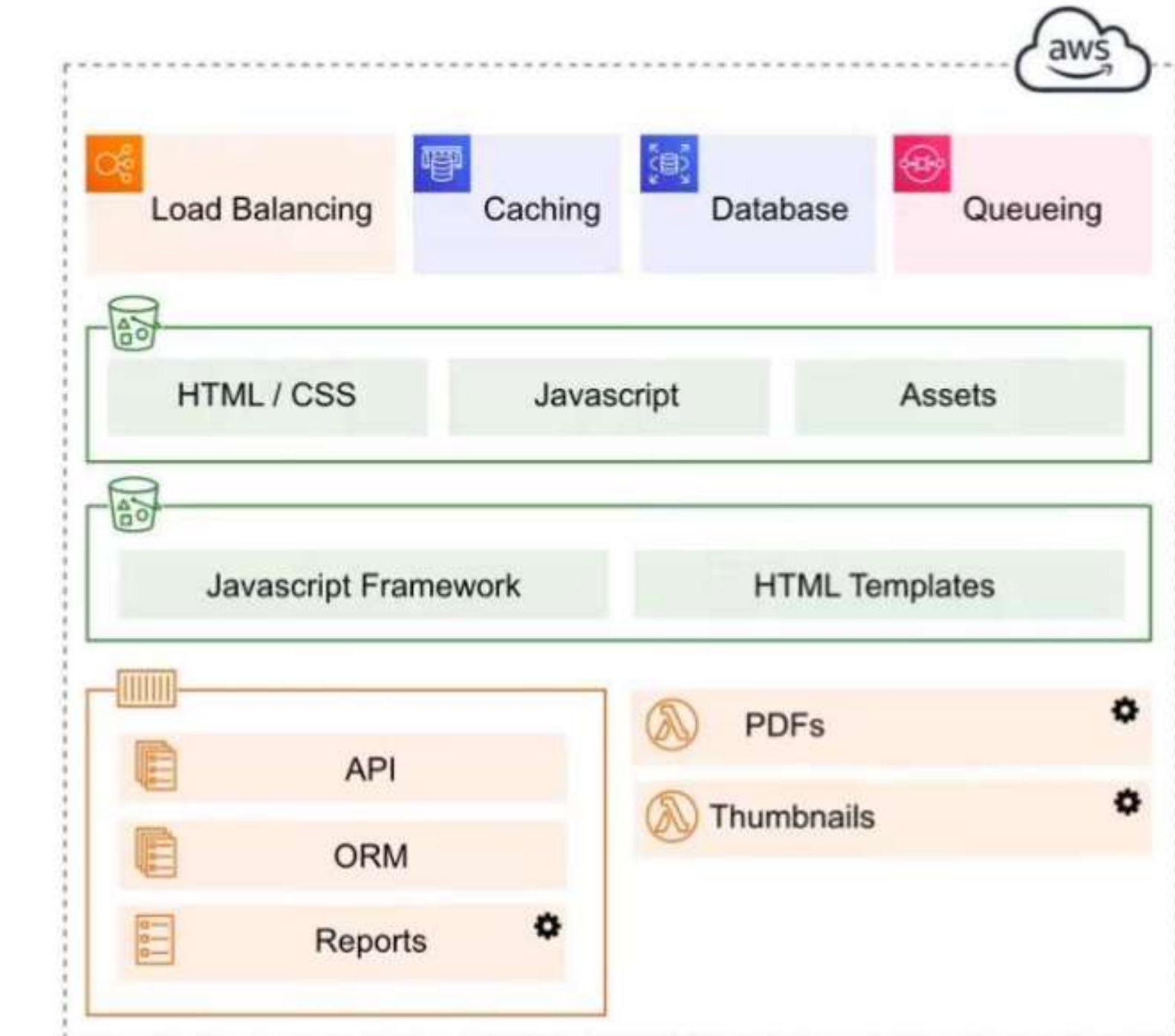
One app which is responsible for everything
Functionality is tightly coupled



VS

Microservices Architecture

Multiple apps which are each responsible for one thing
Functionality is isolate and stateless



Kubernetes



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



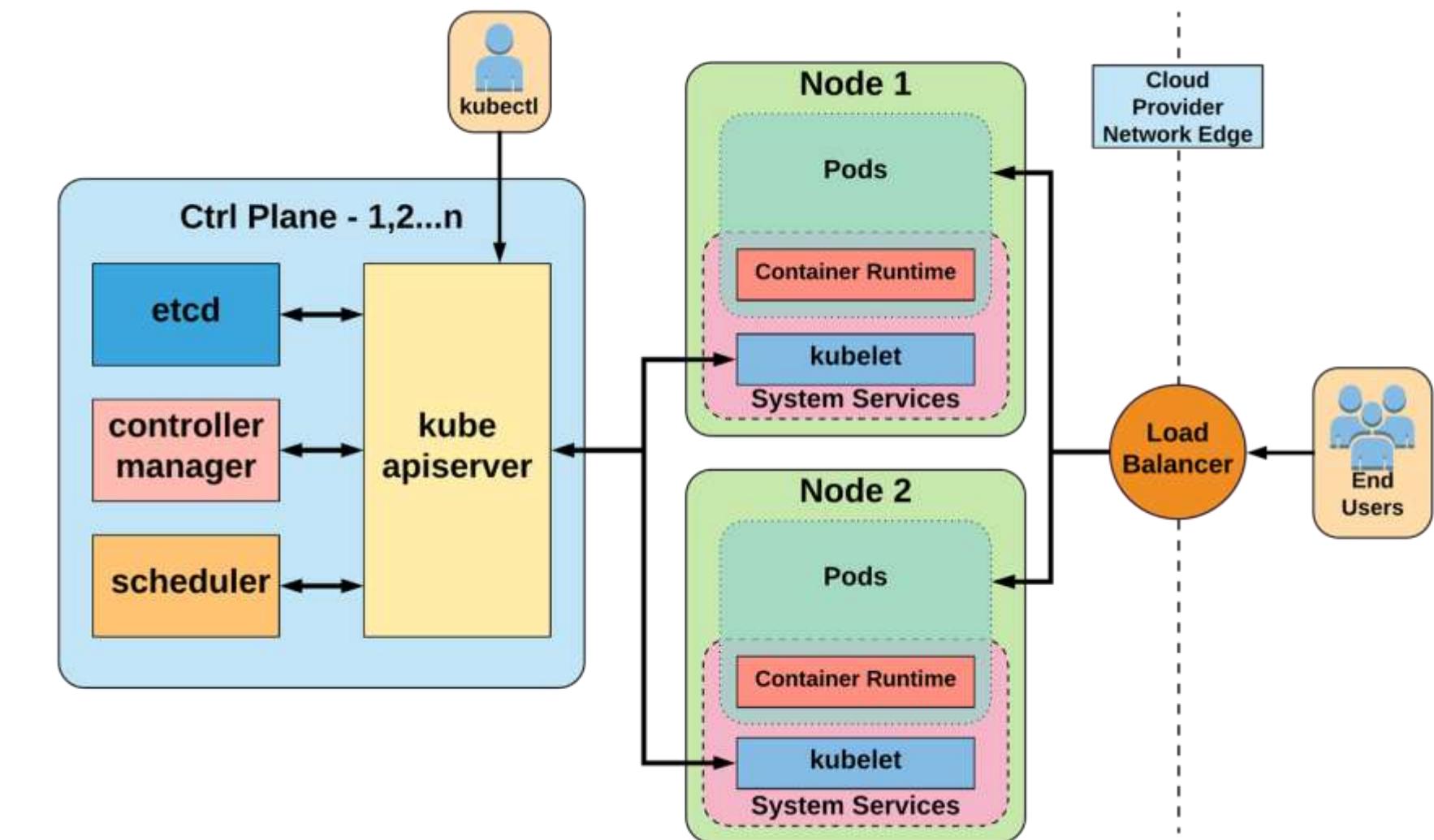
Originally created by Google and now maintained by the **Cloud Native Computing Foundation (CNCF)**

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

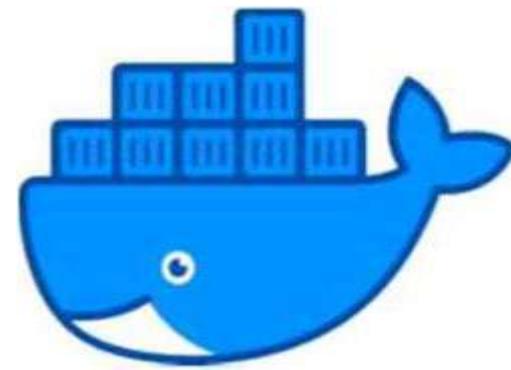
A unique component of Kubernetes are **Pods**.

A pod is a group of one more containers with shared storage, network resources and other shared settings.



Kubernetes is ideally for micro-service architectures where a company has tens to hundreds of services they need to manage

Docker



Docker is a set of Platform as a Service (PaaS) products that use OS-level virtualization to deliver software in packages called containers.

Docker was the earliest popularized open-source container platform.

When people think of containers, they think of Docker.

```
FROM python:3.8-alptne3.12
COPY . /app WORKDIR /app
RUN pip Install -r requirements.txt
CMD ["python3", "app.py"]
```

Docker CLI - CLI commands to download, upload, build run and debug containers

Dockerfile - a configuration file on how to provision a container

Docker Compose - is a tool and 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 and runtime. Docker established the OCI and it is now maintained by the Linux Foundation.

Docker has been losing favor with developers due to their handling of introducing a paid open-source model and alternative like Podman are growing.

Podman, Buildah and Skopeo



Podman is a container engine that is OCI-compliant and is a drop-in replacement for Docker.

Podman is daemon-less where Docker uses a contained deamon Podman allows you to create pods like K8, Docker does not have pods
Podman only replaces one part of Docker. Podman is to be used alongside Buildah and Skopeo



Buildah is a tool used to build OCI Images



Skopeo 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 and 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 and operate production ready
containerized applications on AWS App Runner,
Amazon ECS, and AWS Fargate

Supporting Services



Elastic Container Registry (ECR)
Repos for your Docker Images



X-Ray
Analyze and debug between
microservices



Step Functions
Stitch together Lambdas and ECS tasks

Organizations and Accounts



AWS Organizations allow the creation of new AWS accounts. Centrally manage billing, control access, compliance, security, and share resources across your AWS accounts.



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

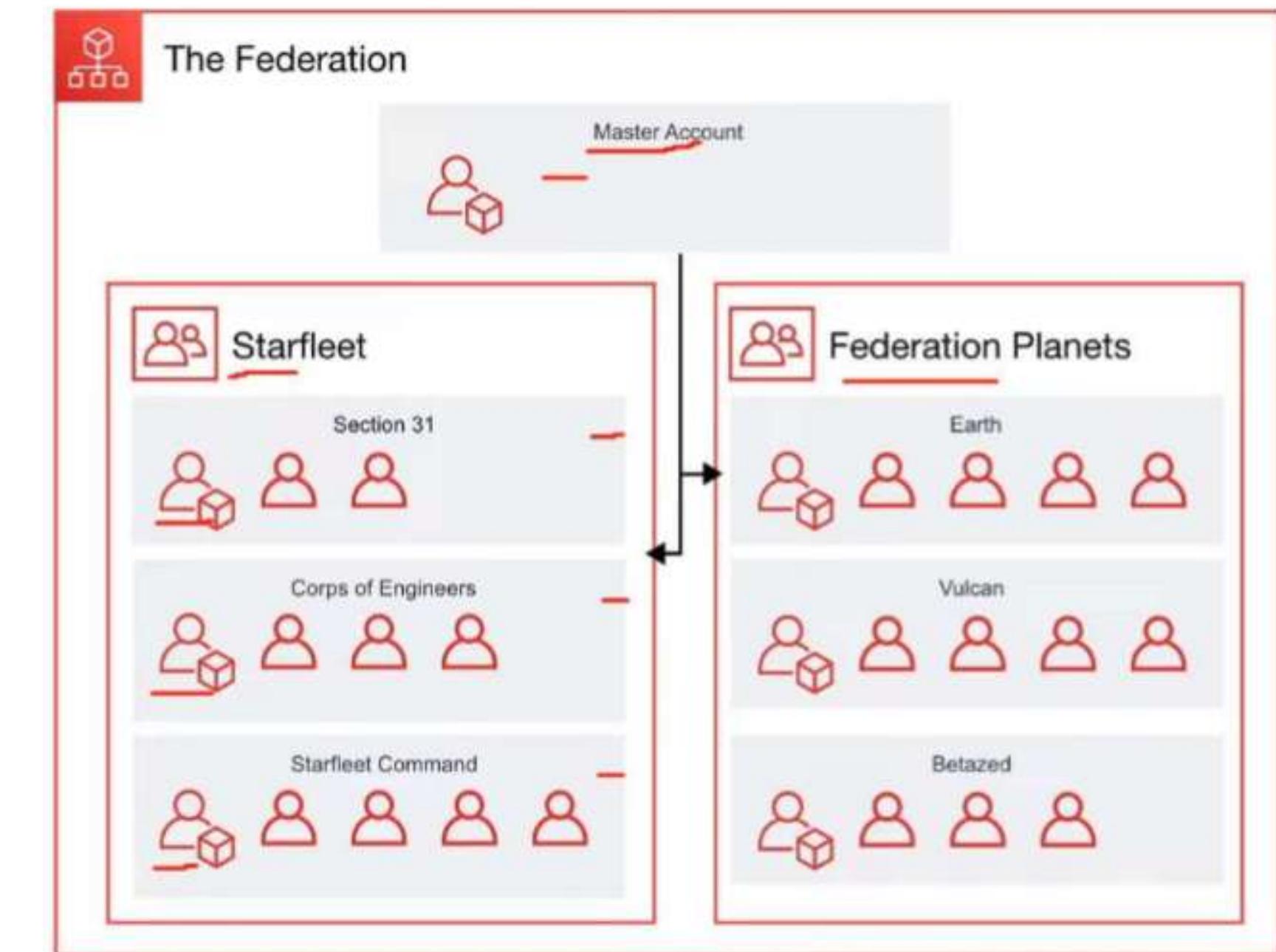


Organization Units are a group of AWS accounts within an organization which can also contain other organizational units - creating a hierarchy

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

AWS Organizations must be turned on, once turned it cannot be turned off.

You can create as many AWS Accounts as you like, one account will be the Master/Root Account

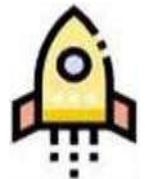


AWS Control Tower

▲



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-architected and best practices to start launching production ready workloads.

- AWS SSO enabled, Centralized logging for AWS CloudTrail, cross-account security auditing



Account Factory

- automates provisioning of new accounts in your organization
- standardize the provisioning of new accounts with pre-approved account configurations
- configure your account factory with pre-approved network configuration and region selections
- enable self-service for your builders to configure and provision new accounts using AWS Service Catalog



Guardrails

pre-packaged governance rules for security, operations, and compliance that customers can select and apply enterprise-wide or to specific groups of accounts

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

AWS Config

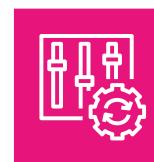
What is Change management?

Change management in the context of Cloud Infrastructure is when we have **formal process** to:

- monitor changes
- enforce changes
- Remediate changes

What is Compliance-as-code (CaC)?

Compliance as code is when we utilize programming to automate the monitoring, enforcing and remediating changes to stay compliant with a compliance programs or expected configuration.

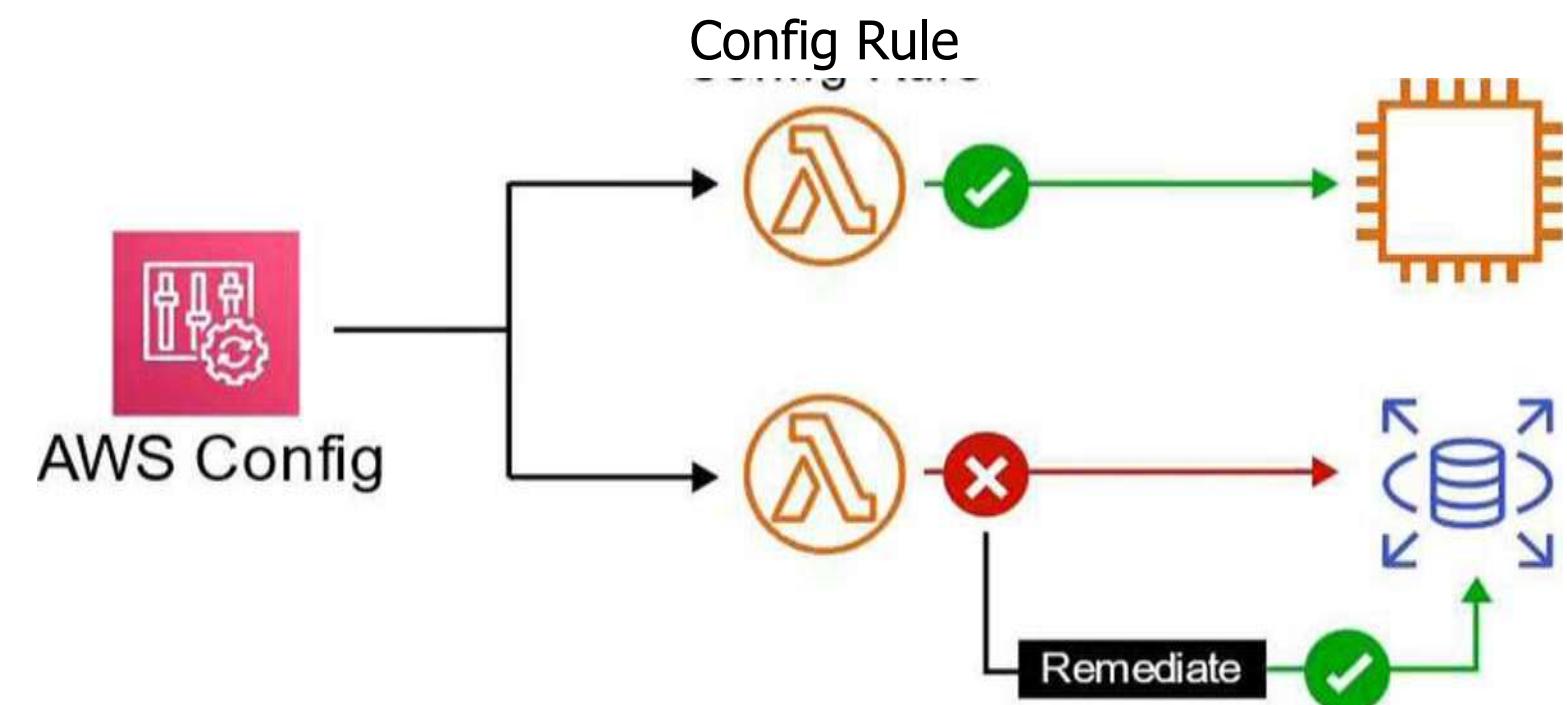


What is AWS Config?

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

When should you use AWS Config?

- I want this **resource** to stay **configured 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 use **analyze potential security** weaknesses, you need detailed historical information.



AWS Quick Starts

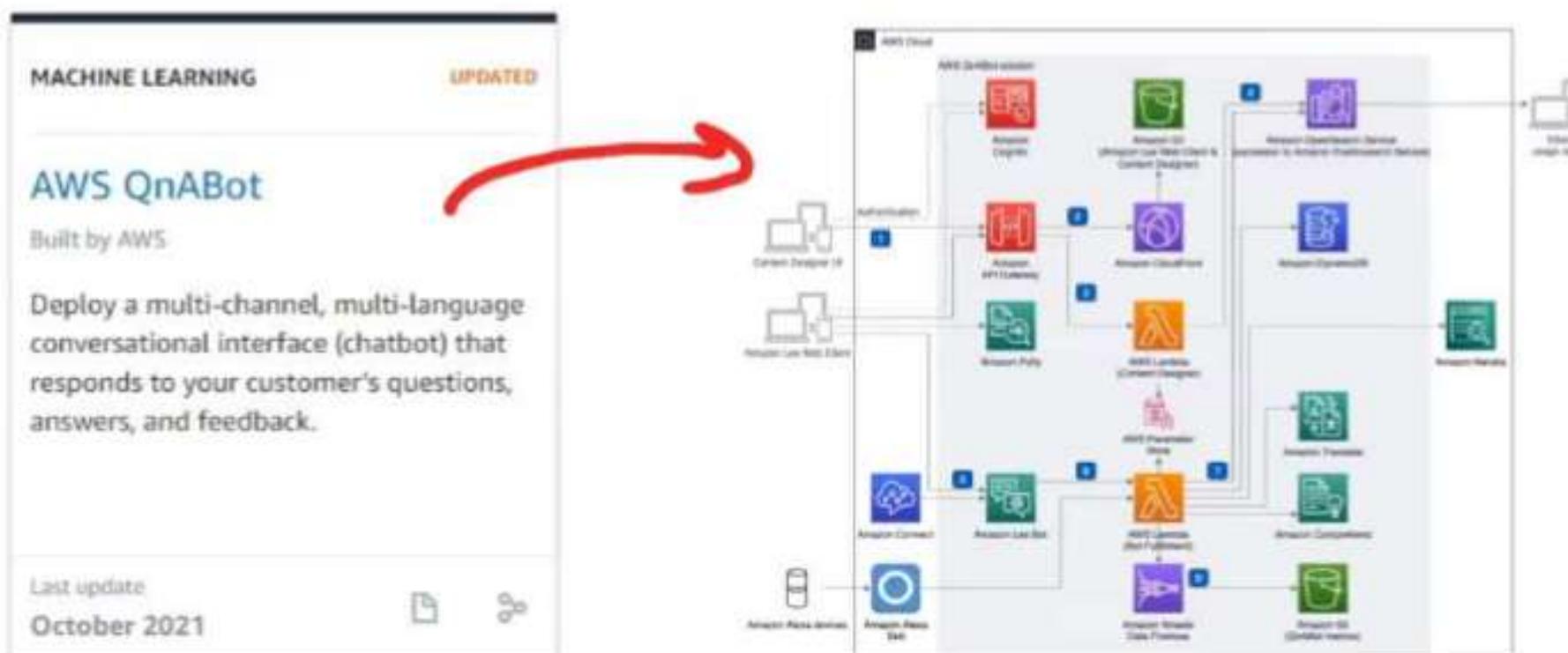


AWS Quick Starts are Prebuilt templates by AWS and AWS Partners to help deploy wide range of stacks

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 and configure the deployment
 3. A deployment guide explaining the architecture and 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 and value pair** that you can assign to AWS resources.

The screenshot shows the 'Tags (2) - optional' section of the AWS console. It displays two tags: one with 'Project' as the key and 'Enterprise' as the value, and another tag with empty 'Key' and 'Value' fields. A red arrow points from the text 'A tag is a key and value pair that you can assign to AWS resources.' to the 'Enterprise' value field.

Key	Value - optional
Project	Enterprise
Key	Value

Add tag

Tag Examples

Dept = Finance Status =

Approved Team =

Compliance Environment =

Production Project =

Enterprise Location =

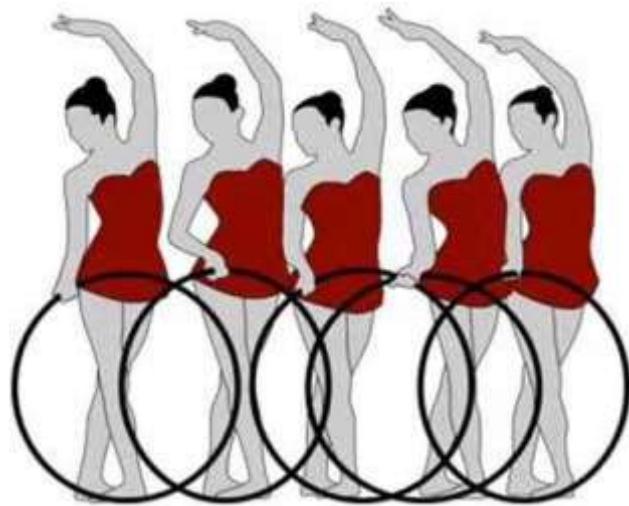
Canada

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

- **Resource management**
 - specific workloads, environments eg. Developer Environments
- **Cost management and optimization**
 - Cost tracking, Budgets, Alerts
- **Operations management**
 - Business commitments and SLA operations eg. Mission-Critical Services
- **Security**
 - Classification of data and security impact
- **Governance and regulatory compliance**
- **Automation**
- **Workload optimization**

Resource Groups

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



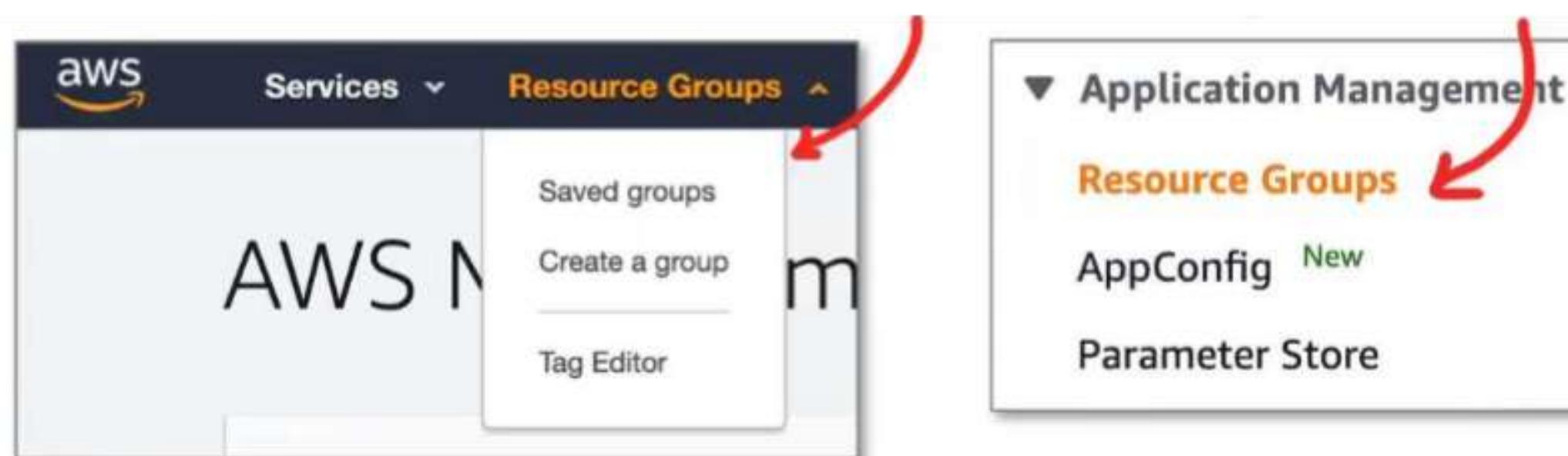
Helps you organize and consolidate information based on your project and the resources that you use.

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 appears in the **Global Console Header** and Under **Systems Manager**





Business Centric Services



Amazon Connect is a **virtual call center service**. You can create workflow 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 **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 centralized storage to share content and files.

It is similar to Microsoft SharePoint. Think of it as a shared folder where the company has ownership



Chime is **video-conference service**. It is similar to Zoom or Skype. You can screenshare, have multiple people on the call. It is secure by default and it can show you a calendar of your upcoming calls.



WorkMail is a **managed business email, contacts, and calendar service** with support for existing desktop and 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, and 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 common template, track open-rates, keep track of your reputation.



QuickSight is a **Business Intelligence (BI) service**. Connect multiple data sources and quickly visualize data in the form of graphs with little to no programming knowledge.

Provisioning Services

What is provisioning?

The allocation or creation of resources and services to a customer.

AWS Provisioning Services are responsible for setting up and 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, and 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** and **Puppet**.



CloudFormation is a **infrastructure modeling and 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 and configure your AWS compute, network, storage, and other services required to deploy a workload on AWS



AWS Marketplace - a **digital catalogue** of **thousands** of software listings from independent software vendors you can use to find, buy, test, and deploy software.



AWS Amplify is a **mobile and web-application framework**, that will provision multiple AWS services as your backend.

Provisioning Services



AWS App Runner

A fully managed service that makes it easy for developers to quickly deploy containerized web applications and APIs, at scale and with no prior infrastructure experience required



AWS Copilot

AWS Copilot is a command line interface (CLI) that enables customers to quickly launch and 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)

An Infrastructure as Code (IaC) tool. Allows you to use your favourite programming language. Generates out CloudFormation templates as the means for IaC.

AWS Elastic Beanstalk

What is Platform as a Service? (PaaS)

a PaaS allows customers to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an app

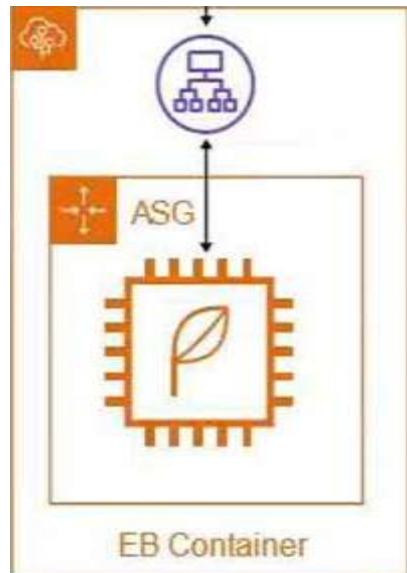


Elastic Beanstalk is a PaaS for deploying web-applications with little-to-no knowledge of the underlying infrastructure so you can focus on writing application code instead of setting up an automated deployment pipeline and DevOps tasks.

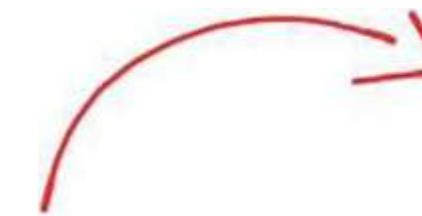
Choose a platform, upload your code and it runs with little knowledge of the infrastructure.

Not Recommended for "Production" applications

AWS is talking about enterprise, large companies.



- Elastic Beanstalk is powered by a CloudFormation template **setups** for you:
- Elastic Load Balancer
 - Autoscaling Groups
 - RDS Database
 - EC2 Instance preconfigured (or custom) platforms
 - Monitoring (CloudWatch, SNS)
 - In-Place and Blue/Green deployment methodologies
 - Security (Rotates passwords)
 - Can run **Dockerized** environments



✓ — Choose a platform --
Generic Docker
Multi-container Docker
Preconfigured
Elastic Beanstalk Packer Builder
Go
.NET (Windows/IIS)
Java
Node.js
Ruby
PHP
Python
Tomcat
Preconfigured - Docker
GlassFish Go
Python

Serverless Services

What is Serverless?

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



DynamoDB is a serverless **NoSQL key/value and 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 very large and 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 **serverless orchestration container service**. It is the same as ECS expect you pay-on-demand per running container (With ECS you have to keep a 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 and how long 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 coordinate 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*

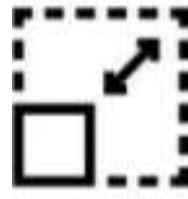
What is Serverless?

What is Serverless?

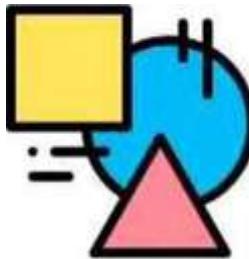
Serverless architecture generally describe fully managed cloud services.

The classification of a cloud service being serverless is not a Boolean answer (yes or no), but a answer on a scale where a cloud service has a degree of serverless.

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



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



Abstracts away the underlying infrastructure and are billed based on the execution of your business task.

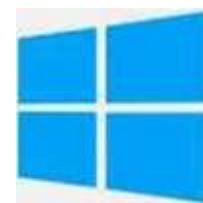


Serverless can **Scale-to-Zero** meaning when not in use the serverless resources cost nothing.

Pay-for-Value (you don't pay for idle servers).



An analogy of serverless could be similar to an energy rating labels which allows consumers to compare the energy efficiency of a product. Some services are more serverless than others.



Windows On AWS

AWS has multiple cloud services and tools to make it easy for you 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

lets you run **Microsoft Active Directory (AD) as a managed service**



AWS License Manager makes it easier to manage your software licenses from software vendors such as Microsoft.



Amazon FSx for Windows File Server is a **fully managed scalable storage** built for Windows.



AWS Software Development Kit (SDK) allows you to write code in your favorite language to interact with AWS API.

The SDK supports **.NET** a language favorite for Windows Developers



Amazon Workspaces allows you to run a virtual desktop. You can launch a **Windows 10 desktop** to a provide secure and 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 from moving large enterprise.

AWS has Amazon Partners that specialize in providing professional services for MAP.

AWS License Manager

What is Bring-Your-Own-License? (BYOL)

The process of reusing an existing software license to run vendor software on a cloud vendor's computing service. BYOL allows companies to save money since they may have purchased the license in bulk or at a time that provided a greater discount than if purchased again.

eg. **License 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 licenses from software vendors centrally across AWS and 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, and other vendors

AWS License Manager works with:

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

The screenshot shows a dropdown menu for selecting a license type. The options listed are vCPUs, Cores, Sockets, and Instances. The 'vCPUs' option is currently selected, indicated by a blue highlight. Below the dropdown is a checkbox labeled 'Enforce license limit' with a descriptive note underneath.

License type
The counting model used for the license. This may not track the terms of your agreement with your licensor for vCPUs.

vCPUs

vCPUs

Cores

Sockets

Instances

Enforce license limit
Helps prevent usage after available license types are exhausted, e.g. an instance launch requiring new license types. Not supported for RDS.

For Microsoft Windows Server and Microsoft SQL Server license you generally need to use a **Dedicated Host**

Logging Services



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

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 centralized 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. every hour take snapshot of server
- CloudWatch **Alarms** triggers notifications based on metrics
- CloudWatch **Dashboard** create visualizations 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, and if it failed to move forward.

AWS CloudTrail



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

AWS CloudTrail is used to monitor API calls and Actions made on an AWS account.

Easily identify which users and accounts made the call to AWS eg.

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

```
1  {"Records": [{}  
2  "eventVersion": "1.0",  
3  "userIdentity": {  
4  "type": "IAMUser",  
5  "principalId": "EX_PRINCIPAL_ID",  
6  "arn": "arn:aws:iam:: 123456789012: user/Worf",  
7  "accountId": "123456789012",  
8  "accessKeyId": "EXAMPLE_KEY_ID",  
9  "userName": "Worf"  
10 },  
11 "eventTime": "2014-03-24T21:11:59Z",  
12 "eventSource": "iam.amazonaws.com",  
13 "eventName": "CreateUser",  
14 "awsRegion": "us-east-1",  
15 "sourceIPAddress": "127.0.0.1",  
16 "userAgent": "aws-cli/1.3.2 Python/2.7.5 Windows/10",  
17 "requestParameters": {"userName": "LaForge"},  
18 "responseElements": {"user": {  
19 "createDate": "Mar 24, 2014 9:11:59 PM",  
20 "userName": "LaForge",  
21 "arn": "arn:aws:iam::123456789012:user/LaForge",  
22 "path": "/",  
23 "userId": "EXAMPLEUSERID"  
24 }}  
25 }]]}
```

AWS CloudTrail

CloudTrail is already logging by default and will collect logs for **last 90 days** via **Event History**

If you need more than 90 days you need to create a **Trail**

Trails are output to S3 and do not have GUI like Event History.

To analyze a Trail you'd have to use **Amazon Athena.**



CloudTrail

Event history

Your event history contains the activities taken by people, groups, or AWS services in supported services. It filters out read-only events. You can change or remove that filter, or apply other filters.

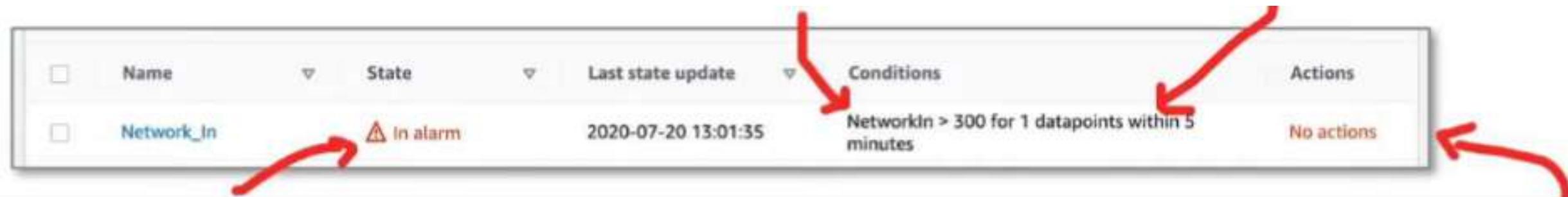
You can view the last 90 days of events. Choose an event to view more information about it. To view a complete trail and then go to your Amazon S3 bucket or CloudWatch Logs. Learn more

Can't find what you're looking for? Run advanced queries in Amazon Athena

Event time	User name	Event name
2019-09-01, 09:33:07 PM	i-014d0d0e482491e69	UpdateInstanceInformation
2019-09-01, 09:30:07 PM	i-08ece9e263d3edfb	UpdateInstanceInformation
2019-09-01, 09:28:07 PM	i-0984241e0f6a0f9ca	UpdateInstanceInformation
2019-09-01, 09:25:07 PM	i-07a9e824ebb4d5f2b	UpdateInstanceInformation
2019-09-01, 09:23:34 PM	exampro-events	CreateLogStream
2019-09-01, 09:23:07 PM	i-014d0d0e482491e69	UpdateInstanceInformation
2019-09-01, 09:20:07 PM	i-0f5f9d47f3c1cf6d	UpdateInstanceInformation
2019-09-01, 09:18:07 PM	i-08ece9e263d3edfb	UpdateInstanceInformation
2019-09-01, 09:15:07 PM	i-07a9e824ebb4d5f2b	UpdateInstanceInformation
2019-09-01, 09:13:51 PM	exampro-metrics	CreateLogStream

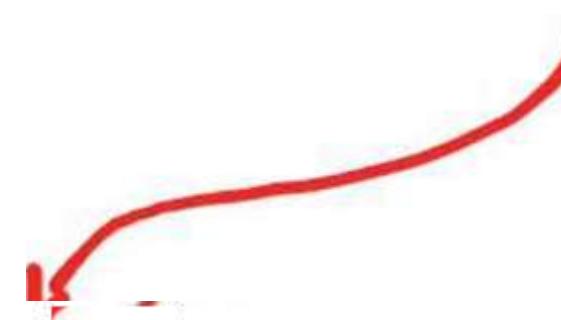
CloudWatch Alarms

A CloudWatch Alarm monitors a CloudWatch Metric based on a defined threshold.



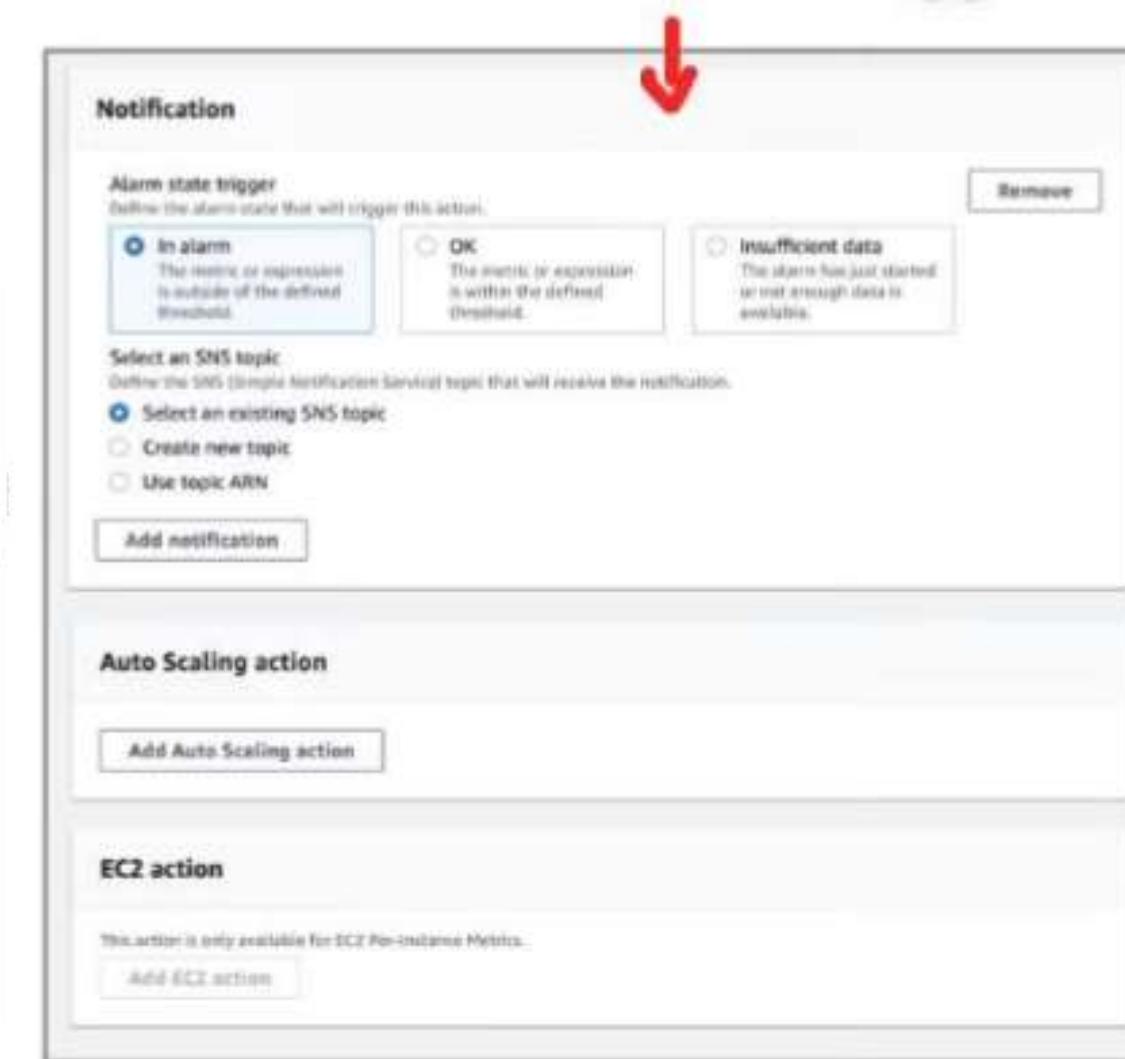
When alarm breaches (goes outside the defined threshold) than it changes state.

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



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



- Notification
- Auto Scaling Group
- EC2 Action

CloudWatch Alarms - Anatomy of an Alarm

Threshold Condition

Defines when a datapoint is breached

Threshold type:

Static
Use a value as a threshold

Whenever NetworkIn is...
Define the alarm condition.

Greater
> threshold

Greater
>= threshold

than...
Define the threshold value.

100000
Must be a number

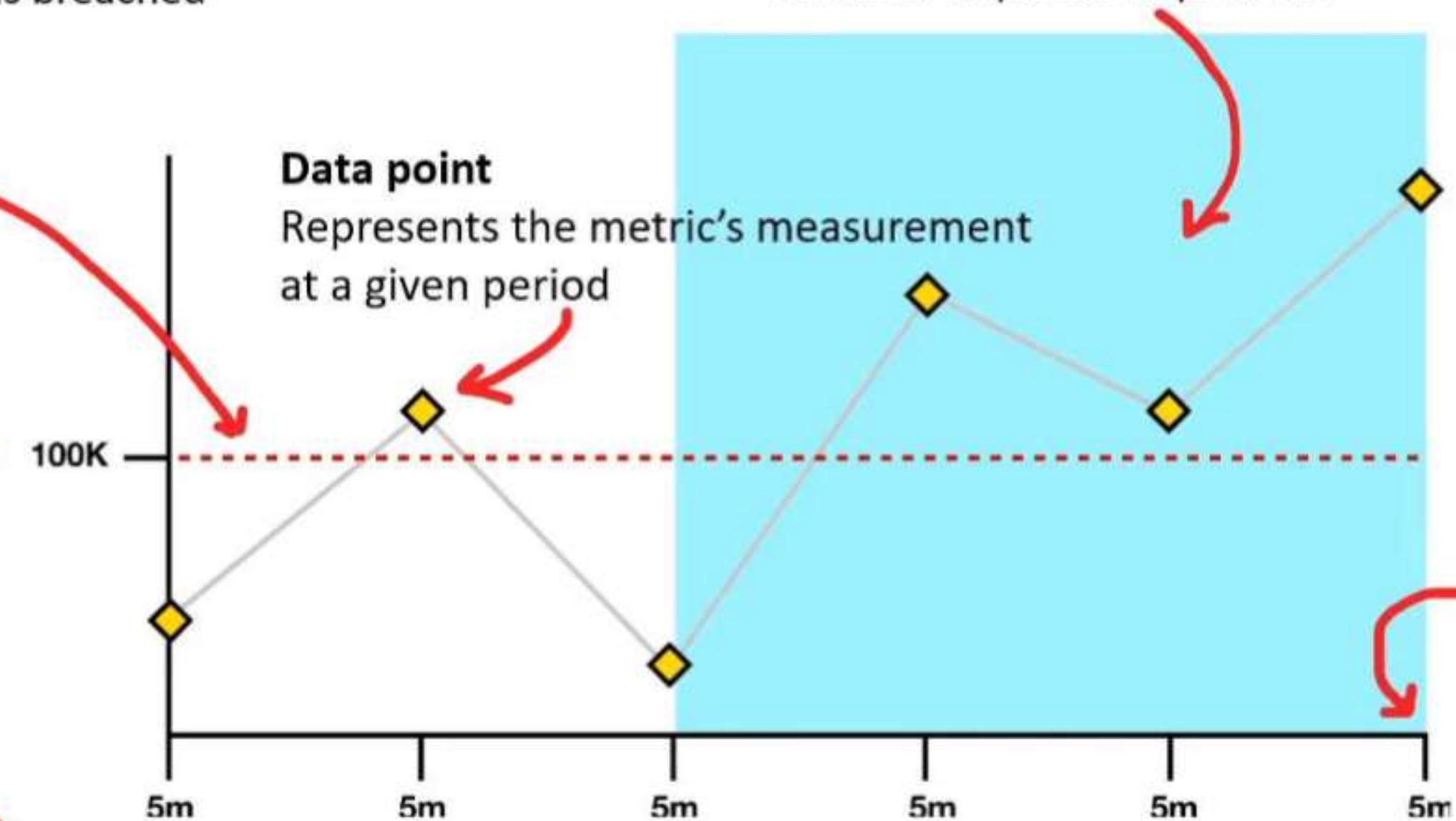
Metric
The actual data we are measuring

NetworkIn
The volume of incoming network traffic.
measured in Bytes. When using 5min monitoring divide by 300 to get Bytes/second

The screenshot shows the 'Threshold type' section with 'Static' selected. It defines the condition 'Whenever NetworkIn is Greater than 100000'. A red arrow points from the 'Metric' label to the 'NetworkIn' condition field. Another red arrow points from the 'Threshold Condition' heading to the 'Threshold type' section.

Evaluation Periods

number of previous periods



Period

How often it checks to evaluate the Alarm

Datapoints to alarm

Define the number of datapoints within the evaluation period that must be breaching to cause the alarm to go to ALARM state.

1 out of 4

The screenshot shows the 'Datapoints to alarm' section with '1 out of 4' selected. A red arrow points from the 'Datapoints to alarm' label to this field. Another red arrow points from the 'Datapoints to alarm' label to the explanatory text below.

Datapoints 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

Log Streams

A log stream represents a **sequence of events** from a **application or instance being monitored**



You can create Log Streams manually but generally this is automatically done by the service you are using

Here is a Log Group for a **Lambda function**

You can see here the Log Streams are named after the **running instance**. Lambdas frequency run on new instances so the stream streams contain timestamps

<input type="checkbox"/> Log stream	Last event time
2020/07/06/[LATEST]ebca38579fac4842b531b260d5c35e0e	7/6/2020, 7:41:24 PM
2020/07/06/[LATEST]7679ba0f37b14a3da994cd243963ca60	7/6/2020, 6:14:42 PM
2020/07/06/[LATEST]bb7edeb95cb345b48dd151a79367a5d6	7/6/2020, 3:52:56 PM
2020/07/06/[LATEST]e1544efd95a4492585b9c8d27ddaea4b	7/6/2020, 1:30:09 PM
2020/07/06/[LATEST]86a8ec4dcff746628effaeb82b41e1cc	7/6/2020, 12:28:00 PM
2020/07/06/[LATEST]a06263c73f8242e5a0e35b366b3bbdf9	7/6/2020, 10:08:43 AM

<input type="checkbox"/> Log stream	Last event time
I-0761fcbbdb19ffc8	7/6/2020, 6:56:31 PM
I-09239615bc7f3f552	7/5/2020, 9:40:42 PM
I-06c9e4fb5469e17a4	7/5/2020, 9:21:08 PM
I-0450c5ca38bdcd125	7/5/2020, 8:27:30 PM
I-01a34b0a12504edfa	7/5/2020, 12:42:24 AM
I-0e4f5ec7610f21d08	7/5/2020, 12:32:35 AM

Here is a Log Group for an **application logs running on EC2** You can see here the Log Streams are named after the **running instance's Instance ID**

<input type="checkbox"/> Log stream	Last event time
exampro-events-crawler	6/30/2019, 12:57:11 PM
exampro-waf-logs	6/26/2019, 9:00:49 AM
exampro-leads-crawler	3/24/2019, 7:57:03 PM
dynamodb-events-tracking	3/13/2019, 4:38:00 PM
cloudtrail	2/24/2019, 4:12:18 PM

Here is a Log Group for **AWS Glue**. You can see here the Log Streams are named after the **Glue Jobs**.

CloudWatch Logs — Log Events

Log Events

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

▶ 2020-07-06T20:12:18.079-04:00	START RequestId: e4b5bd10-5d88-4d7b-870c-daf793159b88 Version: \$LATEST
▶ 2020-07-06T20:12:18.082-04:00	{"records_size":1}
▶ 2020-07-06T20:12:18.093-04:00	{"failed_put_count":0}
▶ 2020-07-06T20:12:18.127-04:00	END RequestId: e4b5bd10-5d88-4d7b-870c-daf793159b88
▶ 2020-07-06T20:12:18.127-04:00	REPORT RequestId: e4b5bd10-5d88-4d7b-870c-daf793159b88 Duration: 45.32 ms Billed Duration: 100 ms Memory Size: 128

You can use filter events to filter out logs based on simple or pattern matching syntax:

The screenshot shows the CloudWatch Logs interface. At the top, there is a search bar with the text 'D' and a clear button. Below the search bar are time range buttons: 'Clear', '1m', '30m', '1h', '12h', 'custom', and a calendar icon. To the right of these buttons is a gear icon. On the left side, there is a sidebar with a 'Log events' section and a 'Filter events' input field. A red arrow points from the 'Filter events' input field to the search bar. The main area displays a list of log events with columns for 'Timestamp' and 'Message'. Each log entry starts with a '▶' symbol. The first few entries are: '2020-07-05T21:39:26.857-04:00 D, [2020-07-06T01:39:26.596187 #3979] DEBUG -- : [1m [35ms (0.4ms)]', '2020-07-05T21:39:26.857-04:00 D, [2020-07-06T01:39:26.614381 #3979] DEBUG -- : [1m [35ms (1.5ms)]', '2020-07-05T21:39:26.857-04:00 D, [2020-07-06T01:39:26.621670 #3979] DEBUG -- : [1m [36ms ActiveRec...', and '2020-07-05T21:39:26.857-04:00 D, [2020-07-06T01:39:26.626819 #3979] DEBUG -- : [1m [35ms (0.4ms)]'. The last entry is partially cut off.

Timestamp	Message
2020-07-05T21:39:26.857-04:00	D, [2020-07-06T01:39:26.596187 #3979] DEBUG -- : [1m [35ms (0.4ms)]
2020-07-05T21:39:26.857-04:00	D, [2020-07-06T01:39:26.614381 #3979] DEBUG -- : [1m [35ms (1.5ms)]
2020-07-05T21:39:26.857-04:00	D, [2020-07-06T01:39:26.621670 #3979] DEBUG -- : [1m [36ms ActiveRec...
2020-07-05T21:39:26.857-04:00	D, [2020-07-06T01:39:26.626819 #3979] DEBUG -- : [1m [35ms (0.4ms)]
2020-07-05T21:39:26.857-04:00	D, [2020-07-06T01:39:26.627990 #3979] DEBUG -- : [1m [35ms (0.4ms)]

CloudWatch Logs - Log Insights

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

- more robust filtering than using the simple Filter events in a Log Stream
- Less burdensome than having to export logs to S3 and analyze them via Athena.

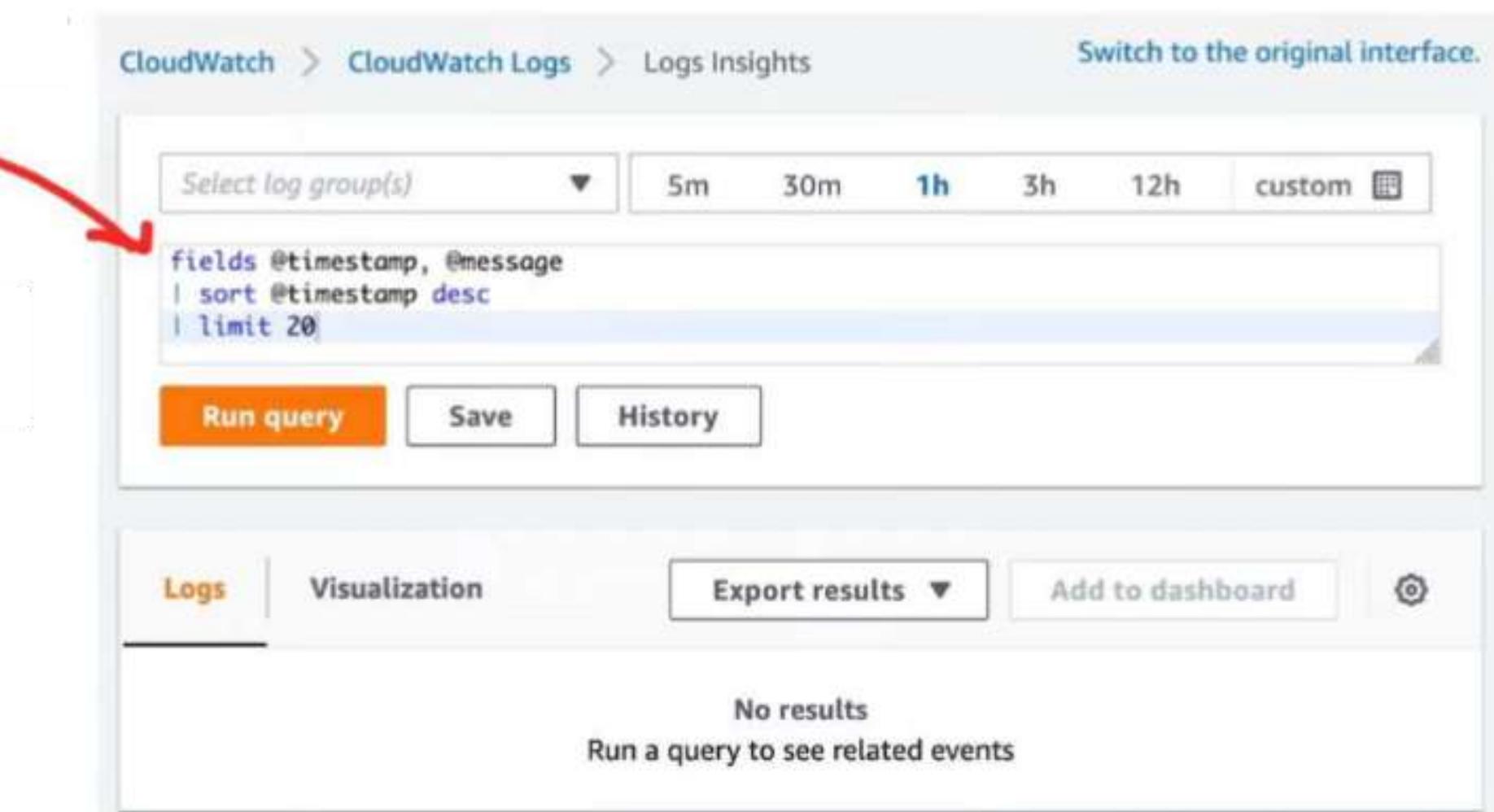
CloudWatch Logs Insights supports all types of logs.

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**

```
filter action="REJECT"
| stats count(*) as numRejections by srcAddr
| sort numRejections desc
| limit 20
```



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

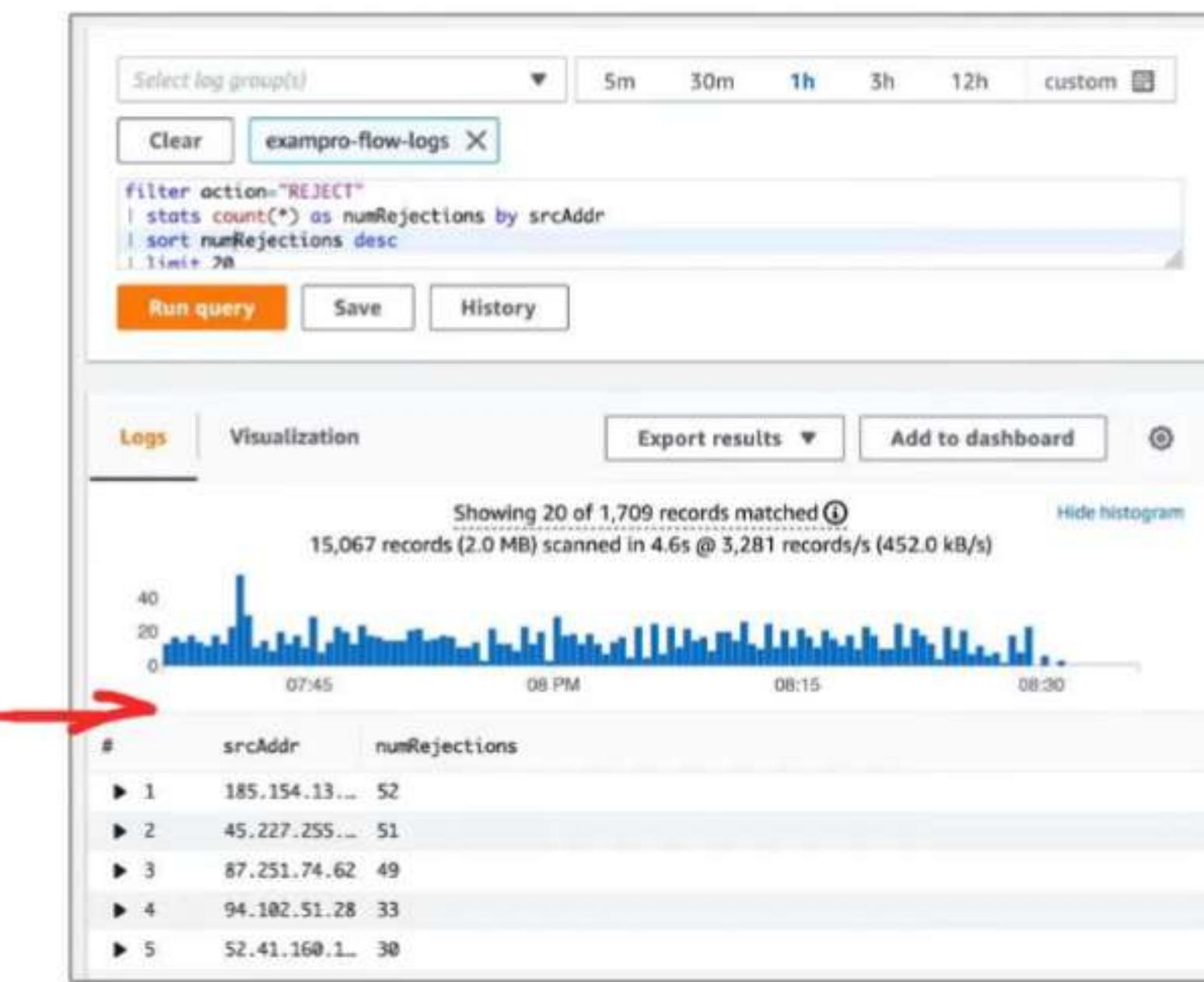
CloudWatch Logs - Log Insights

The screenshot shows the AWS CloudWatch Log Insights interface. On the left, there's a sidebar with 'Saved queries' and a search bar. Below it is a 'Create query' button. The main area is titled 'Sample queries' and includes sections for 'Lambda' and 'VPC Flow Logs'. Under 'VPC Flow Logs', there are four items: 'Average, min, and max byte transfers by source and destination IP addresses', 'IP addresses using UDP transfer protocol', 'Top 10 byte transfers by source and destination IP addresses', and 'Top 20 source IP addresses with highest number of rejected requests'. A red arrow points from the 'Top 20 source IP addresses...' item to the bottom section of the screenshot. The bottom section shows the query code:

```
filter action="REJECT"
| stats count(*) as numRejections by srcAddr
| sort numRejections desc
| limit 20
```

 and an 'Apply' button.

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



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

CloudWatch Metrics

A CloudWatch Metric represents a **time-ordered set of data points**

Its a **variable** that is **monitored over time**.

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

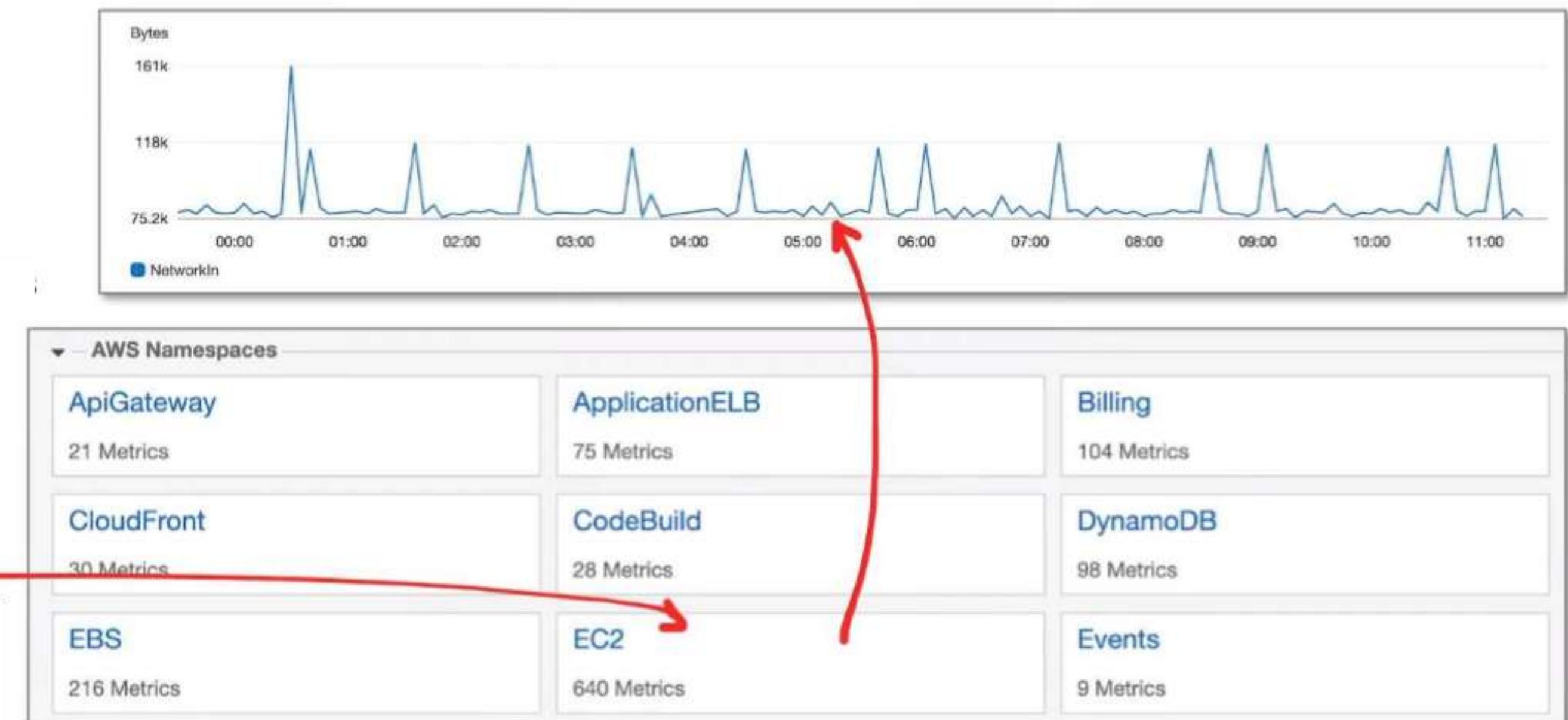


EC2 Per-Instance Metrics

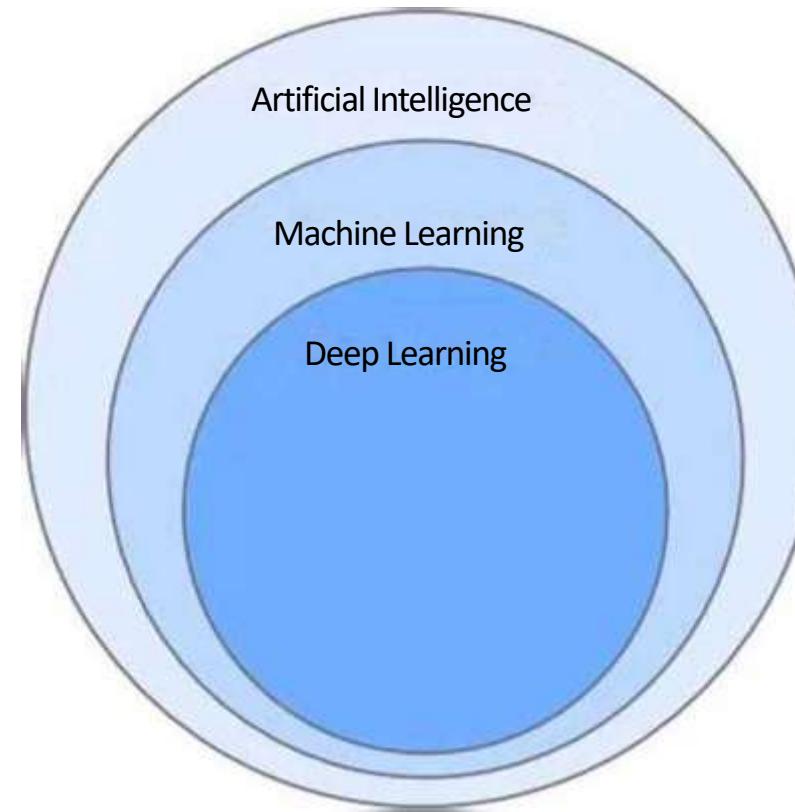
CPUUtilization
DiskReadOps
DiskWriteOps
DiskReadBytes
DiskWriteBytes

NetworkIn

NetworkOut
NetworkPacketsIn
NetworkPacketsOut



Machine Learning and AI Services



What is Artificial Intelligence (AI)?

Machines that perform jobs that mimic human behavior

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, and 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 **data-labeling 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 uses machine Learning to make a prediction is not confident it has the right answer queue up the predication for human review.

Machine Learning and AI Services



Amazon CodeGuru is a **machine-learning code analysis service**. CodeGuru performs code-reviews and 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 and text chatbots**



Amazon Personalize 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 and an audio file spoken by synthesized voice is generated.



Amazon Rekognition is **image and video recognition service**. Analyze images and videos to detect and label objects, people, celebrities.



Amazon Transcribe is a **speech-to-text service**. Upload your audio file and it is converted



Amazon Textract and OCR (extract text from scanned documents) service. When you have paper forms and you want to digitally extract the data.



Amazon Translate neural machine learning translation service. Uses deep learning models to deliver more accurate and natural sounding translations.



Amazon Comprehend is a **Natural Language Processor (NLP) service**. Find relationships between text to produce insights. Looks at data such as Customer emails, support tickets, social media and makes predictions.

Machine Learning and AI Services



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** and interfaces such as TensorFlow, PyTorch, Apache MXNet, Chainer, Gluon, Horovod, and Keras



AWS Deep Learning Containers Docker images instances pre-install with popular deep learning frameworks and interfaces such as TensorFlow, PyTorch, and Apache MXNet.



AWS DeepComposer is 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 a service**, identify potentially fraudulent online activities such as online payment fraud and the creation of fake accounts.



Amazon Kendra enterprise machine learning search engine service. Uses natural language to suggest answers to question instead of just simple keyword matching

Big Data and 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 and process** using traditional database and software techniques.



Amazon Athena is a **serverless interactive query service**. It can take a bunch of CSV or JSON files in a S3 Bucket and 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 add search to your website*



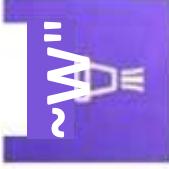
Amazon Elasticsearch Service (ES) is a **managed Elasticsearch cluster**. Elasticsearch is a open-source full-text search engine. It is more robust than CloudSearch but requires more server and operational maintaince.



Amazon Elastic MapReduce (EMR) is for data processing and analysis. Its 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, click streams, ingesting data from a fleet of IOT Devices



Kinesis Firehose is serverless and a simpler version of Data Streams, You pay-on-demand based on how much data is consumed through the stream and 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 and analysis on emerging data.



Amazon Kinesis Video Streams allows you to analyze or apply processing on real-time streaming video.

Big Data and Analytics Services



Managed Kafka Service (MSK) a fully managed Apache Kafka service. Kafka is an open-source platform for building real-time streaming data pipelines and 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 and a large amount of data and get that data back very fast.

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



Amazon QuickSight is **business intelligence (BI) dashboard**. You can use it to create business dashboards to power business decisions. It requires little to no programming knowledge and connect and ingest to many different types of databases



AWS Data Pipeline automates the **movement of data**. You can reliably move data between compute and storage services.



AWS Glue is an **Extract, Transform, Load (ETL) service**. Moving data from one location to another and where you need to perform transformations before the final destination. Similar to Database Migration Service (DMS) but more robust

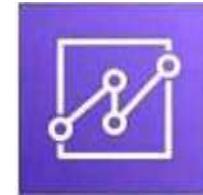


AWS Lake Formation is as a **centralized, curated, and secured repository that stores all your data**.

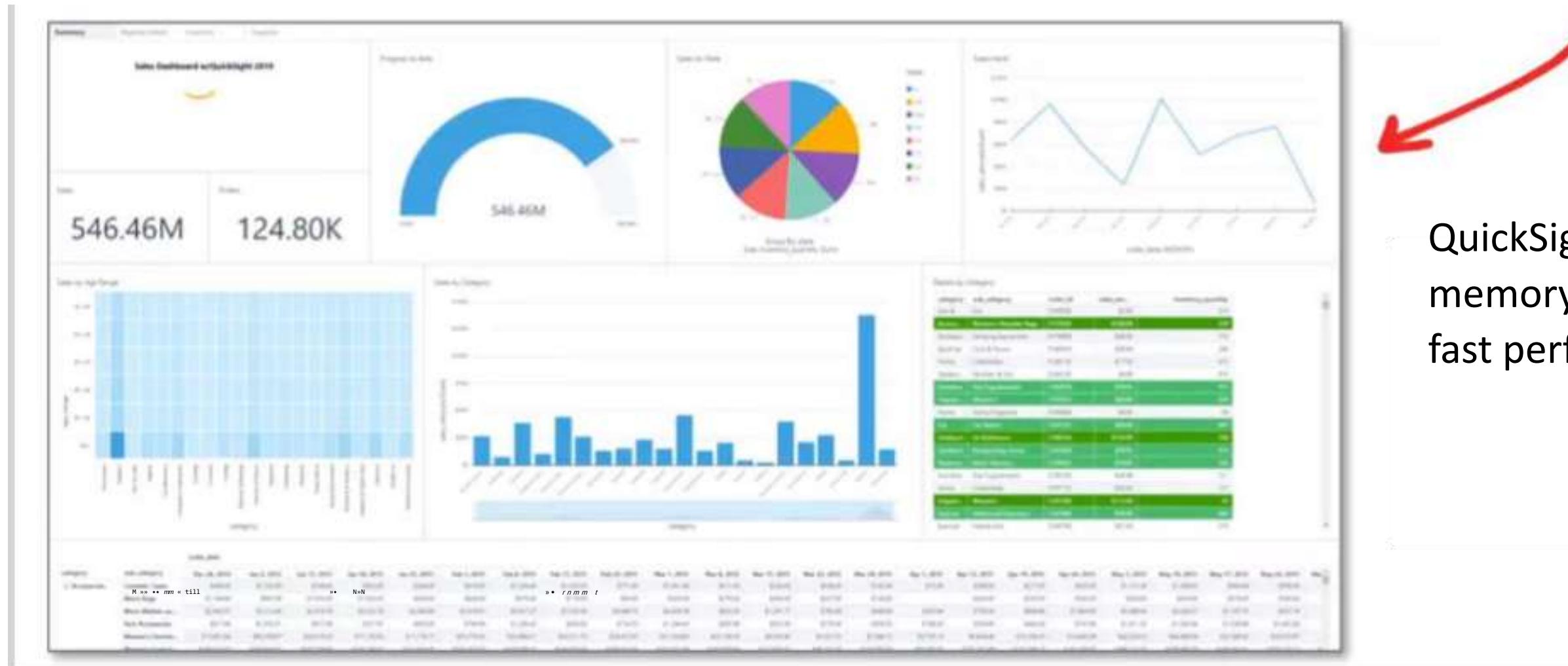
A **data lake** is a storage repository that holds a vast amount of raw **data** in its native format until it is needed.

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 and 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 visualize 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

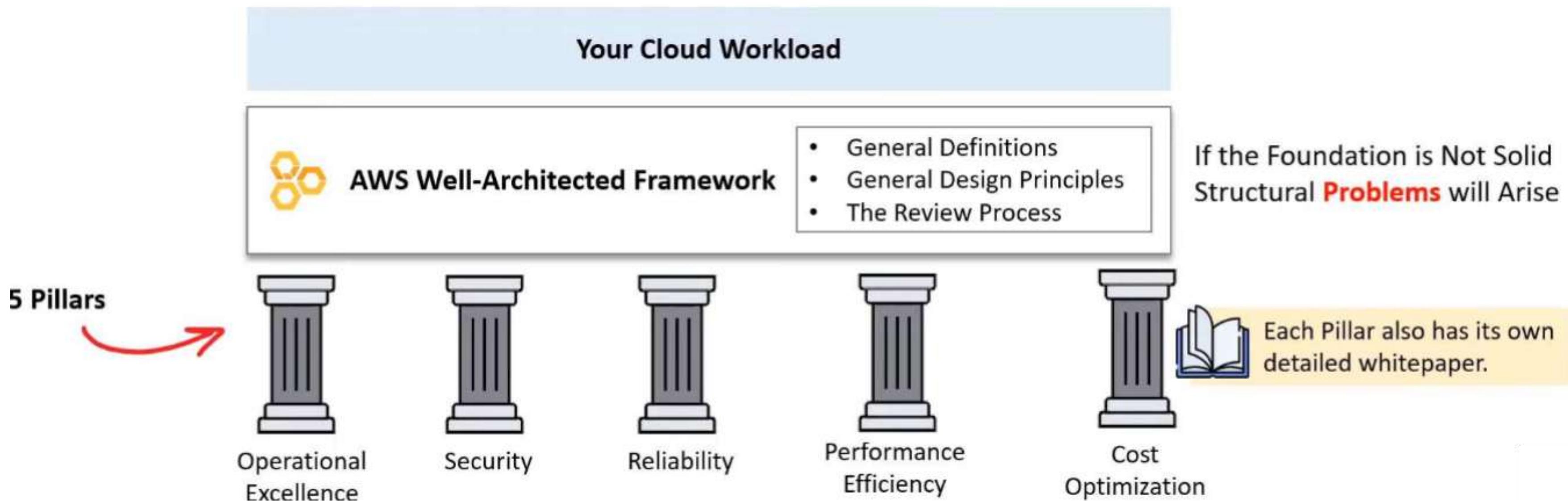
Amazon QuickSight ML Insights - Detect Anomalies, Perform accurate forecasting, Generate Natural Language Narratives. **Amazon QuickSight Q** - Ask question using natural language, on all your data, and receive answers in seconds.

AWS Well-Architected Framework

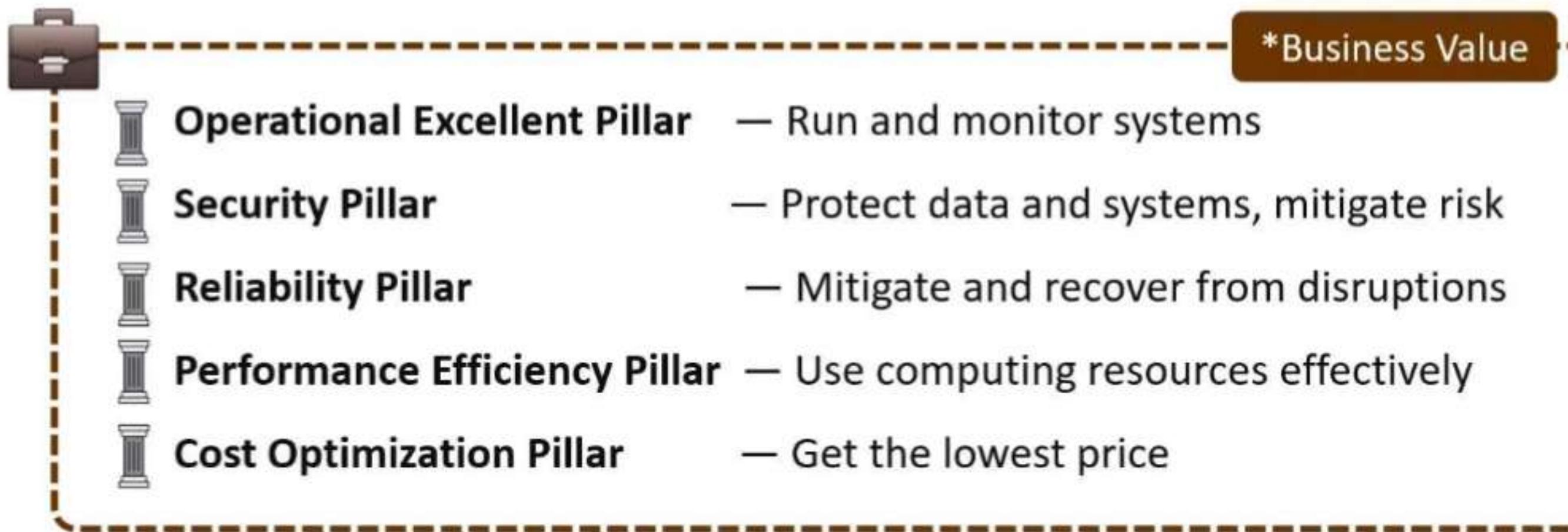
The AWS Well-Architected Framework is a Whitepaper created by AWS to help customers build using best-practices defined by AWS.

aws.amazon.com/architecture/well-architected

The framework is divided into 5 sections called pillars which address different aspects or "lenses" that can be applied to a cloud workload.



AWS Well-Architected - General Definitions



*Trade-Off Pillars Based on Business Context

General Definitions

Component — Code, Configuration and 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

AWS Well-Architected - Architecture

The AWS Well-Architected Framework is designed around a different kind of team structure.

Enterprises generally have centralized teams with specific roles where AWS has distributed teams with flexible roles. Distributed teams can come with new risks, AWS mitigates these with Practices, Mechanisms and Leadership Principles



On-Premise Enterprise

VS

Amazon Web Services

Centralized team consisting of:

- Technical Architect (infrastructure)
- Solution Architect (software)
- Data Architect
- Networking Architect
- Security Architect

Managed by either **TOGAF** or **Zachman Framework**

Distributed teams consisting of:

- Practices
 - Team Experts (Raise the Bar)
- Mechanisms
 - Automated Checks for Standards
- *Amazon Leadership Principle

Supported by a virtual community of **SMEs, Principle Engineers**
eg. lunchtime talks - recycled into onboarding material

Amazon Leadership Principles



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

1. Customer Obsession
2. Ownership
3. Invent and Simplify
4. Are Right, A Lot
5. Learn and Be Curious
6. Hire and 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 and Commit
14. Deliver Results
15. Strive to be Earth's Best Employer
16. Success and Scale Bring Broad Responsibility



You can read about all 16 principles here
<https://www.amazon.jobs/en/principles>

AWS Well-Architected - General Design Principles

Stop guessing your capacity needs

eg. Cloud computing you use as little or much based **on demand**.

Test systems at production scale

eg. Clone production env to testing, Tear down testing not in use to save money.

Automate to make architectural experimentation easier

eg. Using CloudFormation with ChangeSets, StackUpdate and 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

AWS Well-Architected - Anatomy of a Pillar



Performance
Efficiency

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**
 - overview of the best practice categories
- **Best Practices**
 - detailed information about each best practice with AWS Services
- **Resources**
 - Additional documentation, whitepapers and videos to implement this pillar

▼ Performance Efficiency

Design Principles

Definition

► Best Practices

Resources



AWS Well-Architected - Design Principles



Operational Excellence Design Principles

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 and 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 servers to test recovery

Learn from all operational failures

share lessons learned in a knowledge base for operational events and failures across your entire organization

AWS Well-Architected - Design Principles



Security Design Principles

Implement a strong identity foundation

Implement Principle of Least Privilege (PoLP). Use Centralized identity. Avoid Long-lived credentials

Enable traceability

Monitor alert and audit actions and changes to your environment in real-time Integrate log and metric collection and automate investigation and remediation

Apply security at all layers

Take Defense 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 and at rest

Keep people away from data Prepare for security events

Incident management systems and investigation policy and processes. Tools to detect, investigate and recover from incidences

AWS Well-Architected - Design Principles



Reliability Design Principles

Automatically recover from failure

Monitor Key Performance Indicators (KPIs) and trigger automation when threshold is breached.

Test recovery procedures

Test how your workload fails, and 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

In on-premise it takes a lot of guess work 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 and review infrastructure

AWS Well-Architected - Design Principles



Performance Efficiency Design Principles

Democratize advanced technologies:

Focus on product development rather than procurement, provisioning and management of services.

Take advantage of advanced technology specialized and optimized 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 and a better experience for your customers at minimal cost.

Use serverless architectures:

Serverless architectures remove the need for you to run and maintain physical servers for traditional compute activities.

Removes the operational burden of managing physical servers, and can lower transactional costs because managed services operate at cloud scale.

Experiment more often:

With virtual and 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 and 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.

AWS Well-Architected - Design Principles



Cost Optimization Design Principles

Implement Cloud Financial Management:

Dedicate time and resources to build capability Cloud Financial Management and Cost Optimization tooling.

Adopt a consumption model

Pay only for the computing resources that you require and increase or decrease usage depending on business requirements

Measure overall efficiency

Measure the business output of the workload and the costs associated with delivering it.

Use this measure to know the gains you make from increasing output and reducing costs.

Stop spending money on undifferentiated heavy lifting

AWS does the heavy lifting of data center operations like racking, stacking, and powering servers.

It also removes the operational burden of managing operating systems and applications with managed services.

This allows you to focus on your customers and business projects rather than on IT infrastructure.

Analyze and attribute expenditure

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

AWS Well-Architected 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.

The screenshot shows the AWS Well-Architected Framework review interface. On the left, a sidebar lists six operational priorities (OPS 1-6). The main content area displays the details for OPS 1: "How do you determine what your priorities are?". It includes a question about shared goals for resources and a list of actions to select from. A red arrow points to the "Mark best practice(s) that don't apply to this workload" button at the bottom. To the right, a "Helpful resources" panel provides links to AWS Support and Cloud Compliance, and contains sections on evaluating customer needs, internal customer needs, governance requirements, and compliance requirements.

Operational Excellence
6/11

OPS 1. How do you determine what your priorities are?

OPS 2. How do you structure your organization to support your business outcomes?

OPS 3. How does your organizational culture support your business outcomes?

OPS 4. How do you design your workload so that you can understand its state?

OPS 5. How do you reduce defects, ease remediation, and improve flow into production?

OPS 6. How do you mitigate deployment risks?

Well-Architected Tool > Workloads > ExamPro > AWS Well-Architected Framework > Review workload

AWS Well-Architected Framework

Add a link to your architectural design.

OPS 1. How do you determine what your priorities are?

Everyone needs to understand their part in enabling business success. Have shared goals in order to set priorities for resources. This will maximize the benefits of your efforts.

Question does not apply to this workload

Select from the following:

Evaluate external customer needs

Evaluate internal customer needs

Evaluate governance requirements

Evaluate compliance requirements

Evaluate threat landscape

Evaluate tradeoffs

Manage benefits and risks

None of these

Mark best practice(s) that don't apply to this workload

Helpful resources

AWS Support

AWS Cloud Compliance

Evaluate external customer needs

Involve key stakeholders, including business, development, and operations teams, to determine where to focus efforts on external customer needs. This will ensure that you have a thorough understanding of the **operations** support that is required to achieve your desired business outcomes.

Evaluate internal customer needs

Involve key stakeholders, including business, development, and operations teams, when determining where to focus efforts on internal customer needs. This will ensure that you have a thorough understanding of the **operations** support that is required to achieve business outcomes.

Evaluate governance requirements

Ensure that you are aware of guidelines or obligations defined by your organization that may mandate or emphasize specific focus. Evaluate internal factors, such as organization policy, standards, and requirements. Validate that you have mechanisms to identify changes to governance. If no governance requirements are identified, ensure that you have applied due diligence to this determination.

Evaluate compliance requirements

Its essentially **a checklist**, with nearby references to help you assemble a report to share with executives and key stake-holders

AWS Architecture Center

The AWS Architecture Center is a web-portal that contains **best practices** and **reference architectures** for a variety of different workloads.
aws.amazon.com/architecture

Security, Identity, & Compliance
Learn how to meet your security and compliance goals using AWS infrastructure and services.

Identity & Access Management
Manage access to AWS services and resources.

Detection
Learn how to detect suspicious activity in your AWS account.

Infrastructure Protection
Monitor and control your network infrastructure.

Data Protection
Operate the security services that protect your data.

Compliance
Implement compliance controls with AWS.

Incident Response
Learn how to automate incident response and recovery.

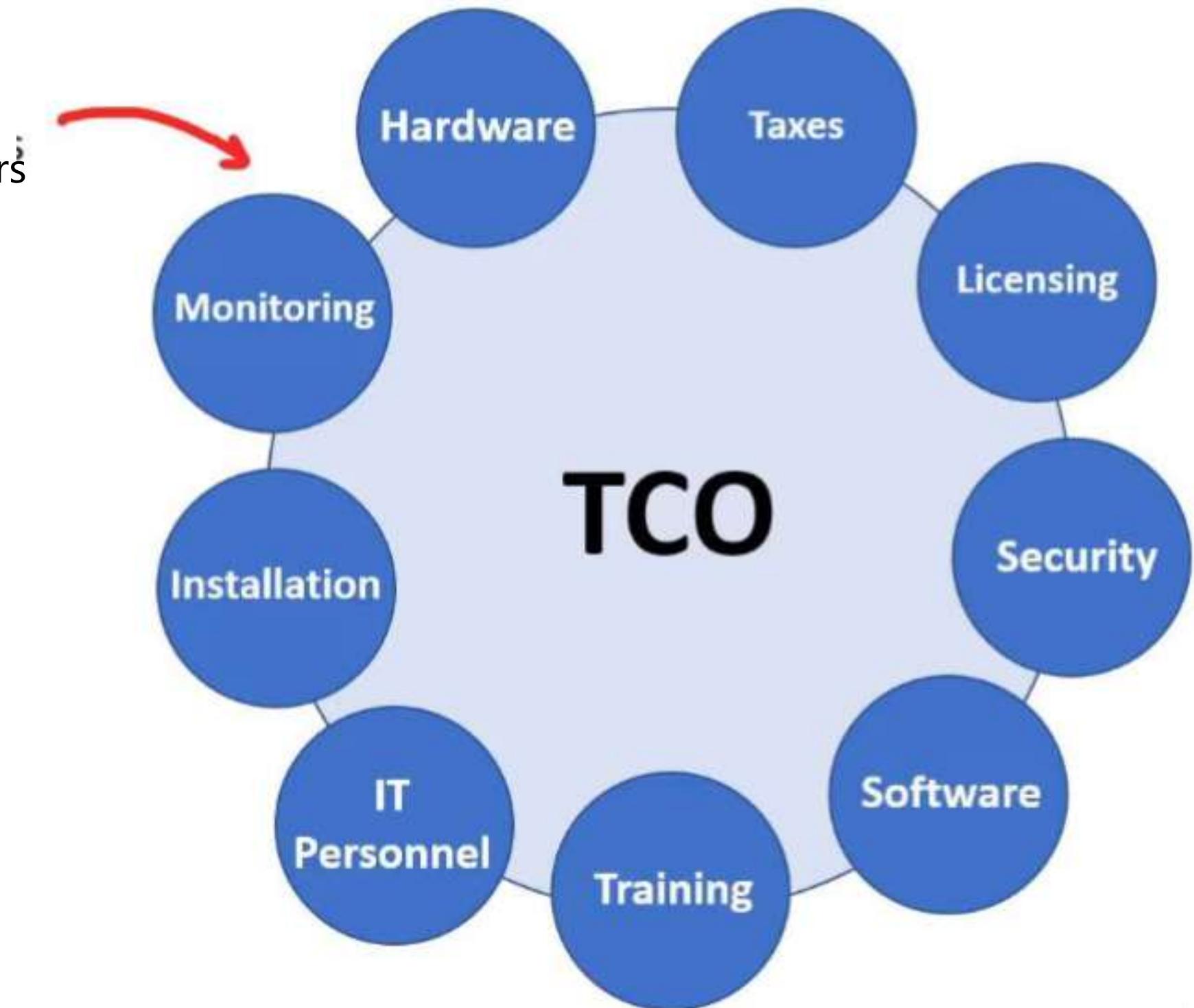


Total Cost of Ownership (TCO)

What is the Total Cost of Ownership? (TCO)?

TCO is a **financial estimate** intended to help buyers and owners determine the direct and 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.

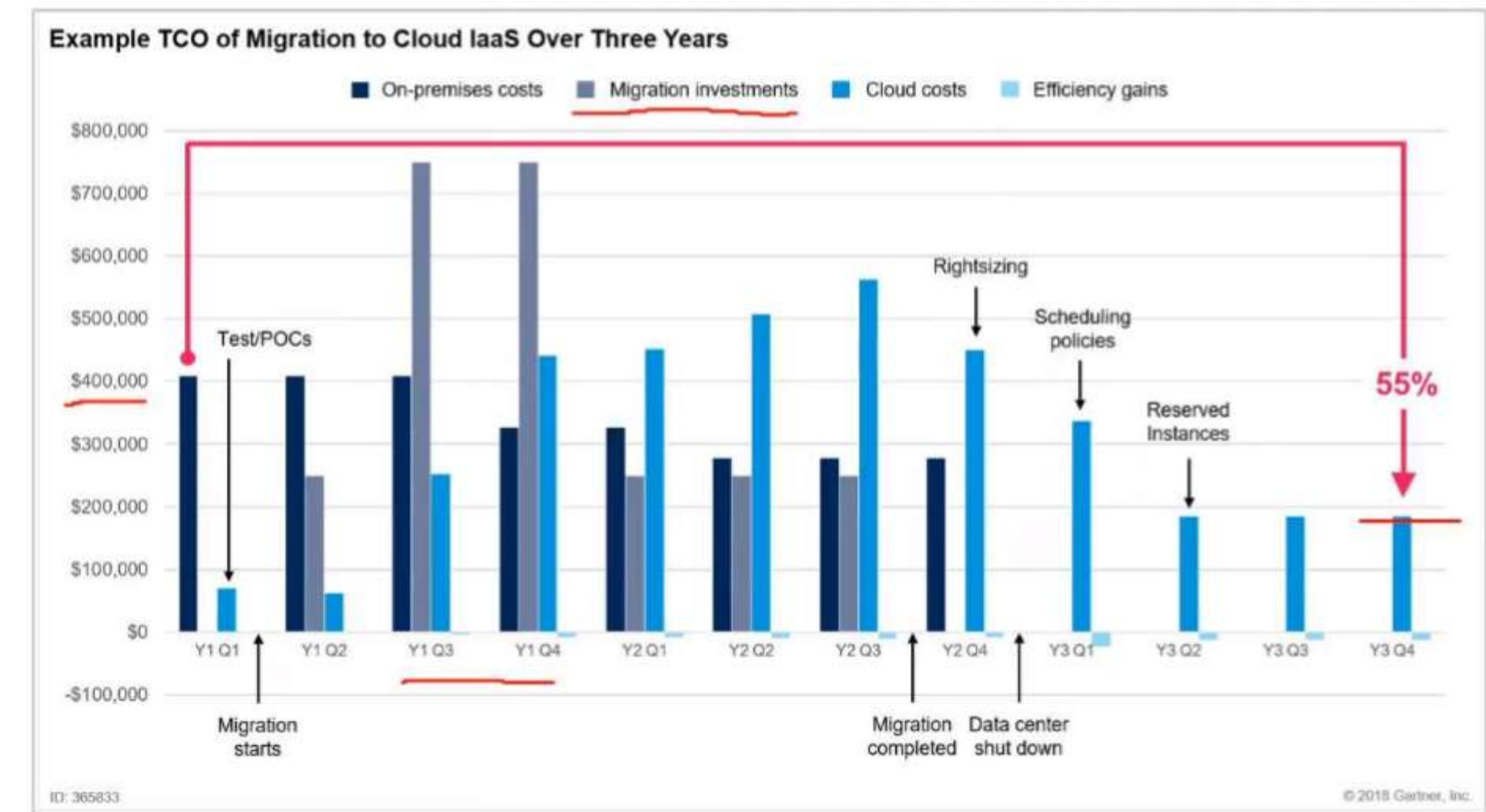


Total Cost of Ownership (TCO)

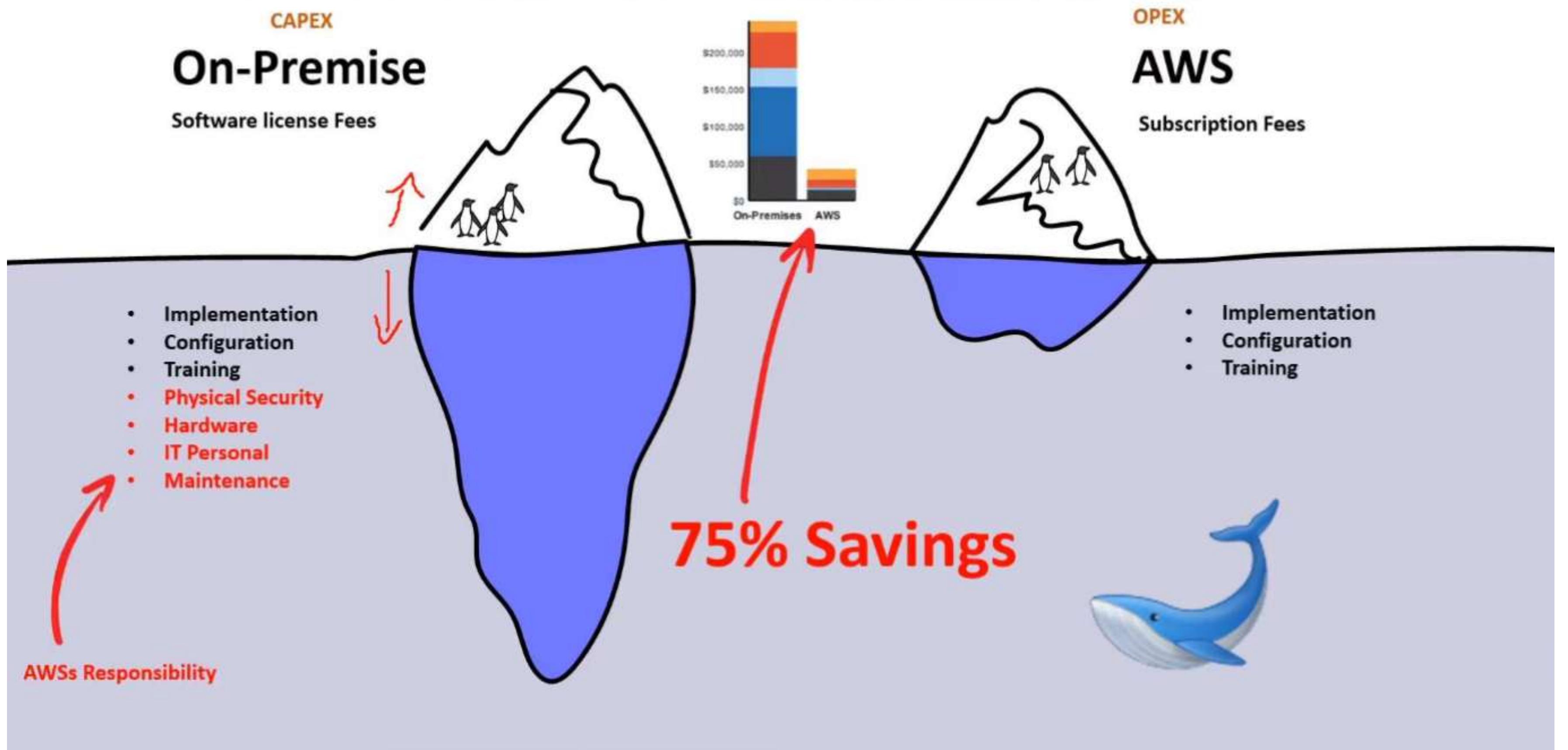
According to research stated by Garter:

"cloud services can initially be more expensive than running on-premises data centers. [However, it also proves that] cloud services can become cost-effective over time if organizations learn to use and operate them more efficiently"

Example of a 2,500
Virtual Machines (VMs)
moved to Amazon EC2



Total Cost of Ownership (TCO)



Capital vs Operational Expenditure

Capital Expenditure (CAPEX)

Spending money upfront on **physical infrastructure**

Deducting that expense from your tax bill over time.

Server Costs (computers)

Storage Costs (hard drives)

Network Costs (Routers, Cables, Switches)

Backup and Archive Costs Disaster Recovery Costs

Datacenter Costs (Rent, Cooling, Physical Security)

Technical Personal

With Capital Expenses **you have to guess upfront** what you plan to spend

Operational Expenditure (OPEX)

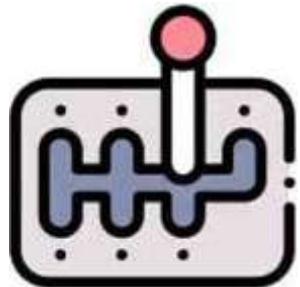
The costs associated with an on-premises datacenter that has shifted the cost to the service provider. The customer only has to be concerned with **non-physical costs**.

- Leasing Software and Customizing features
- Training Employees in Cloud Services
- Paying for Cloud Support
- Billing based on cloud metrics eg.
 - compute usage
 - storage usage

With Operation Expenses you can try a product or service **without investing in equipment**

Does Cloud Make IT Personnel Redundant?

A company is considering migrating their 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. Does cloud make IT Personnel redundant?



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 and 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 a various AWS services.

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

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



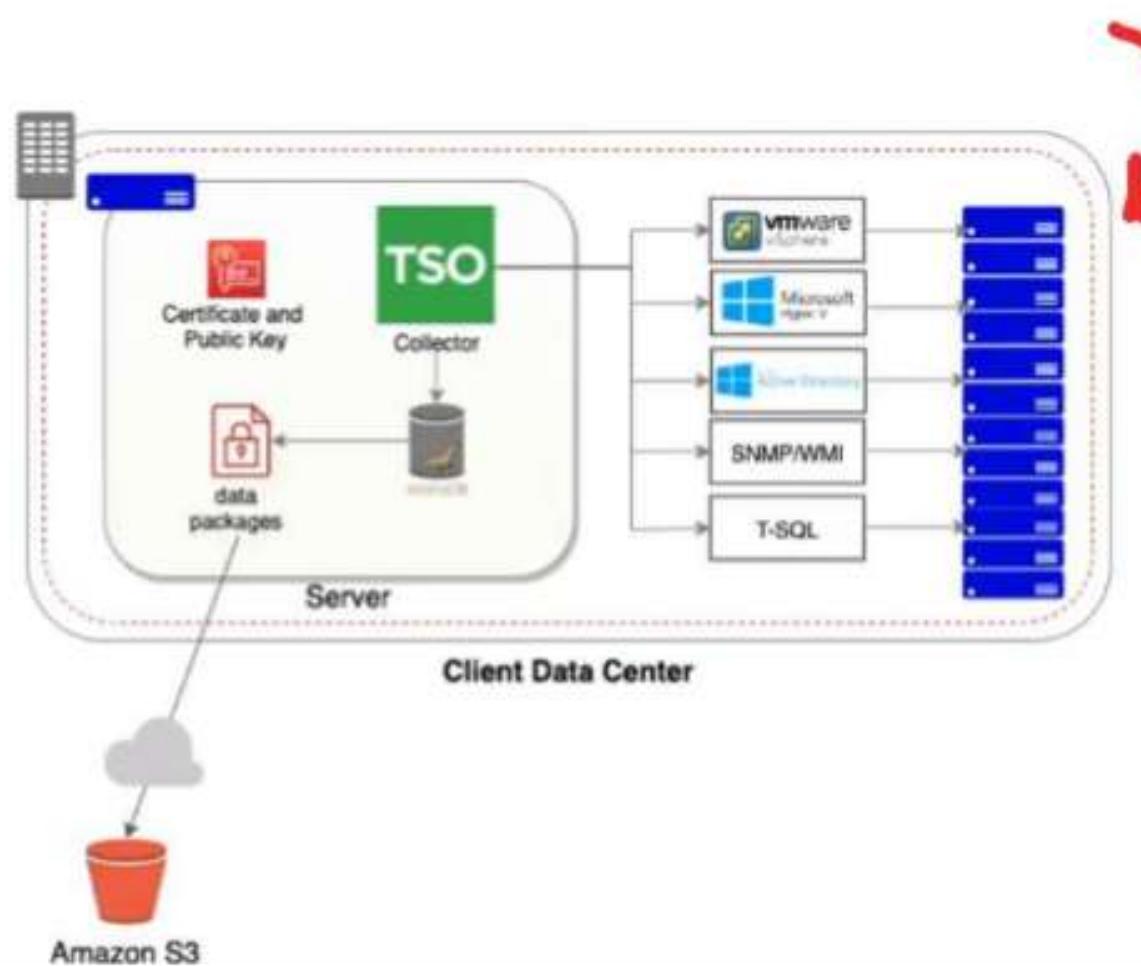
You can export your final estimate to a CSV.

The screenshot shows the AWS Pricing Calculator interface. At the top, there's a navigation bar with links like 'calculator.aws', 'AWS Pricing Calculator', 'My Estimate', 'Add service', 'Add support', 'Add group', 'Clear estimate', 'Export estimate', and 'Share'. A red arrow points from the text above to the 'Share' button. Below the navigation, there's a summary section titled 'Estimate summary' with three boxes: 'Upfront cost: 0.00 USD', 'Monthly cost: 4,751.53 USD', and 'Total 12 months cost: 57,018.36 USD'. To the right, there's a 'Getting Started with AWS' sidebar with 'Contact Us' and 'Sign in to the Console' buttons. The main content area shows a 'Support' section for a 'Business support plan' (Region: All Regions) which supports 24/7 phone, chat, and email access to Cloud Support Engineers for unlimited contacts with a response time of less than 1 hour. Below that is a 'Services' section listing 'Amazon RDS for PostgreSQL' (Region: US East (Ohio)) and 'RDS for PostgreSQL'. Both services have 'Edit' and 'Action' buttons. The total monthly cost for the services is listed as 2,139.96 USD.

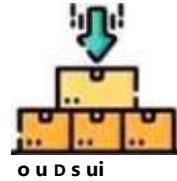
Migration Evaluator

AWS Migration Evaluator (formally known as TSO Logic) is an **estimate tool** used to determine an organization 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



Prepare your Virtual
Image for Upload



Upload your Virtual
Image to S3



Use the AWS CLI to Import your Image
It will generate an Amazon Machine Image (AMI)

```
aws ec2 import-image \  
--disk-containers Format=ova,UserBucket="{S3Bucket=my-vm,S3Key=vm.ova}
```

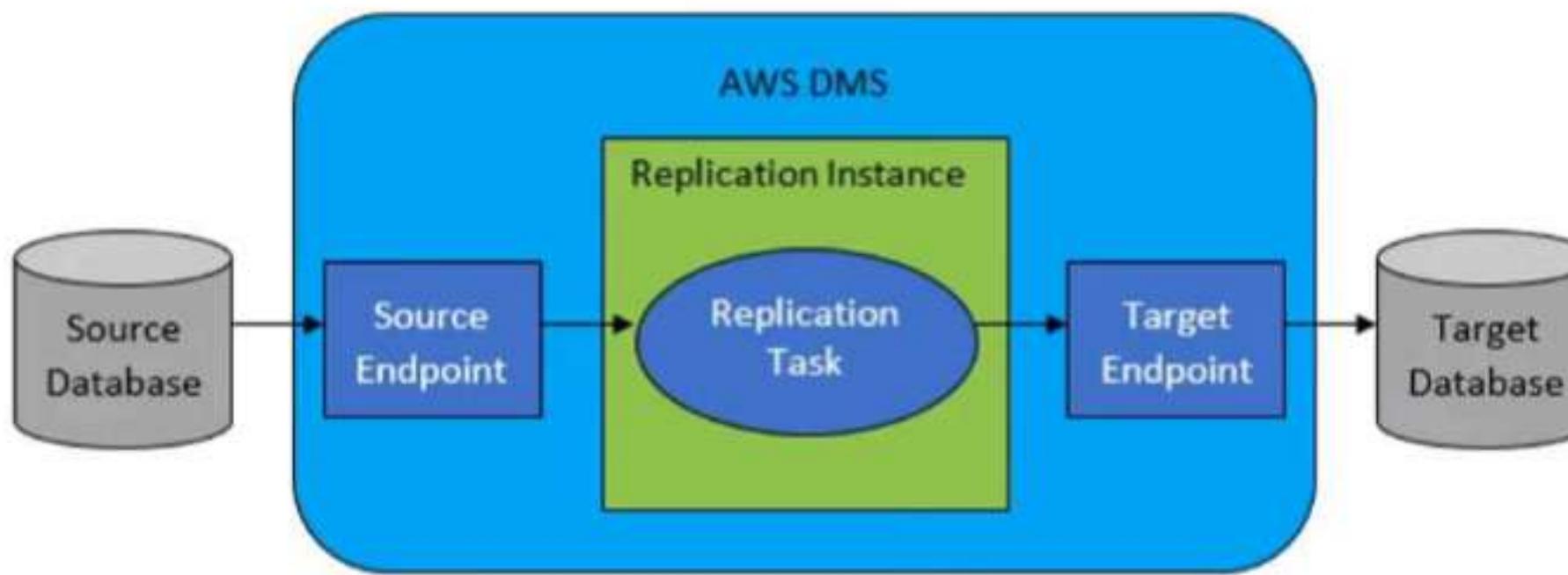
Database Migration Service (DMS)



AWS Database Migration Service (DMS) allows you to quickly and 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



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 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 combination of sources and targets are possible.

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 organizes guidance into **six focus areas**.



1 Business Perspective e.g. Business Managers, Finance Managers, Budget Owners, and Strategy Stakeholders.

How to update the staff skills and organizational processes to optimize 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 and organizational processes to optimize ensure that the architecture deployed in the cloud aligns to the organization's and maintain their workforce, and ensure competencies are in place at security control requirements, resiliency, and compliance requirements, the appropriate time.

3 Governance Perspective e.g. CIO, Program Managers, Project Managers, Enterprise Architects, Business Analysts

how to update the staff skills and organizational processes that are necessary to ensure business governance in the cloud, and manage and measure cloud investments to evaluate their business outcomes.

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

how to update the staff skills and organizational processes that are necessary to deliver and optimize cloud solutions and services.

5 Security Perspective e.g. CISO, IT Security Managers, IT Security Analysts.

how to update the staff skills and organizational processes that are necessary to ensure system health and reliability during the move of operations to the cloud and then to operate using agile, ongoing, cloud computing best practices.

6 Operations Perspective e.g. IT Operations Managers, IT Support Managers.

AWS Free Services

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

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

AWS Support Plans

Basic

Email Support only
For Billing and Account

Developer

Tech Support via **Email** ~24 hours until reply

No third party support

General Guidance

System Impaired

7 Trusted Advisor Checks

\$0 USD /month

Business

Tech Support via **Chat, Phone** Anytime 24/7

Production System Impaired

Production System **DOWN!**

Business-Critical System **DOWN!** < 15m

☀️ Personal Concierge

👓 TAM

All Trusted Advisor Checks

***\$100 USD** / month

***\$15,000 USD** / month

Enterprise

AWS Support Plans

Developer	Business	Enterprise
*\$29 USD /month or 3% of monthly AWS usage <i>whichever is greater</i>	*\$100 USD / month or 10% of monthly AWS usage for the first \$0–\$10K 7% of monthly AWS usage from \$10K–\$80K 5% of monthly AWS usage from \$80K–\$250K 3% of monthly AWS usage over \$250K <i>whichever is greater</i>	*\$15,000 USD / month or 10% of monthly AWS usage for the first \$0–\$150K 7% of monthly AWS usage from \$150K–\$500K 5% of monthly AWS usage from \$500K–\$1M 3% of monthly AWS usage over \$1M <i>whichever is greater</i>
eg. Monthly Spend is \$500 $3\% \text{ of } 500 = \$15 \text{ USD } (\$29)$	eg. Monthly Spend is \$1000 $10\% \text{ of } 1000 = \100 USD	
Monthly Spend is \$1000 $3\% \text{ of } 1000 = \$30 \text{ USD}$	Monthly Spend is \$5000 $10\% \text{ of } 5000 = \500 USD	Monthly Spend is \$12,000 $10\% \text{ of } 10,000 = \1000 USD $7\% \text{ of } 2,000 = 140 \text{ USD}$ $\$1140 \text{ USD}$

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 prelaunch 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 organization 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 visualize usage for consolidated billing

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

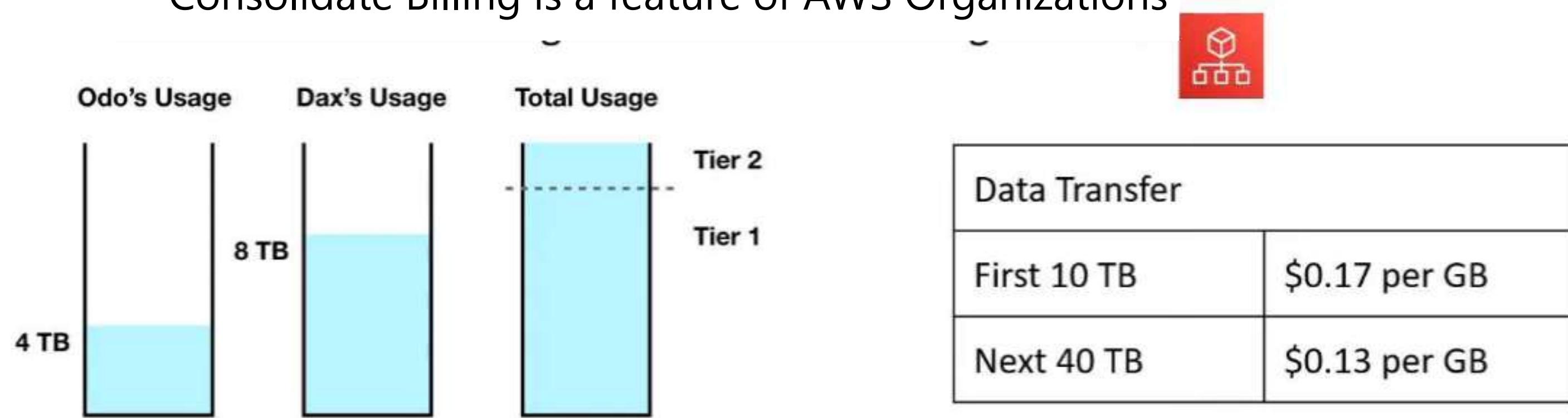


Consolidated Billing - Volume Discounts

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



$$\text{Odo} \quad (4 * 1024) * 0.17 = \$696.32$$

$$\text{Dax} \quad (8 * 1024) * 0.17 = \$1392.64$$

$$\text{Unconsolidated} \quad 696.32 + 1392.64 = \$2088.96$$

$$\text{Consolidated} \quad ((10 * 1024) * 0.17) + ((2 * 1024) * 0.13) = \$2007.04$$



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

1 TB = 1024 GB

AWS Trusted Advisor



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

The screenshot shows the AWS Trusted Advisor dashboard. On the left, a sidebar lists categories: Dashboard, Cost optimization, Performance, Security, Fault tolerance, Service limits, and Preferences. The 'Dashboard' item is selected. The main area is titled 'Checks summary' and shows two items: 'Action recommended' (2) and 'Investigation recommended' (1). A red arrow points from the top of the slide to the 'Dashboard' title. Another red arrow points from the 'Download all checks' button to a detailed view of a specific check: 'Security Groups - Specific Ports Unrestricted'. This detailed view includes a description, a link to refresh, and a link to download.



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

AWS Trusted Advisor

AWS Trusted Advisor providers different level of checks based on your AWS Support Plan

Basic

7 Trusted Advisor Checks

Developer

Business

Enterprise

All Trusted Advisor Checks

AWS providers the following checks for free:

1. MFA on Root Account
2. Security Groups - Specific Ports of Unrestricted
3. Amazon S3 Bucket Permissions
4. Amazon EBS Public Snapshots
5. Amazon RDS Public Snapshots
6. IAM Use - discourage the use of root access
7. Service Limits (All Service limits checks are free)

6 Security Checks

AWS Trusted Advisor



Cost Optimization

Amazon EC2 Reserved Instances Optimization
Low Utilization Amazon EC2 Instances
Underutilized Amazon EBS Volumes Amazon
EC2 Reserved Instance Lease Expiration Amazon
RDS Idle DB Instances Amazon Route 53 Latency
Resource Record Sets **Idle Load Balancers**
Unassociated Elastic IP Addresses
Underutilized Amazon Redshift Clusters



Performance

CloudFront Alternate Domain Names
Amazon EBS Provisioned IOPS (SSD) Volume Attachment Configuration
Amazon EC2 to EBS Throughput Optimization
Amazon Route 53 Alias Resource Record Sets
CloudFront Content Delivery Optimization
CloudFront Header Forwarding and Cache Hit Ratio

High Utilization Amazon EC2 Instances

Large Number of EC2 Security Group Rules Applied to an Instance
Large Number of Rules in an EC2 Security Group Overutilized
Amazon EBS Magnetic Volumes



Security

AWS CloudTrail Logging
IAM Password Policy
MFA on Root Account
Security Groups - Specific Ports Unrestricted
Security Groups - Unrestricted Access
Amazon S3 Bucket Permissions
IAM Access Key Rotation
Amazon EBS Public Snapshots
Amazon RDS Public Snapshots
Amazon RDS Security Group Access Risk
Amazon Route 53 MX Resource Record Sets and Sender Policy Framework
CloudFront Custom SSL Certificates in the IAM Certificate Store
CloudFront SSL Certificate on the Origin Server
ELB Listener Security
ELB Security Groups
Exposed Access Keys
IAM Use

AWS Trusted Advisor



Fault Tolerance

Amazon EBS Snapshots
Amazon RDS Multi-AZ
Amazon S3 Bucket Logging
Amazon S3 Bucket Versioning
Amazon Aurora DB Instance Accessibility
Amazon EC2 Availability Zone Balance

Amazon RDS Backups

Amazon Route 53 Deleted Health Checks
Amazon Route 53 Failover Resource Record Sets
Amazon Route 53 High TTL Resource Record Sets
Amazon Route 53 Name Server Delegations

Auto Scaling Group Health Check

Auto Scaling Group Resources

ELB Connection Draining

ELB Cross-Zone Load Balancing

Load Balancer Optimization

VPN Tunnel Redundancy

AWS Direct Connect Connection Redundancy

AWS Direct Connect Location Redundancy

AWS Direct Connect Virtual Interface Redundancy

EC2Config Service for EC2 Windows Instances

ENA Driver Version for EC2 Windows Instances

NVMe Driver Version for EC2 Windows Instances



Service Limits

Auto Scaling Groups
Auto Scaling Launch Configurations
CloudFormation Stacks
DynamoDB Read Capacity
DynamoDB Write Capacity
EBS Active Snapshots
EBS Active Volumes
EBS Cold HDD (sc1) Volume Storage
EBS General Purpose SSD (gp2) Volume Storage
EBS Magnetic (standard) Volume Storage
EBS Provisioned IOPS (SSD) Volume Aggregate IOPS
EBS Provisioned IOPS SSD (io1) Volume Storage
EBS Throughput Optimized HDD (st1) Volume Storage
EC2 Elastic IP Addresses
EC2 On-Demand Instances
EC2 Reserved Instance Leases
ELB Active Load Balancers
IAM Group
IAM Instance Profiles
IAM Policies
IAM Roles
IAM Server Certificates IAM Users
Kinesis Shards per Region

RDS Cluster Parameter Groups
RDS Cluster Roles
RDS Clusters
RDS DB Instances
RDS DB Parameter Groups
RDS DB Security Groups
RDS DB Snapshots Per User
RDS Event Subscriptions
RDS Max Auths per Security Group
RDS Option Groups
RDS Read Replicas per Master
RDS Reserved Instances
RDS Subnet Groups
RDS Subnets per Subnet Group
RDS Total Storage Quota
Route 53 Hosted Zones
Route 53 Max Health Checks
Route 53 Reusable Delegation Sets
Route 53 Traffic Policies
Route 53 Traffic Policy Instances
SES Daily Sending Quota
VPC
VPC Elastic IP Address VPC Internet Gateways

Service Level Agreement

What is a Service Level Agreement (SLA)?

A 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 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 A 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%
- 99.99999999% (commonly called **Nine nines**)
- 99.999999999% (commonly called **Nine elevens**)

AWS Service Level Agreements

DynamoDB SLA

AWS will use commercially reasonable efforts to make DynamoDB available with a Monthly Uptime Percentage for each AWS region, during any monthly billing cycle, of (a) at least 99.999% if the Global Tables SLA applies, or (b) at least 99.99% if the Standard SLA applies

In the event DynamoDB does not meet the Service Commitment, you will be eligible to receive a Service Credit as described below

	Monthly Uptime Percentage	Service Credit Percentage
<i>Global Tables SLA</i>	Less than 99.999% but equal to or greater than 99.0%	10%
	Less than 99.0% but equal to or greater than 95.0%	25%
	Less than 95.0%	100%
<i>Standard SLA</i>	Less than 99.99% but equal to or greater than 99.0%	10%
	Less than 99.0% but equal to or greater than 95.0%	25%
	Less than 95.0%	100%

AWS Service Level Agreements

Compute SLAs

- Amazon Elastic Compute Cloud (Amazon EC2)*
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic Container Service (Amazon ECS)
- AWS Fargate for Amazon ECS and Amazon EKS

AWS makes two SLA commitments for the Included Services:

1. a Region-Level SLA that governs Included Services deployed across multiple AZs or regions, and
2. an Instance-Level SLA that governs Amazon EC2 instances individually.

	Monthly Uptime Percentage	Service Credit Percentage
<i>Region-Level SLA</i>	Less than 99.99% but equal to or greater than 99.0%	10%
	Less than 99.0% but equal to or greater than 95.0%	30%
	Less than 95.0%	100%
<i>Instance-Level SLA</i>	Less than 99.5% but equal to or greater than 99.0%	10%
	Less than 99.0% but equal to or greater than 95.0%	30%
	Less than 95.0%	100%

AWS Service Level Agreements

RDS SLA

AWS will use commercially reasonable efforts to make Multi-AZ instances available with a Monthly Uptime Percentage of at least 99.95% during any monthly billing cycle

In the event Amazon RDS does not meet the Monthly Uptime Percentage commitment, you will be eligible to receive a Service Credit as described below.

Monthly Uptime Percentage

Service Credit Percentage

Less than 99.95% but equal to or greater than 99.0%

10%

Less than 99.0% but equal to or greater than 95.0%

25%

Less than 95.0%

100%

Service Health Dashboard

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

The screenshot shows the AWS Service Health Dashboard interface. At the top, there's a navigation bar with the AWS logo and the title "SERVICE HEALTH DASHBOARD". Below the navigation, a breadcrumb trail says "Amazon Web Services » Service Health Dashboard" and a call-to-action button "Get a personalized view of AWS service health" with a link "Open the Personal Health Dashboard".

Current Status - Oct 19, 2021 PDT

A descriptive text explains that Amazon Web Services publishes up-to-the-minute information on service availability. It encourages users to check back for current status or subscribe to an RSS feed for notifications. All dates and times are Pacific Time (PST/PDT).

The main content area features regional status tables:

- North America**: Shows "Recent Events" (No recent events) and "Remaining Services": Alexa for Business (N. Virginia) is "Service is operating normally", Amazon API Gateway (Montreal) is "Service is operating normally", and Amazon API Gateway (N. California) is "Service is operating normally".
- South America**, **Europe**, **Africa**, **Asia Pacific**, and **Middle East** sections are also present but show no recent events.

Each service entry includes a green checkmark icon and a "Details" link under the "Service is operating normally" status.

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.

The screenshot shows the AWS Personal Health Dashboard interface. At the top, there's an overview section with three metrics: 0 Open issues (Past 7 days), 1 Scheduled changes (Upcoming and past 7 days), and 1 Other notifications (Past 7 days). A red arrow points from the text above to this overview section. To the right is a 'Set up alerts' section with a 'Create rule' button. Below the overview is a 'Scheduled changes' section for an 'EC2 persistent instance retirement scheduled' event. This section includes tabs for 'Details' (selected) and 'Affected resources'. Under 'Event data', it shows the event name 'EC2 persistent instance retirement scheduled', status 'Upcoming', region 'us-east-1', and category 'Scheduled change'. It also lists the start time as March 19, 2021 at 6:35:40 PM UTC-7 and end time as March 19, 2021 at 6:36:40 PM UTC-7. The 'Affected resources' section shows one resource. The 'Description' section explains that EC2 has detected degradation of the underlying hardware hosting the instance and will stop it after a specified time. It also provides a link to the AWS Management Console for more information. The bottom section, 'What will happen to my instance?', states that the instance will be stopped after the retirement date, and any data on local instance-store volumes will be lost.

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.



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 which allows you to use AWS at no cost for the first 12 months of signup
Or free usage up to a certain monthly limit forever

EC2 Web Server

t2.micro 750 hours per month for 1 year



RDS Database (MySQL or Postgres) **t2.db.micro** 750 hours per month for 1 year

ELB Load Balancer

750 hours per month for 1 year

The Best Deals

Amazon CloudFront Homepage Video 50 GB data-transfer out in total for 1 year

Amazon Connect Toll Free Number 90 minutes of call-time per month for 1 year

Amazon ElastiCache Caching **cache.t3.micro** 750 hours per month for 1 year

Amazon ElasticSearch Service Full Text Search

750 hours per month for 1 year

PinPoint Campaign / Marketing Emails

5.000 targeted users per month for 1 year

SES Emails sent by your web-application

62.000 emails per month forever

AWS CodePipeline CI/CD 1 Pipeline free

AWS CodeBuild Building Code 100 build minutes per month forever

AWS Lambda Serverless Compute

1M free request per month

3.2M seconds of compute time per month

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 several ways:

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

Summary

Total amount remaining	Total amount used
\$500.00	\$332.00

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 organization up to business opportunities and allows exclusives trainings and marketing events

When joining the APN you can either be a:
Consulting Partner - you help companies utilize AWS
Technology Partner - you build technology ontop 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

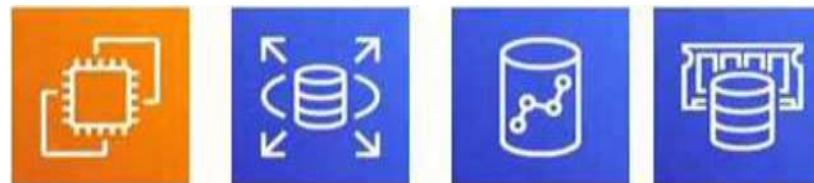


AWS Budgets give you the ability to setup alerts if you **exceed** or are **approaching** your defined budget

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 is 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

Choose your budget amount in \$\$\$

Budgeted amount

 Last month's cost \$126.59

Usage unit(s)

 Usage Type Group
 EC2: Running Hours (Hrs) X
 Usage Type

Choose based a different kind of unit

Budgeted amount

 Hrs Last month's usage 2260.54 Hrs

AWS Budgets

You have a list of budgets:



Budgets (3) <small>Info</small>							
<input type="text"/> Find a budget		Thresholds		Budget	Amount used	Forecasted am...	Current vs. budgeted
Name	Thresholds	Exceeded (1)	OK	\$200.00	\$318.79	\$392.07	159.39%
AWS Credits Budgets	Exceeded (1)	\$200.00	\$318.79	\$392.07	159.39%	196.02	
MinecraftServerBudget	OK	\$100.00	\$0.00	-	0.00%	-	
Overall Costs	Exceeded (1)	\$100.00	\$393.77	\$539.28	393.77%	539.28	

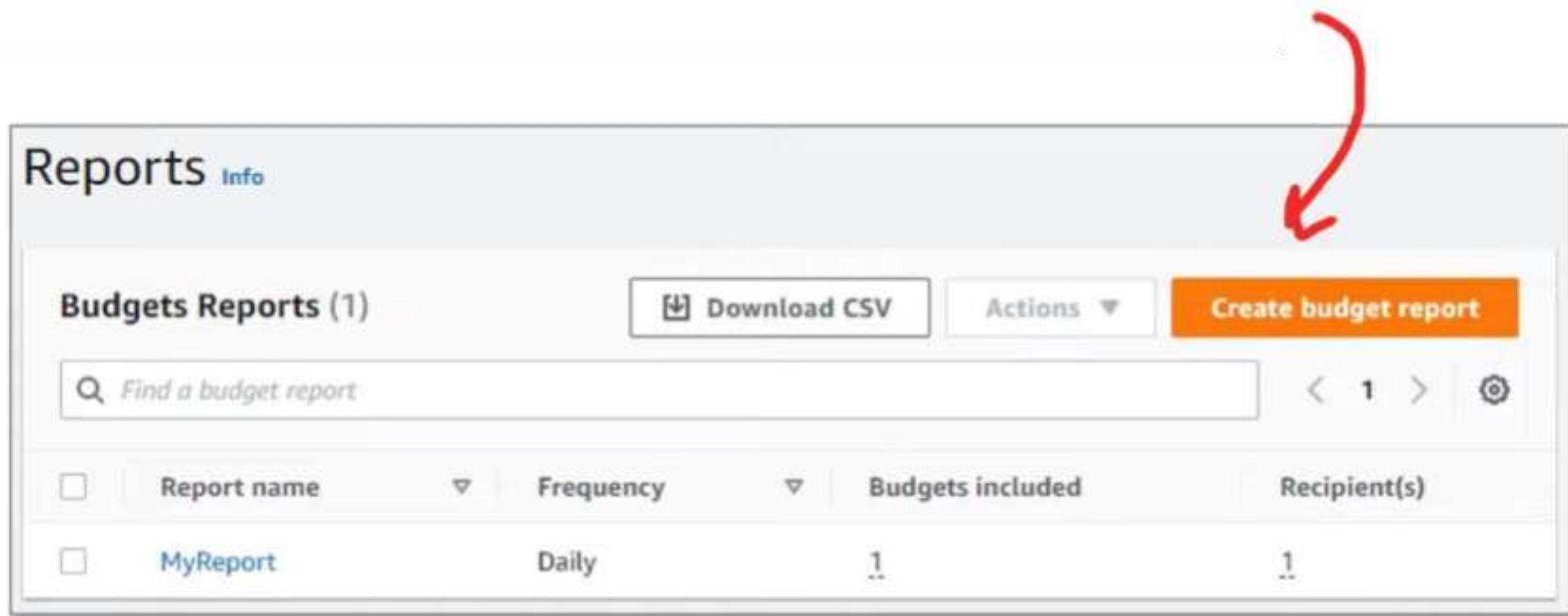
You can see your budget history, download it as a CSV



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

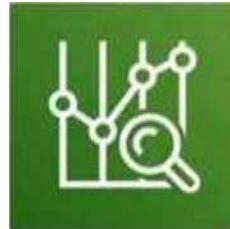
AWS Budget Reports

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.

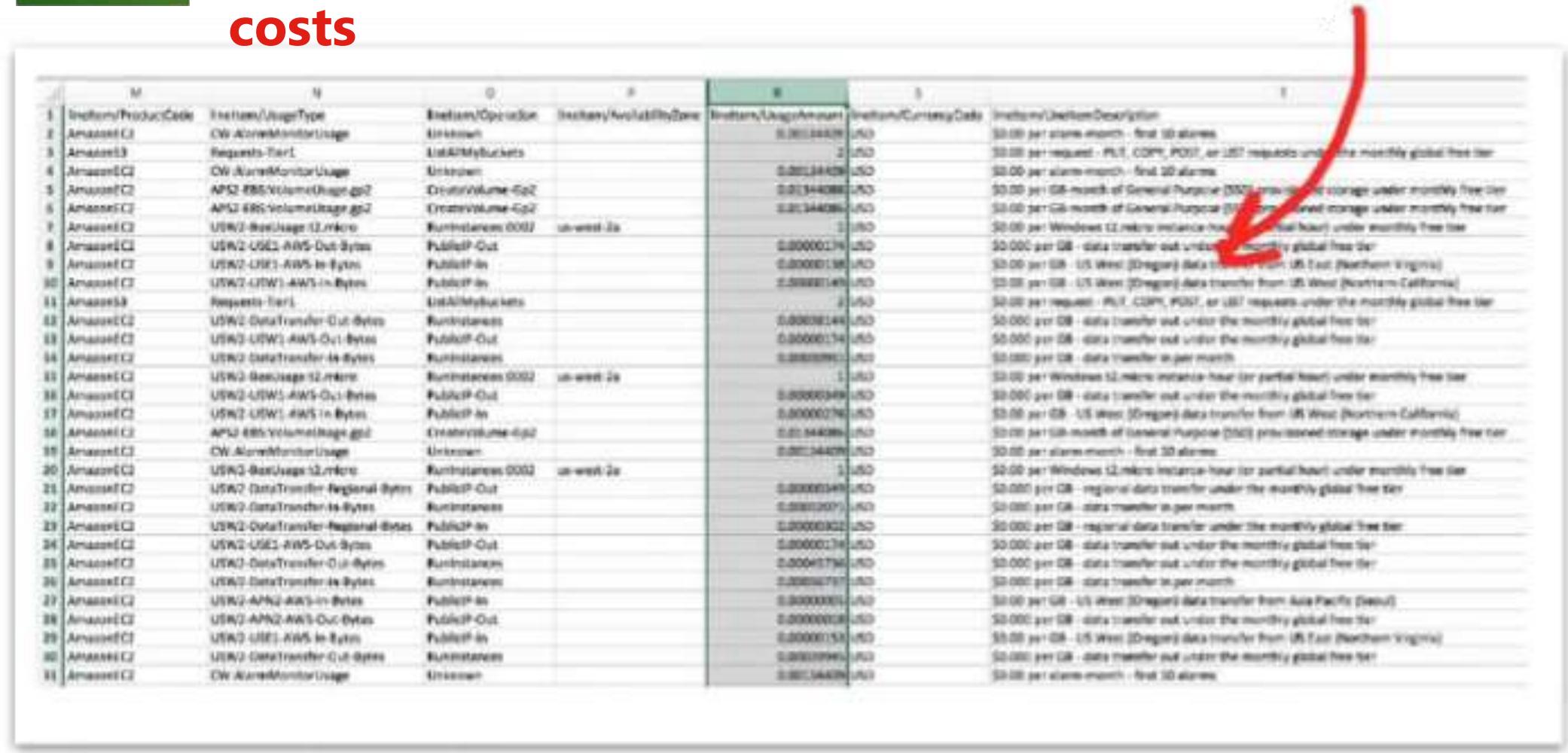


AWS Budget Reports serve as a more convenient way of staying on top of reports since they are delivered to your email instead of logging into the AWS Management Console

AWS Cost and Usage Reports (CUR)



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



The screenshot shows a Microsoft Excel spreadsheet titled 'CostAndUsageReport.csv'. The columns are labeled M through T. A red arrow points from the text 'choose the granularity of your data by selecting hourly, daily or monthly' to the column header 'Granularity' (T). The data includes various AWS services like Amazon S3, Amazon EC2, and Amazon RDS, along with their usage details and costs.

M	N	O	P	Q	R	S	T
Item/ProductCode	ItemName/UsageType	ItemName/Operator	ItemName/AvailabilityZone	ItemName/UsageAmount	ItemName/Currency	ItemName/Description	Granularity
AmazonEC2	CW AlarmsMonitorUsage	Unknown		0.00034408	USD	\$0.00 per alarm-month - first 10 alarms	
AmazonS3	Requests-Hr1	UnlmtdBuckets		2 USD	USD	\$0.00 per request - PUT, COPY, POST, or LIST requests under the monthly global free tier	
AmazonEC2	CW AlarmMonitorUsage	Unknown		0.00034408	USD	\$0.00 per alarm-month - first 10 alarms	
AmazonEC2	API2-KBS-VolmmtUsage-gp2	CreateVolume-Gp2		0.01344088	USD	\$0.00/month of General Purpose (GP2) provisioned storage under monthly free tier	
AmazonEC2	API2-KBS-VolmmtUsage-gp2	CreateVolume-Gp2		0.01344088	USD	\$0.00/month of General Purpose (GP2) provisioned storage under monthly free tier	
AmazonEC2	USW2-Batchage (2)Micro	RunInstances 0002	us-west-2a	2 USD	USD	\$0.00 per Windows 12 micro instance-hour (or partial hour) under monthly free tier	
AmazonEC2	USW2-USG2-AWS-Dst-Bytes	PublicIP-Out		0.00000279	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	USW2-USG2-AWS-In-Bytes	PublicIP-In		0.0000138	USD	\$0.00 per GB - US West (Oregon) data transfer to US East (Northern Virginia)	
AmazonEC2	USW2-USW1-AWS-In-Bytes	PublicIP-In		2.0000147	USD	\$0.00 per GB - US West (Oregon) data transfer from US West (Northern California)	
AmazonS3	Requests-Hr1	UnlmtdBuckets		2 USD	USD	\$0.00 per request - PUT, COPY, POST, or LIST requests under the monthly global free tier	
AmazonEC2	USW2-DataTransfer-Out-Bytes	RunInstances		0.00000249	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	USW2-USW1-AWS-Out-Bytes	PublicIP-Out		0.00000249	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	USW2-DataTransfer-In-Bytes	RunInstances		0.00000249	USD	\$0.00 per GB - data transfer in per month	
AmazonEC2	USW2-Batchage (2)Micro	RunInstances 0002	us-west-2a	1 USD	USD	\$0.00 per Windows 12 micro instance-hour (or partial hour) under monthly free tier	
AmazonEC2	USW2-USG2-AWS-Out-Bytes	PublicIP-Out		0.00000249	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	USW2-USW1-AWS-In-Bytes	PublicIP-In		0.00000279	USD	\$0.00 per GB - US West (Oregon) data transfer from US West (Northern California)	
AmazonEC2	API2-KBS-VolmmtUsage-gp2	CreateVolume-Gp2		0.01344088	USD	\$0.00/month of General Purpose (GP2) provisioned storage under monthly free tier	
AmazonEC2	CW AlarmsMonitorUsage	Unknown		0.00034408	USD	\$0.00 per alarm-month - first 10 alarms	
AmazonEC2	USW2-Batchage (2)Micro	RunInstances 0002	us-west-2a	1 USD	USD	\$0.00 per Windows 12 micro instance-hour (or partial hour) under monthly free tier	
AmazonEC2	USW2-DataTransfer-Regional-Bytes	PublicIP-Out		0.00000249	USD	\$0.00 per GB - regional data transfer under the monthly global free tier	
AmazonEC2	USW2-DataTransfer-In-Bytes	RunInstances		0.00000207	USD	\$0.00 per GB - data transfer in per month	
AmazonEC2	USW2-DataTransfer-Regional-Bytes	PublicIP-In		0.00000207	USD	\$0.00 per GB - regional data transfer under the monthly global free tier	
AmazonEC2	USW2-USG2-AWS-Dst-Bytes	PublicIP-Out		0.00000279	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	USW2-DataTransfer-Out-Bytes	RunInstances		2.0000179	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	USW2-DataTransfer-In-Bytes	RunInstances		2.0000179	USD	\$0.00 per GB - data transfer in per month	
AmazonEC2	USW2-APN2-AWS-In-Bytes	PublicIP-In		0.00000311	USD	\$0.00 per GB - US West (Oregon) data transfer from Asia Pacific (Seoul)	
AmazonEC2	USW2-APN2-AWS-Out-Bytes	PublicIP-Out		0.00000311	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	USW2-USG2-AWS-In-Bytes	PublicIP-In		0.0000153	USD	\$0.00 per GB - US West (Oregon) data transfer from US East (Northern Virginia)	
AmazonEC2	USW2-DataTransfer-Out-Bytes	RunInstances		0.00000996	USD	\$0.00 per GB - data transfer out under the monthly global free tier	
AmazonEC2	CW AlarmsMonitorUsage	Unknown		0.00034408	USD	\$0.00 per alarm-month - first 10 alarms	

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



Places the reports into S3



Use Athena to turn the report into a queryable database



QuickSight to visualize your billing data as graphs

Cost Allocation Tags

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

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

There are **two types** of tags:

- User-Defined
 - Eg Project
- AWS Generated
 - E.g. awsxreatedBy

AWS-generated cost allocation tags (17)		
<input type="checkbox"/> Tag key	Status	
<input type="checkbox"/> aws:createdBy	<input checked="" type="checkbox"/> Active	
<input type="checkbox"/> aws:cloudformation:stack-name	<input type="checkbox"/> Inactive	
<input type="checkbox"/> aws:ec2launchtemplate:id	<input type="checkbox"/> Inactive	

Billing Alerts/Alarms



You can create your own Alarms in CloudWatch Alarms to monitor spend. They are commonly called "Billing Alarms"

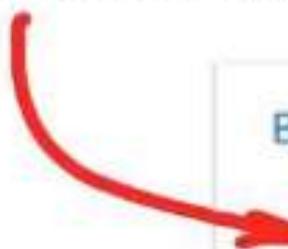
You first need to turn on **Billing Alerts**



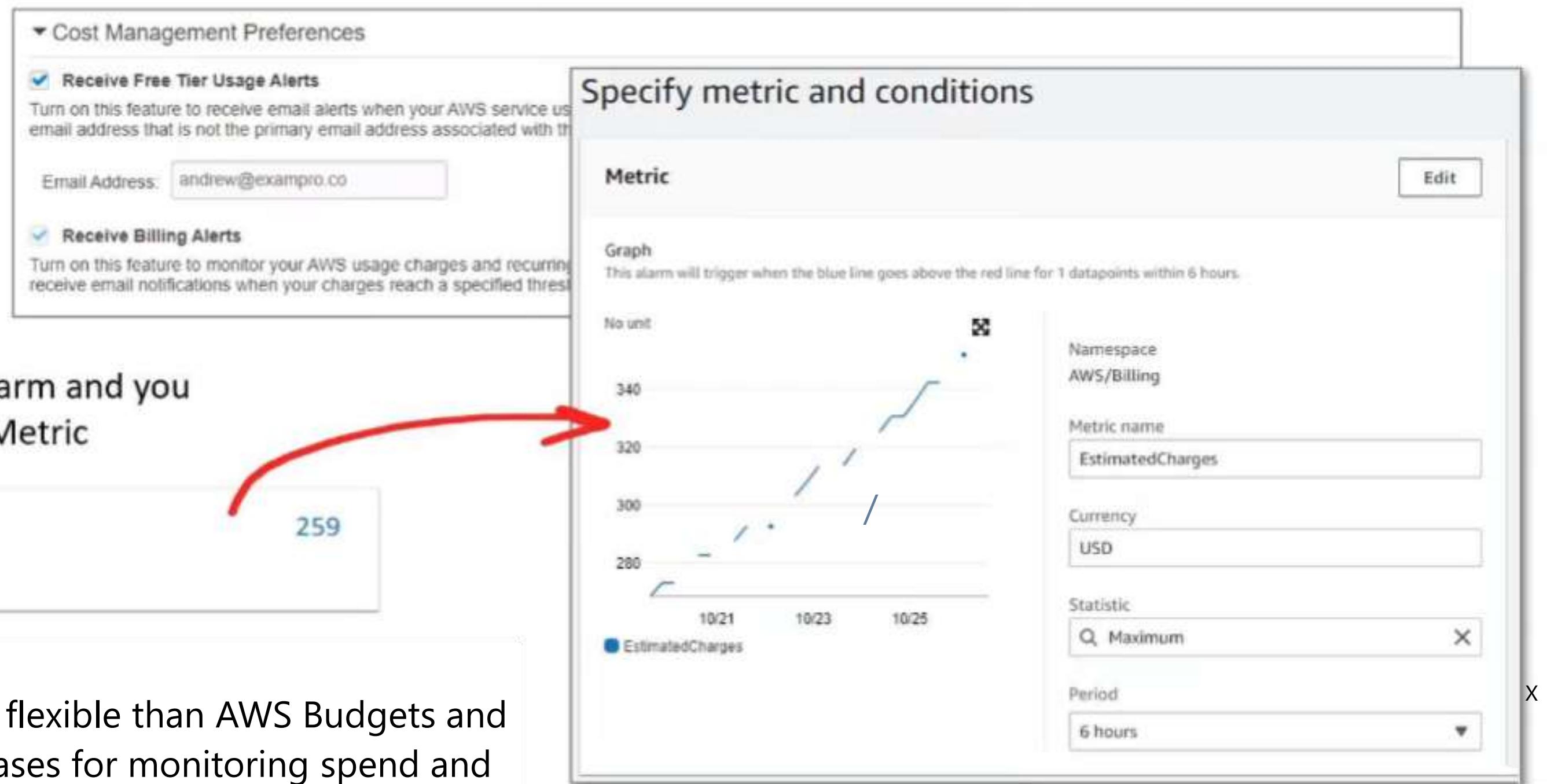
Go create a CloudWatch Alarm and you can choose Billing as your Metric

Billing

259



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

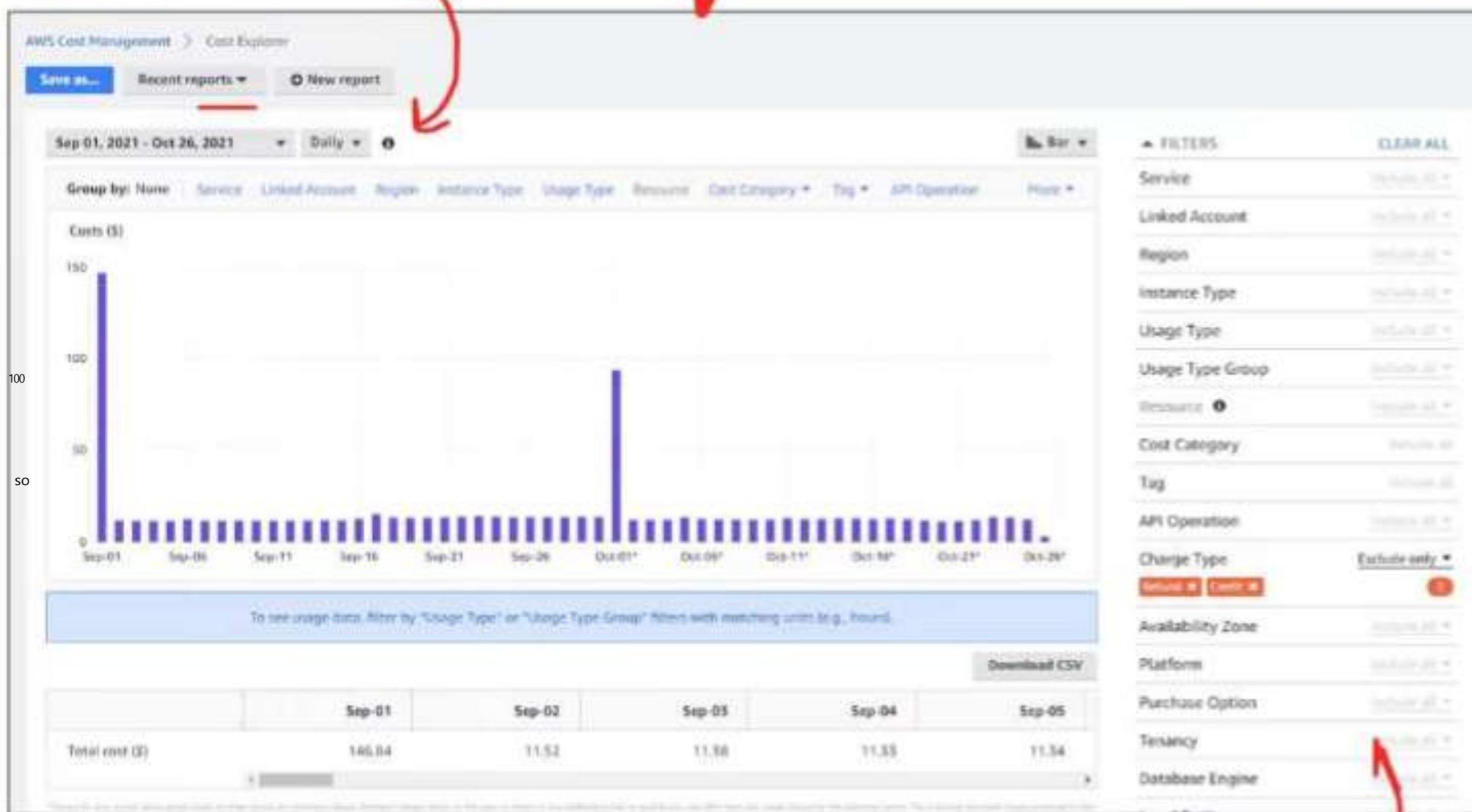


AWS Cost Explorer



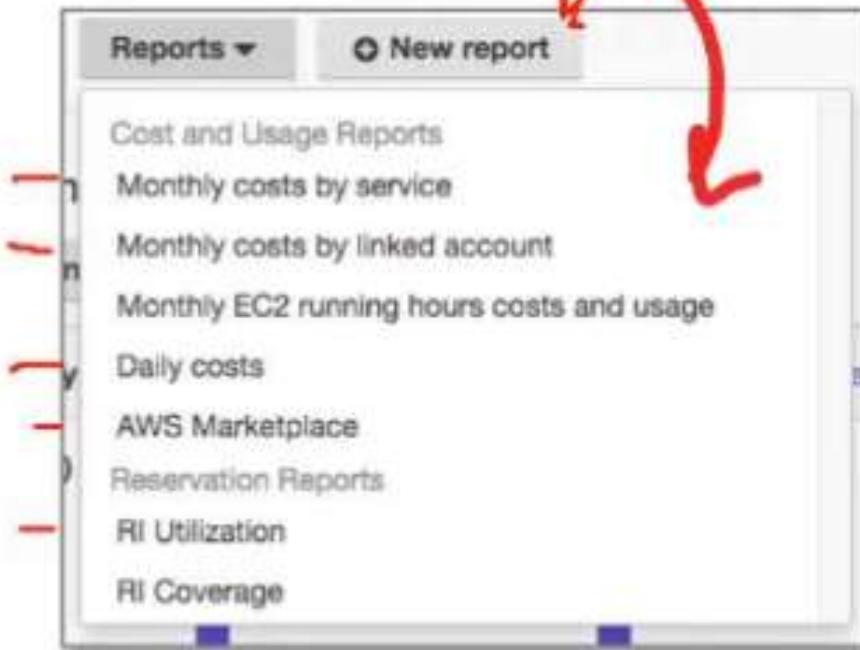
AWS Cost Explorer lets you visualize, understand, and manage

Specific type range
and aggregation



Robust filtering

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



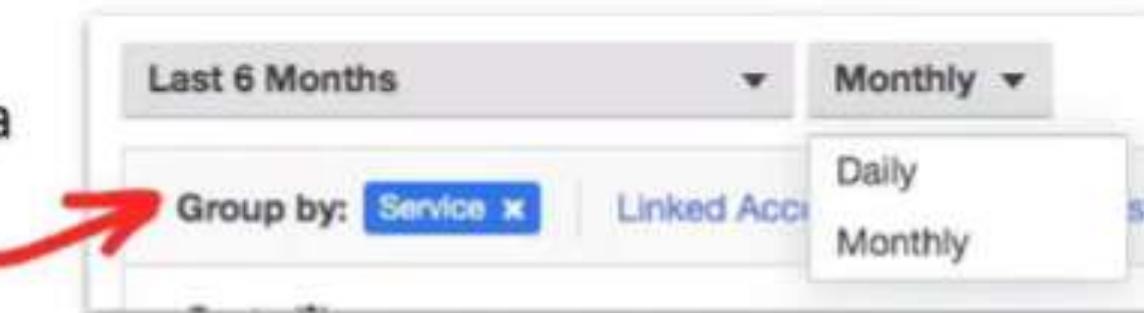
Use **forecasting** to get an idea of future costs

Forecasted month end costs
\$456.94 1%
Over last month

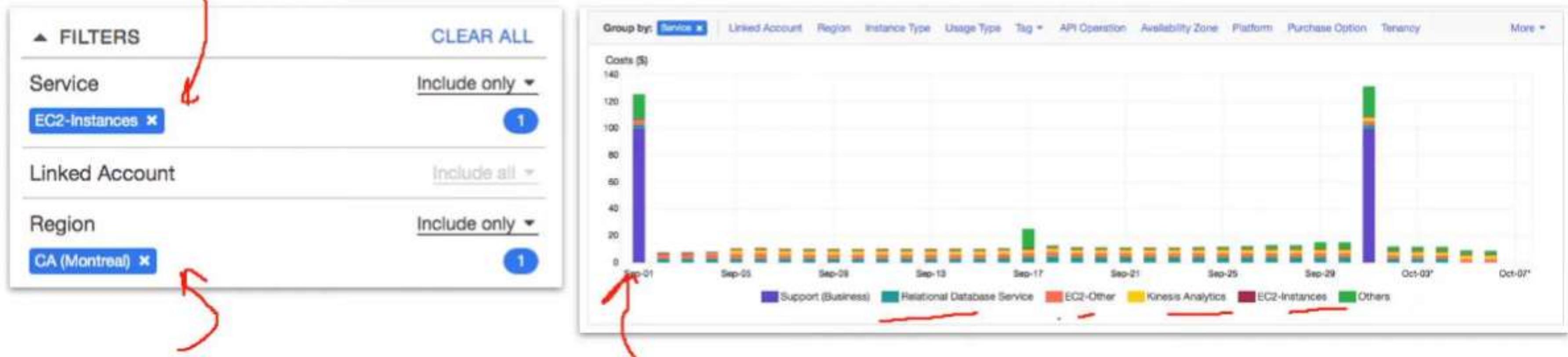
AWS Cost Explorer

£

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.

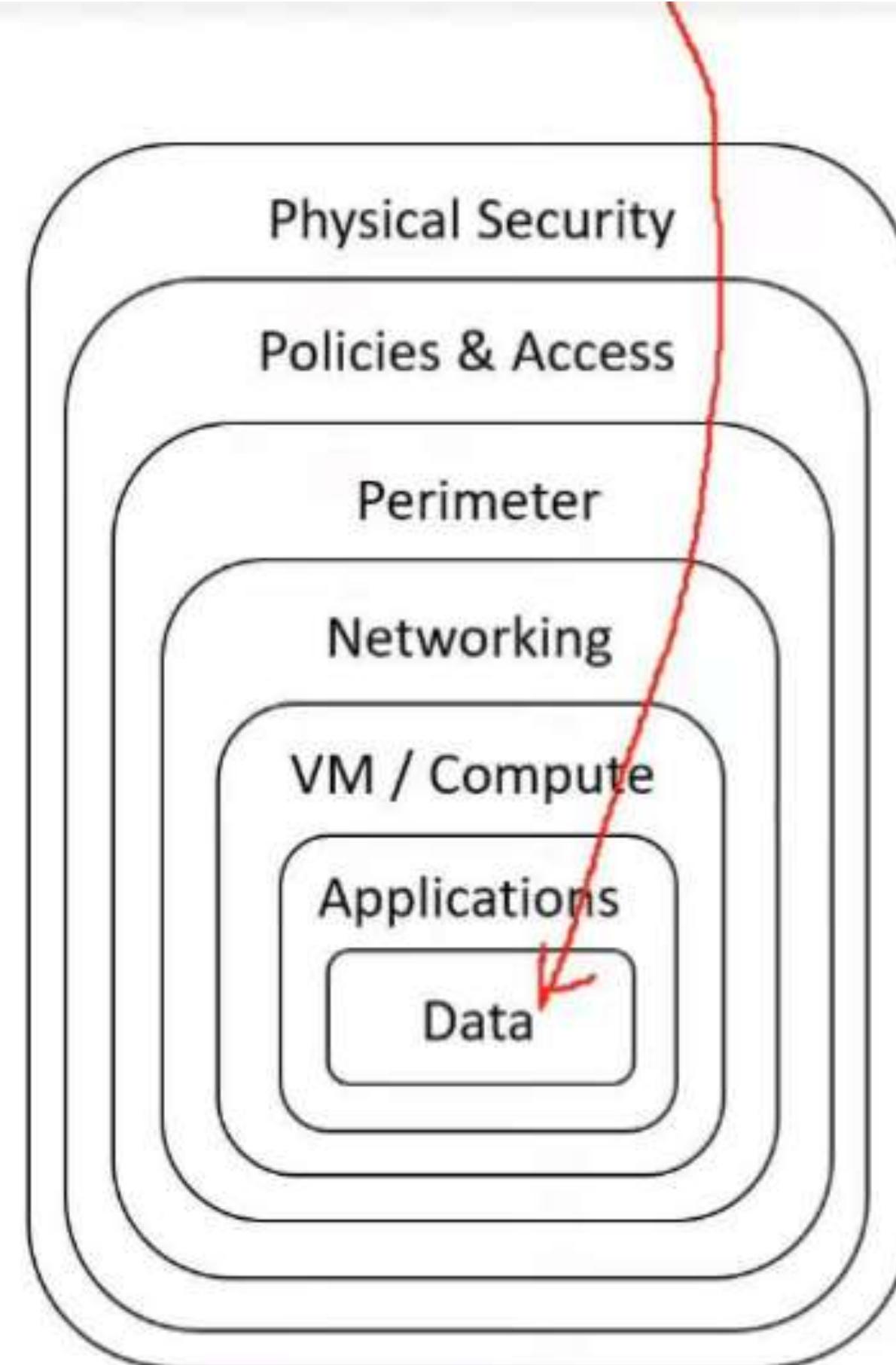
There are two versions of this API:

- 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/vl.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

Defense 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

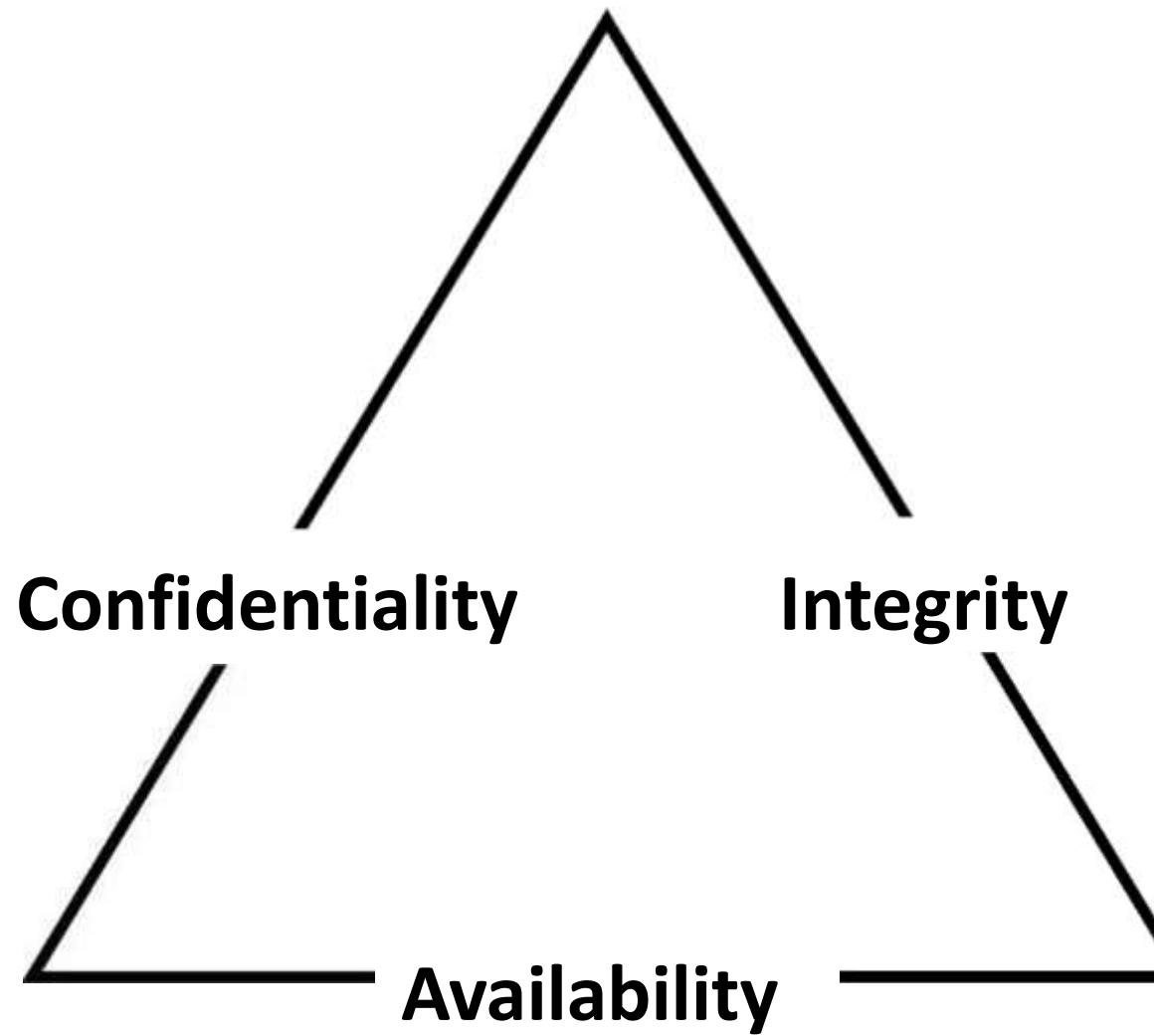
controlling access to infrastructure and change control.

7. Physical

limiting access to a datacenter to only authorized personnel.

Confidentiality, Integrity, Availability (CIA)

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



Confidentiality

confidentiality is a component of privacy that implements to protect our data from unauthorized 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 utilizing ACID compliant databases for valid transactions. Utilizing 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 modernize or suggest alternatives to CIA triad:

- (1998) Six Atomic Elements of Information eg. 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

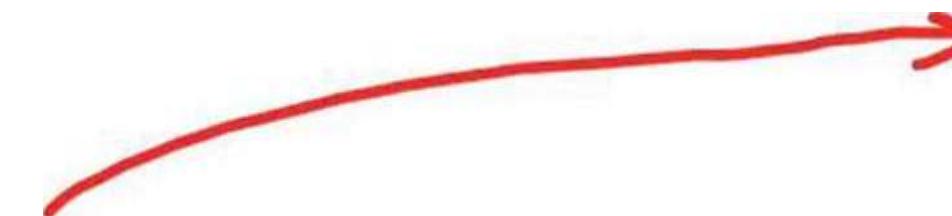
Encryption

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.

Cyphers

What is a cypher?

An algorithm that performs encryption or decryption. Cipher is synonymous with "code"

What is ciphertext

Ciphertext is the result of encryption performed on plaintext via an algorithm



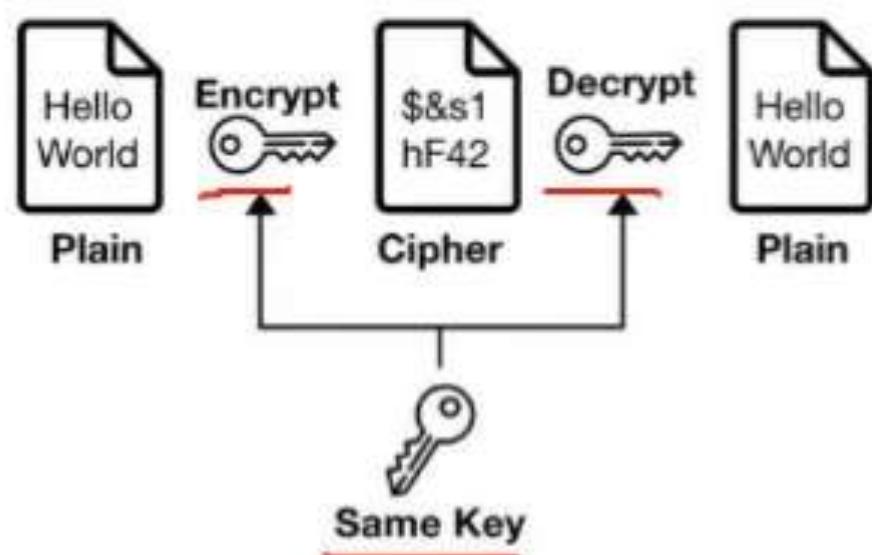
Code word	Code No.	Message or true reading
Ou»»* IMUI*	00 01 02	Authority—Continued Give them authority Authority—Continued
ffiy*	03	you xiti.
Cannot	00	-----
Cannula --	01	-----
Cannulated	02	-----
Canny.....	03	-----
Canoe.....	04	-----
Canoed----	05	-----
Canoeing	06	-----
Canoeist —	07	-----
	08	-----
CkntalouW	38	iHrrtlvrr Verbal authority
Can tar....	39	What ia their mnthority
Cautaro	40	What in your authority
Cantata... (41	Who ia your authority With
natation.	42	authority With our
Cau talon-	43	authority With their
Quitatrice	44	authority Wtth your
Chntod...	45	authority Within't
Canteen ..	46	authority Without our
Canteen*..	47	authority Without their
Canter	48	authority Without your
	49	authilonty
CupaMlly	88	Wo arr tatburitml to W« nut
I'ajwM..	89	authorised to You an? authorized
	90	You are authoris'd to You arc
Giparifini.	91	authorized to answer You an*
Oapadf.T.,	92	authorized to aarurv You arr
Caparbw.	93	authorized to convey You an
C*parit <u>t</u>	94	authorized to atno You arr htrrliv
Capacity..	95	authorized You an? hereby
	96	auUe.rix.-d to Y ou ant not
	97	authorized Authoring

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

Cryptographic Keys

What is a cryptographic key?

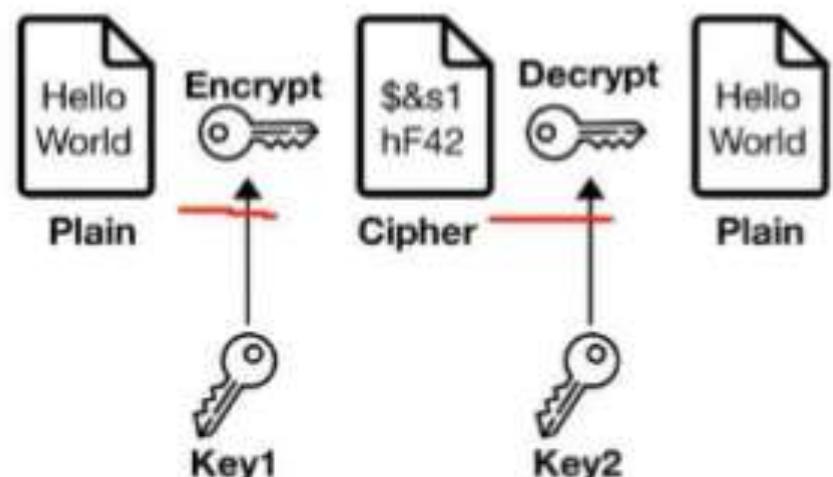
A key is a variable used in conjunction with an encryption algorithm in order to encrypt or decrypt data.



What is symmetric encryption?

The same key is used for encoding and decoding, eg

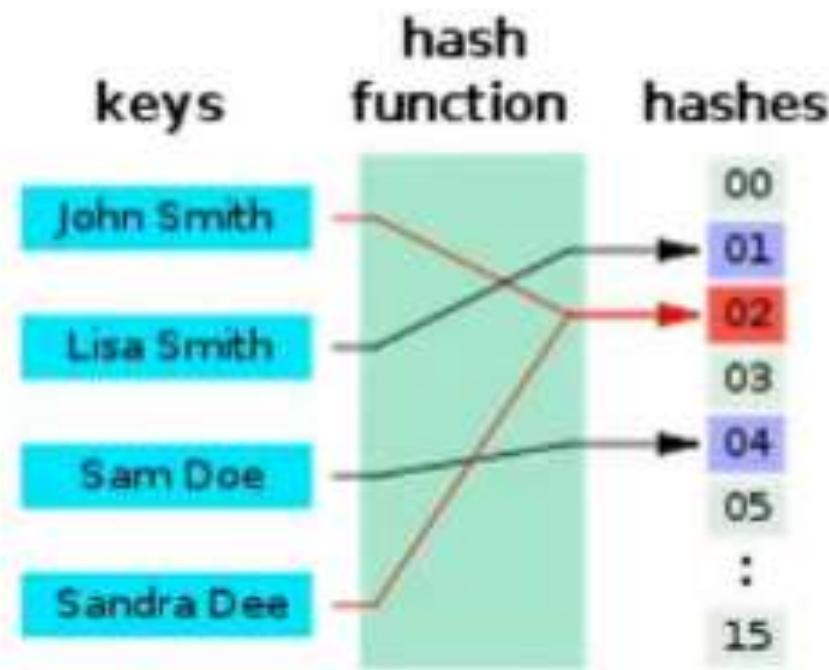
Advanced Encryption Standard (AES)



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 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 successful logged in.

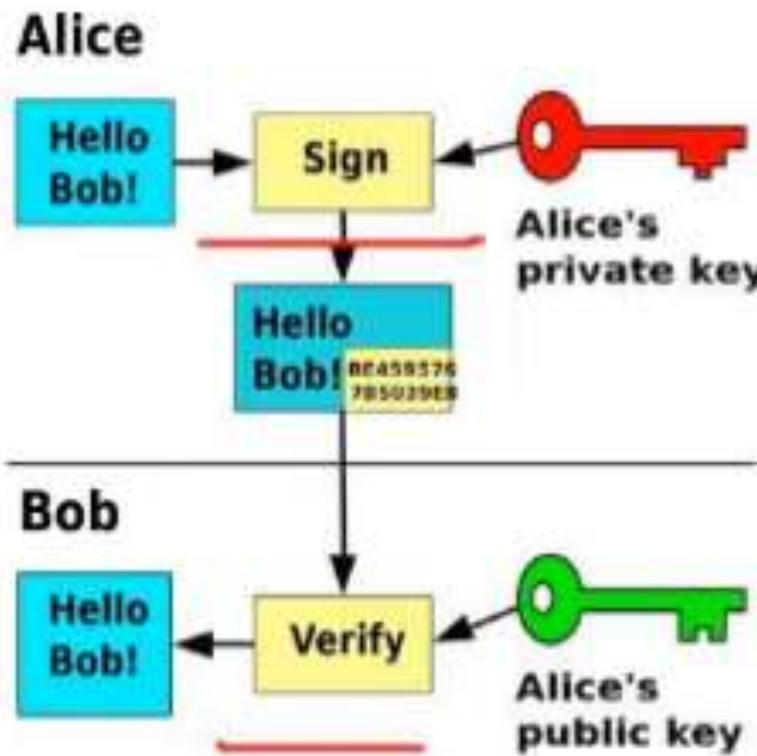
Popular hashing functions are **MD5, SHA256 and Bcrypt**

If an attacker knows what function you are using and stole your database, they could enumerate a dictionary of password 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 is not from the expected sender?

There are three algorithms to 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-keygen -t rsa
```

SSH uses a public and private key to authorize 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

Data that is secure when moving between locations

Algorithms: **TLS, SSL**

Encryption At-Rest

Data that is secure when residing on storage or within a database

Algorithms: **AES, RSA**

Transport Layer Security (TLS)

An encryption protocol for data integrity between two or more communicating computer application.

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 application *SSL 1.0, 2.0 and 3.0 are deprecated*

Common Compliance Programs

Compliance Programs

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

Health Insurance Portability and Accountability Act of 1996) is United States legislation that provide data privacy and security provisions for safeguarding medical information.

The Payment Card Industry Data Security Standard (PCI DSS)

When you want to sell things online and you need to handle credit card information.



Common Compliance Programs



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.



System and Organization Controls (SOC)

SOC 1 — 18 standard and report on the effectiveness of internal controls (SSAE) at a service organization • 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 organization

SOC 3 — A report based on the Trust Services Criteria that can be freely distributed



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.



Federal Information Processing Standard (FIPS) 140-2

US and Canadian government standard that specifies the security requirements for cryptographic modules that protect sensitive information.

Common Compliance Programs



Personal Health Information Protection Act (PHIPA)

An Ontario provincial law (Canada) that regulates patient Protected Health Information



Health Insurance Portability and Accountability Act (HIPAA).

US federal law that regulates patient Protected Health Information



Cloud Security Alliance (CSA) STAR Certification

Independent third-party assessment of a cloud provider's security posture

Common Compliance Programs



Federal Risk and Authorization Management Program (FedRAMP)

US government standardized approach to security authorizations for Cloud Service Offerings



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.



General Data Protection Regulation (GDPR)

A European privacy law. Imposes new rules on companies, government agencies, non-profits, and other organizations that offer goods and services to people in the European Union (EU), or that collect and analyze data tied to EU residents.

Penetration Testing

What is PenTesting?

An authorized 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**



Choose your report

Reports Info

Reports (82)

Q canada

Copy link Download report

Title	Reporting period	Category	Description
Government of Canada (GC) Partner Package	August 25, 2017 to current	Alignment Documents	The Government of Canada (GC) Partner Package is intended for use by partners and customers when building applications and solutions on AWS that need to meet the GC requirements based on the Protected B/Medium Integrity/Medium Availability (PBMM) profile. The documents available in this package include: Partner Package Playbook, Controls Implementation Summary (CIS)/Customer Responsibility Matrix (CRM), and Government of Canada PBMM Security Assessment and Letter of Attestation.

View the PDF



Download the Excel

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.

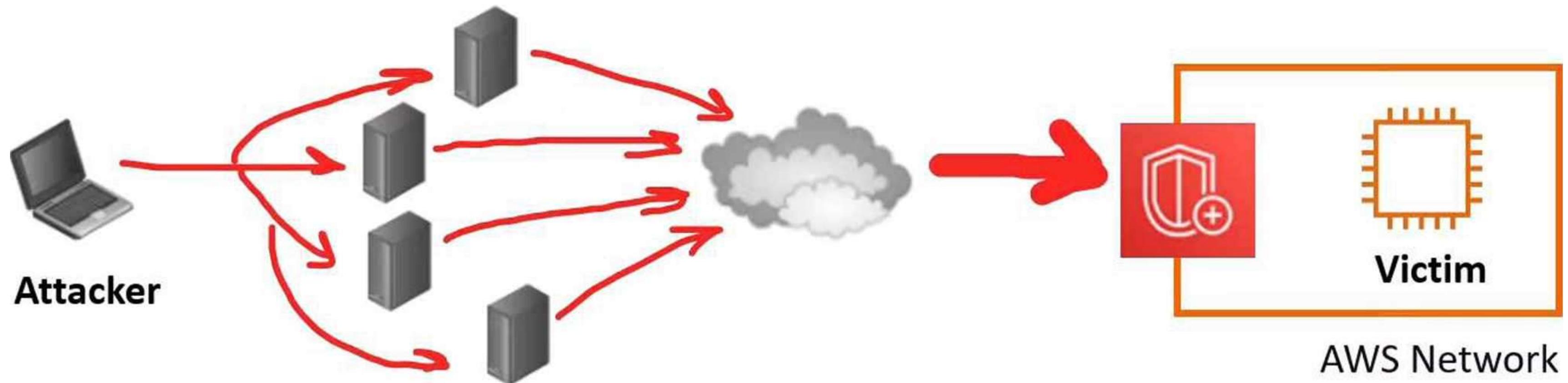
One very popular benchmark you can run is by CIS which has **699 checks!**



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

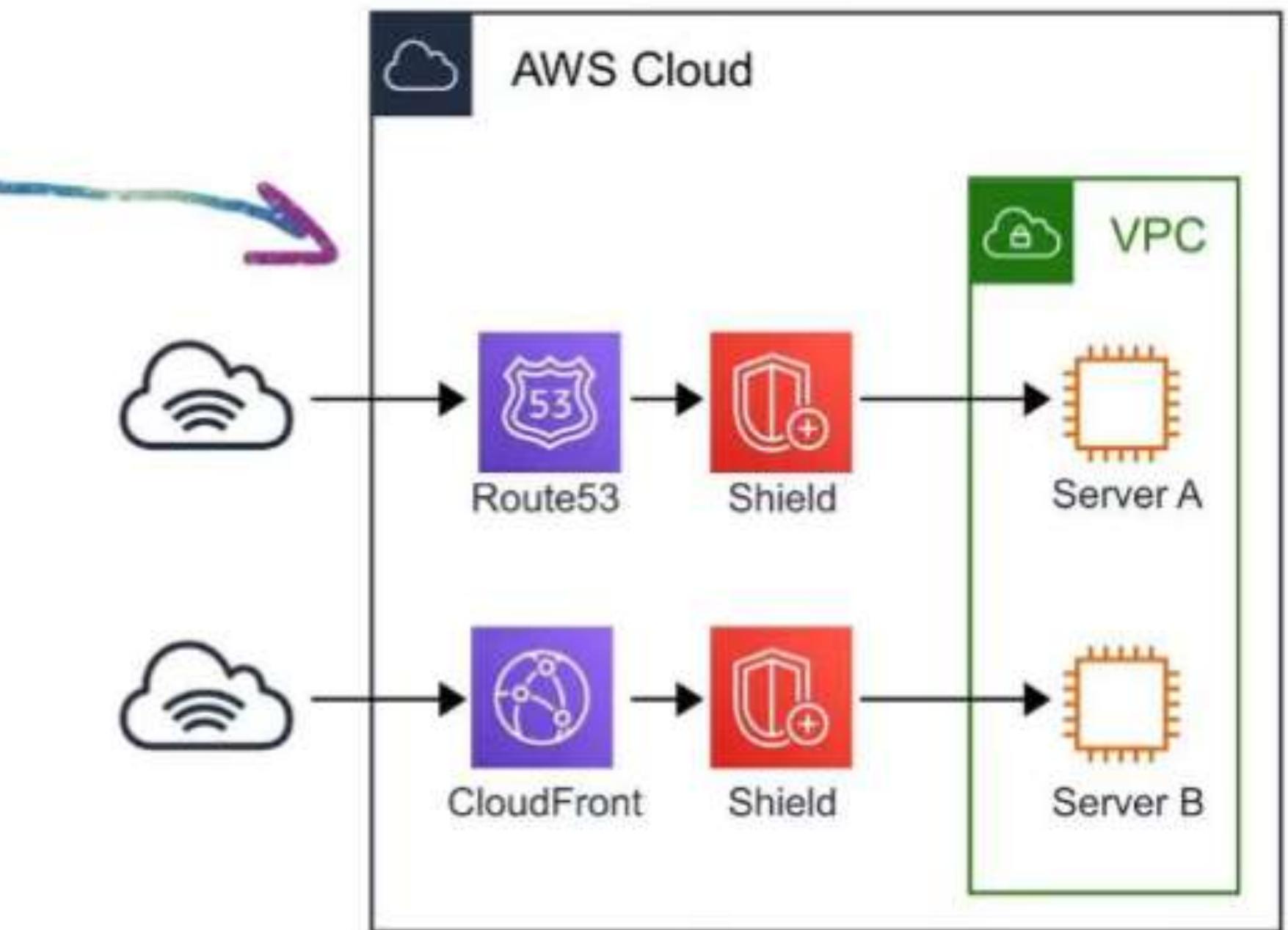


AWS Shield is a **managed** DDoS (Distributed Denial of Service) protection service that safeguards applications running on AWS

When you route your traffic through **Route53** or **CloudFront** you are using **AWS Shield Standard**

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

- 7 Application
- 4 Transport
- 3 Network



AWS Shield

Shield Standard FREE

protection against most common DDoS attacks

- access to tools and best practices to build a DDoS resilient architecture.
- Automatically available on all AWS services.

Shield Advanced *3000 USD / Year

additional protection against larger and more sophisticated attacks

Available On

- Amazon Route 53
- Amazon CloudFront
- Elastic Load Balancing
- AWS Global Accelerator
- Elastic IP (Amazon EC2 and Network Load Balancer)

Notable Features

- Visibility and Reporting on Layer 3,4 and 7
- 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

Amazon Guard Duty

What is IDS/IPS?

Intrusion Detection System and Intrusion Protection System.

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 unauthorized behavior. It uses Machine Learning to analyze the following AWS logs:

- CloudTrail Logs
 - VPC Flow Logs
 - DNS logs

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

The screenshot shows a findings detail page for a specific event. The event details are as follows:

Policy: IAMUser/RootCredentialUsage	
Finding ID:	dcbe0ca20e68085ad8a0c8e049659217
Severity:	Low
Region:	us-east-1
Count:	36
Account ID:	123456789012
Resource ID:	No information available
Created at:	09-24-2021 15:24:26 (a month a...)
Updated at:	09-24-2021 16:59:21 (a month a...)

A red arrow points from the text "Investigate with Detective" in the original slide to the "Investigate with Detective" button in the screenshot.



Amazon Guard Duty

aws Services Resource Groups ▾

GuardDuty

Findings

Settings

Lists

Accounts

What's New •

Usage

Partners

New feature: New Root Credential Detection

Amazon GuardDuty has added a new finding type that notifies you when root credentials are used programmatically in your account. [Learn more](#)

Findings

Showing 33 of 33

Actions Saved filters / Auto-archive [No saved filters](#)

Current Add filter criteria

<input type="checkbox"/>	Finding type	Resource	Last s...	Co...
<input type="checkbox"/>	Recon:EC2/PortProbeUnprotected...	Instance: i-05e8996590e85b1b...	a mon...	280
<input type="checkbox"/>	Recon:EC2/PortProbeUnprotected...	Instance: i-0d89ad53f4d6f3f94...	a mon...	330
<input type="checkbox"/>	Recon:EC2/PortProbeUnprotected...	Instance: i-04fae5b8df570e6ce...	a mon...	310
<input type="checkbox"/>	Recon:EC2/PortProbeUnprotected...	Instance: i-0269e117c812c22fd...	a mon...	253
<input type="checkbox"/>	UnauthorizedAccess:EC2/SSHBrut...	Instance: i-0269e117c812c22fd...	a mon...	1
<input type="checkbox"/>	UnauthorizedAccess:EC2/SSHBrut...	Instance: i-04fae5b8df570e6ce...	a mon...	1
<input type="checkbox"/>	UnauthorizedAccess:EC2/SSHBrut...	Instance: i-0269e117c812c22fd...	a mon...	1
<input type="checkbox"/>	UnauthorizedAccess:EC2/SSHBrut...	Instance: i-04fae5b8df570e6ce...	a mon...	2
<input type="checkbox"/>	UnauthorizedAccess:EC2/SSHBrut...	Instance: i-05e8996590e85b1b...	a mon...	35
<input type="checkbox"/>	UnauthorizedAccess:EC2/SSHBrut...	Instance: i-05e8996590e85b1b...	a mon...	5
<input type="checkbox"/>	UnauthorizedAccess:EC2/SSHBrut...	Instance: i-05e8996590e85b1b...	a mon...	1

Useful?

UnauthorizedAccess:EC2/SSHBruteForce

Finding ID: [bab43ae3b9a5cf5c032aab5f0914a468](#)

76.72.169.18 is performing SSH brute force attacks against i-04fae5b8df570e6ce. Brute force attacks are used to gain unauthorized access to your instance by guessing the SSH password.

Severity	Region	Count
Low	us-east-1	1
Account ID	Resource ID	Created at
655604346524	i-04fae5b8df570e...	01-22-2019 08:37...
Updated at		
01-22-2019 08:37...		

Resource affected

Resource role	Resource type
TARGET	Instance
Instance ID	Port
i-04fae5b8df570e6ce	22
Port name	Instance type
SSH	t2.small
Instance state	Availability zone
running	us-east-1a
Image ID	Image description
ami-06aa276f0e7597475	Agent Installed, Bundle -no-de...
Launch time	
01-10-2019 12:51:45	

Amazon Macie



Macie is a fully managed service that continuously monitors **S3 data access** activity for anomalies, and generates detailed alerts when it detects risk of unauthorized access or inadvertent data leaks.

Macie works by uses Machine Learning to Analyze your CloudTrail logs

Macie has a variety of alerts

- Anonymized Access
- Config Compliance
- Credential Loss
- Data Compliance
- File Hosting
- Identity Enumeration
- Information Loss

- Location Anomaly
- Open Permissions
- Privilege Escalation
- Ransomware
- Service Disruption
- Suspicious Access

Macie's will identify your most at-risk users which could lead to a compromise



AWS Virtual Private Network (VPN)



AWS VPN lets you establish a **secure** and **private tunnel** from your network or device to the AWS global network

AWS Site-to-Site VPN

securely connect on-premises network or branch office site to VPC

AWS Client VPN

securely connect users to AWS or on-premises networks

What is IPsec?

Internet Protocol Security (IPsec) is a secure network protocol suite that authenticates and encrypts the packets of data to provide secure encrypted communication between two computers over an Internet Protocol network. It is used in virtual private networks (VPNs)



AWS WAF



AWS Web Application Firewall (WAF) protect your web applications from common web exploits

Write your own **rules** to ALLOW or DENY traffic based on the contents of an HTTP requests

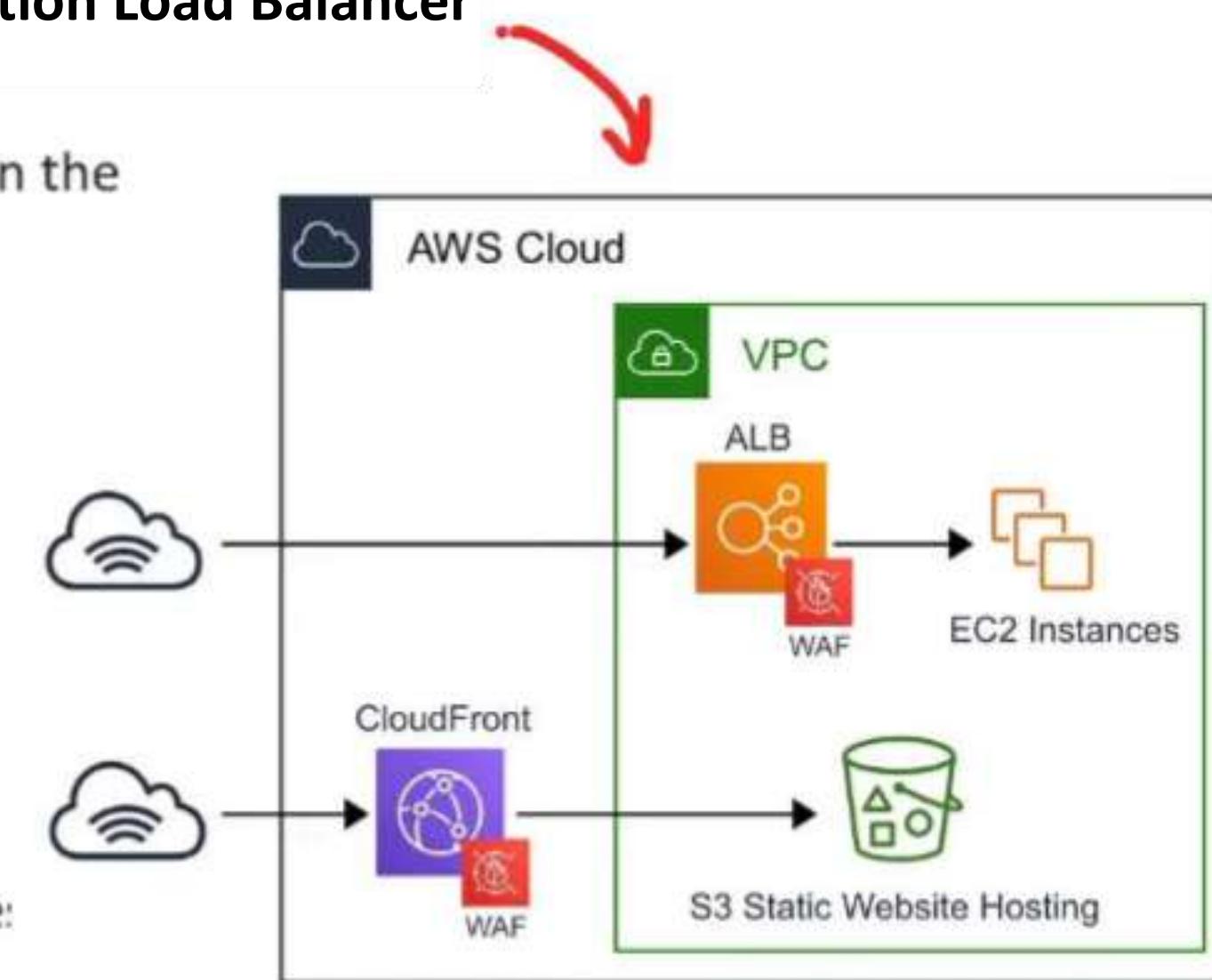
Use a **ruleset** from a trusted AWS Security Partner in the AWS WAF Rules Marketplace

WAF can be attached to either **CloudFront** or an **Application Load Balancer**



Protect web applications from attacks covered in the **OWASP Top 10** most dangerous attacks:

1. Injection
2. Broken Authentication
3. Sensitive data exposure
4. XML External Entities (XXE)
5. Broken Access control
6. Security misconfigurations
7. Cross Site Scripting (XSS)
8. Insecure Deserialization
9. Using Components with known vulnerabilities
10. Insufficient logging and monitoring

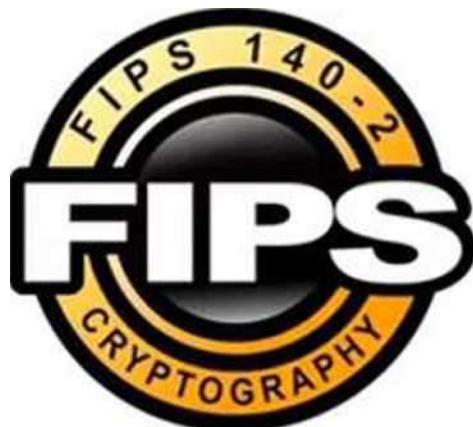


Hardware Security Module (HSM)

A Hardware Security Module (HSM).

Its a piece of hardware designed to store encryption keys.

HSM hold keys in memory and never write 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** are **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)



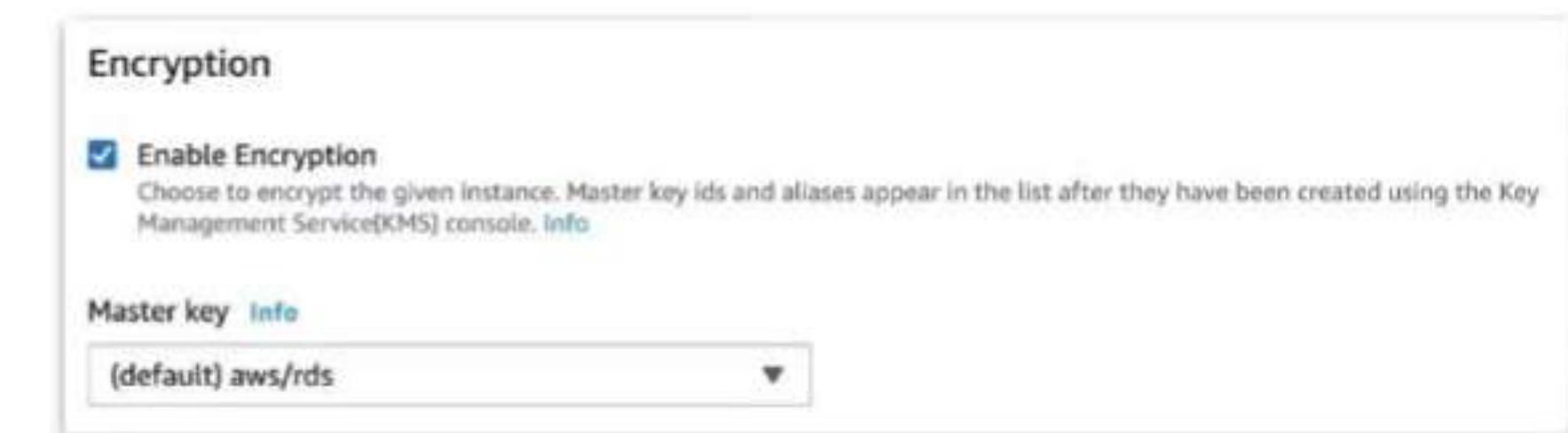
eg. AWS CloudHSM

AWS Key Management Service



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.

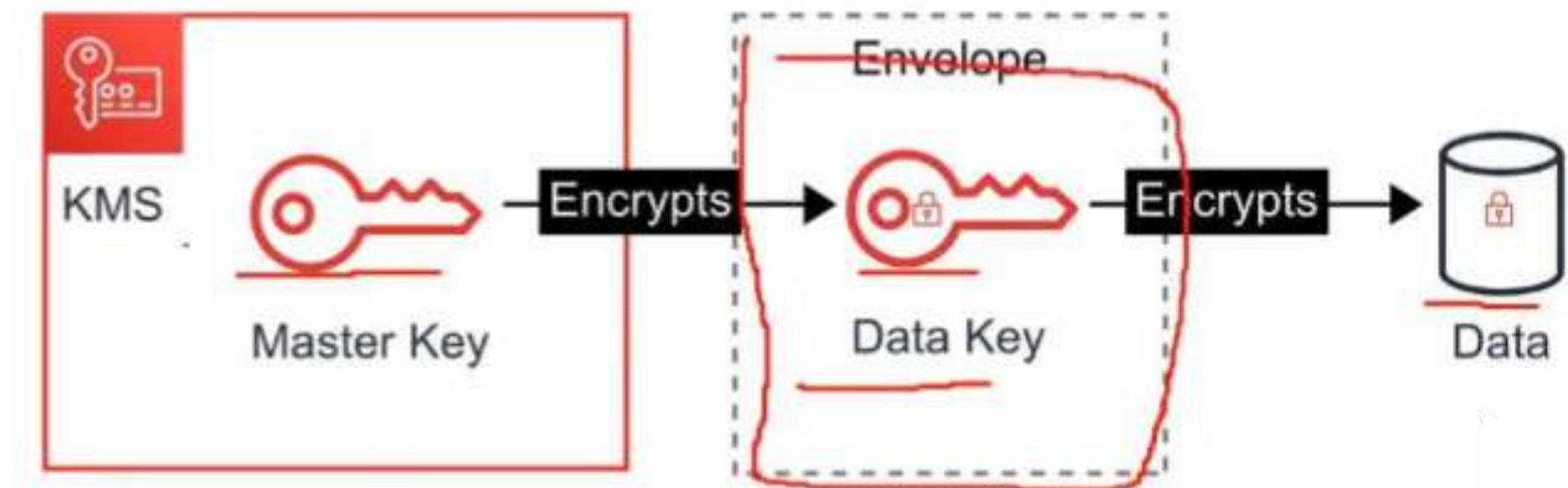
- KMS is a multi-tenant HSM (hardware security module)
- Many AWS services are integrated to use KMS to encrypt your data with a simple checkbox
- KMS uses Envelope Encryption.



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.

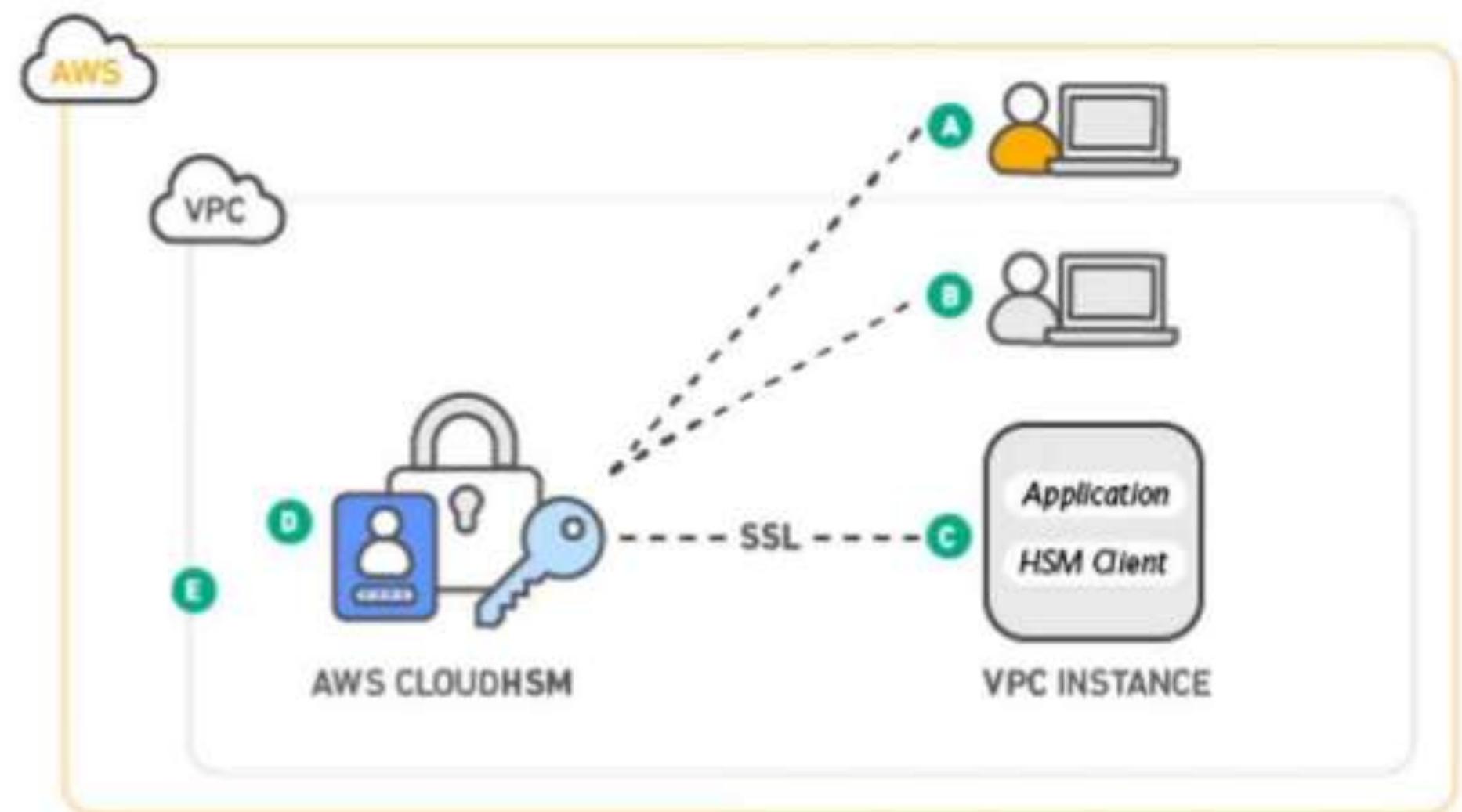
AWS CloudHSM enables you to generate and use your encryption keys on a FIPS 140-2 Level 3 validated hardware.

Built on Open HSM industry standards to integrate with:

- PKCS#11
- Java Cryptography Extensions (JCE)
- Microsoft CryptoAPI (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.



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

MKS Managed **Kafka** Service

RAM AWS Resource Manager

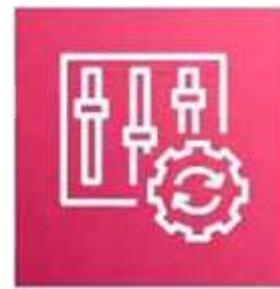
ACM Amazon Certificate Manager

PoLP Principle of Least Privilege

IoT Internet of Things

RI Reserved Instances

AWS Config vs AWS AppConfig

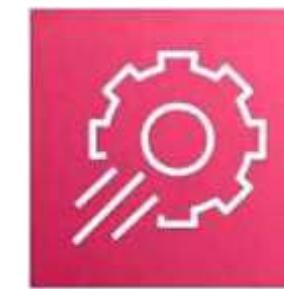


AWS Config

AWS Config is a governance tool for Compliance as Code (CoC).

You can create rules that will check to see if resources are configured the way you expect them to be.

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



AWS AppConfig

AWS App Config is used to automate the process of deploying application configuration variable changes to your web-application(s).

You can write a validator to ensure the changed variable will not break your web-app

You can monitor deployments and automate integrations to catch errors or rollback.

SNS vs SQS

They both Connect Apps via Messages



Simple Notifications Service

Pass Alongs Messages eg. PubSub

Send notifications to **subscribers** of **topics** via multiple protocol, eg, HTTP, **Email**, SQS, SMS

SNS is generally used for sending **plain text emails** which is 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



PubNub



Simple Queue Service

Queue Up Messages, Guaranteed Delivery

Places 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

Really good for delayed tasks, queueing up emails

RabbitMQ



SNS vs SES vs Pinpoint vs Workmail



Simple Notifications Service Practical and Internal Emails

Send notifications to **subscribers** of **topics** via multiple protocol, eg, HTTP, **Email**, SQS, SMS

SNS is generally used for sending **plain text emails** which is 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**



Simple Email Service Transactional Emails

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



Amazon PinPoint Promotional Emails

Emails for marketing

- Segment your contacts
- Create email campaigns
- Create customer journeys via emails
- A/B emailing testing



Amazon Workmail Email Web Client

Similar to Gmail and Outlook. Create company emails, read, write and send emails from a Web Client within AWS Management Console

Amazon Inspector vs AWS Trusted Advisor

Both are **security tools and they both perform audits**



Amazon Inspector

Audits **a single EC2 instance** that you've selected

Generates a report from a long list of security checks i.e 699 checks.



Trusted Advisor

Trusted Advisor **doesn't generate out a PDF report**.

Gives you a **holistic view** of recommendations across multiple services and best practices

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



Direct Connect

- A Dedicated Fiber Optics Connection from your DataCenter to AWS
- Intended for large enterprises with their own datacenter and they need an insanely fast and *private* connection directly AWS.
- If you need a **secure connection** you need apply a 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, setup automated phone systems.
- Interactive Voice System (IVS)



Media Connect

- New Version of Elastic Transcoder, Converts Videos to Different Video Types
- You have 1000 of videos you and you need to transcode them into different videos format, maybe you need to apply watermarks, or insert introduction video in front of every video

Elastic Transcoder vs MediaConvert

Both services **transcodes videos**



Elastic Transcoder

The Old Way

Elastic Transcoder was the original transcoding service. It may have programmatic APIs or workflows not available in MediaConvert.

It exists due to legacy customers still using the platform

- Transcodes videos to streaming formats



AWS Elemental MediaConvert

The New Way

MediaConvert is a more robust transcoding service that can perform various operations during transcoding.

- Transcodes videos to streaming formats
- Overlays images
- Insert video clips
- Extracts captions data
- Robust UI

AWS Artifact vs Amazon Inspector

Both Artifact and Inspector **compile out PDFs**



AWS Artifact

Why should an enterprise trust AWS?

Generates a security report that's based on **global compliance frameworks** such as:

- Service Organization Control (SOC)
- Payment Card Industry (PCI)



Amazon Inspector

How do we know this EC2 instance is Secure? Prove It?

Runs a script that analyzes your EC2 instance, then generates a PDF report telling you which security checks passed.

Audit tool for security of EC2 instances

ELB vs ALB vs NLB vs GWLB vs CLB



Elastic Load Balancer (ELB) has 4 different types of possible load balancers.



Application Load Balancer (ALB)

Layer 7 - HTTP/S Routing Rules

- create rules to change routing based on information found in a HTTP/S

request

Can attach an AWS WAF



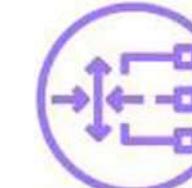
Network Load Balancer (NLB)

Layer 3 and 4 - TCP and UDP

Where extreme performance is required for **TCP and TLS traffic**

Capable of handling millions of requests per second while maintaining **ultra-low latencies**

Optimized for **sudden and volatile traffic** patterns while using a single static IP address per Availability Zone



Gateway Load Balancer (GWLB)

When you need to deploy a fleet of third-party virtual appliances that support GENEVE



Classic Load Balancer (CLB)

Layer 3,4 and 7

Intended for applications that were built within the **EC2~Classic network** Doesn't use Target Groups Retires on Aug 15, 2022