Cloud Computing

What Is Cloud Computing?

Cloud Computing is the on-demand delivery of IT resources, particularly compute power, application hosting, database application, networking, and more.

How Does It Work?

To create resources in the cloud, the client (customer) sends a request to the cloud to create/modify/delete resources.

*It follows a **Client-Server model**, with the cloud acting as the "server".*

Cloud Computing Models

CLOUD	HYBRID	ON-PREMISES
- Utilized heavily by startups of all sizes - All aspects of the application are hosted in the cloud - All existing applications are migrated to the cloud - New applications are built in the cloud only	- Some parts of the application are run on the cloud, while other parts are run on-premise - Some existing applications may be migrated to the cloud, while others remain on-premise - Most new applications are designed and built for the cloud - Fast connection between on-premise and cloud resources	- Cloud is not used at all - All applications are run on their own data centers or on those that are rented out - Responsible for all security and operation - Used by established companies that haven't had a reason to move to the cloud - Infrastructure already in place to manage the infrastructure themselves - Companies that need strict control and security over the entire infrastructure

AWS

A - Amazon
W - WEB
S - Services

What Is AWS?

- Amazon Web Services (AWS) was the first hyper-scale **cloud computing platform released in 2006.**
- AWS allows users to access powerful computing resources, such as servers, storage, databases, and many more, without the need to own or physically operate the infrastructure.
- AWS can be used to host websites and web applications, store data or files, and process large sets of data, among other things.
- AWS is designed to be **highly scalable**, **flexible**, and **cost-effective** so that businesses of all sizes can benefit from cloud computing at affordable pricing.

AWS Core Service Categories

AWS services can be broken down into six major categories.

CATEGORY	DESCRIPTION
Compute	Servers are used to run applications.
Networking and Content Delivery	These are services for managing networking in the cloud.
Storage	These are services used to store data.
Databases	These are services that manage databases.
Security, Identity and Compliance	These handle the security of your AWS infrastructure.
Management and Governance	These ensure that AWS infrastructure is following best practices and meeting regulatory requirements.

Interacting With AWS

WEB GUI	Command Line Interface	Programmatically
Console	AWS CLI	AWS SDK
Beginners	Engineer	Developers

Benefits of the Cloud

- CAPEX can be converted to OPEX
    a) Upfront expenditures are referred to as capital expenditures (CAPEX).
    b) Day-to-day expenses are referred to as operational expenses (OPEX).
    c) In the cloud, you trade upfront expenses for variable day-to-day expenses; therefore, **you pay for only what you use and give back what you don't need.**
- In the cloud, you do not have to predict your future capacity.
- Allows faster deployment of applications
    - Companies don't have to order physical equipment and have it installed for scaling up or upgrading according to requirements.

- High availability and low latency
    - With AWS, you can access their entire global infrastructure and deploy applications on any of their sites with just a click of a button.

Cloud Economics

AWS has five different pricing models, so you can select the one that saves you on cost for your specific workload.

Free Tier	On-Demand	Reserved	Volume Discounts	Price Drops
- Over 100 services available for free - 12 months of free service - Some services are always free	- Pay for what you use or the size you request	- If you know you will be using a service for a long time, you can reserve it ahead of time (for 1-3 years) to save on cost	- Like most things in the world, when you buy more, the price per unit goes down	- AWS drops prices on services fairly regularly - There have been 129 price drops from 2006 to early 2023

Cloud Native Design principles

PRINCIPLE	DESCRIPTION
Design for Failure	- No single point of failure – no single component or location going down should take down your entire application - Add redundancy as much as possible
Decouple Components	- AWS offers Simple Queue Service (SQS) that allows you to move data between different components When components need to communicate with one another, they'll send messages through the queue - Allows you to have individual components go down without loss of data
Implement Elasticity	- Make sure your application and all its components can scale up and down as load varies

Think Parallel	- Have multiple instances running concurrently to finish tasks as quickly as possible

Security and Compliance

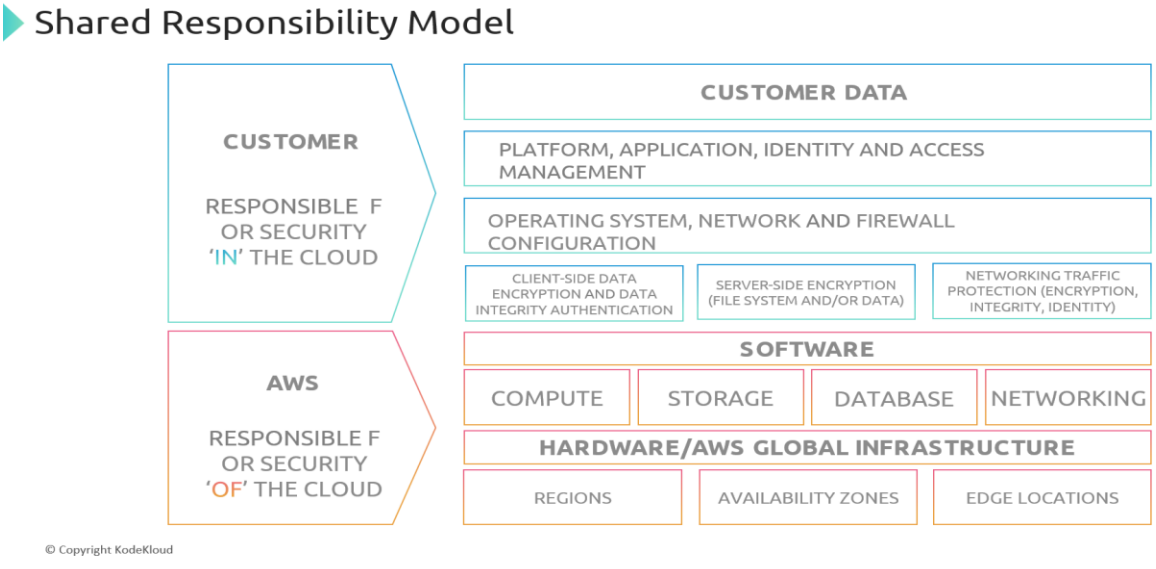
AWS Shared Responsibility Model

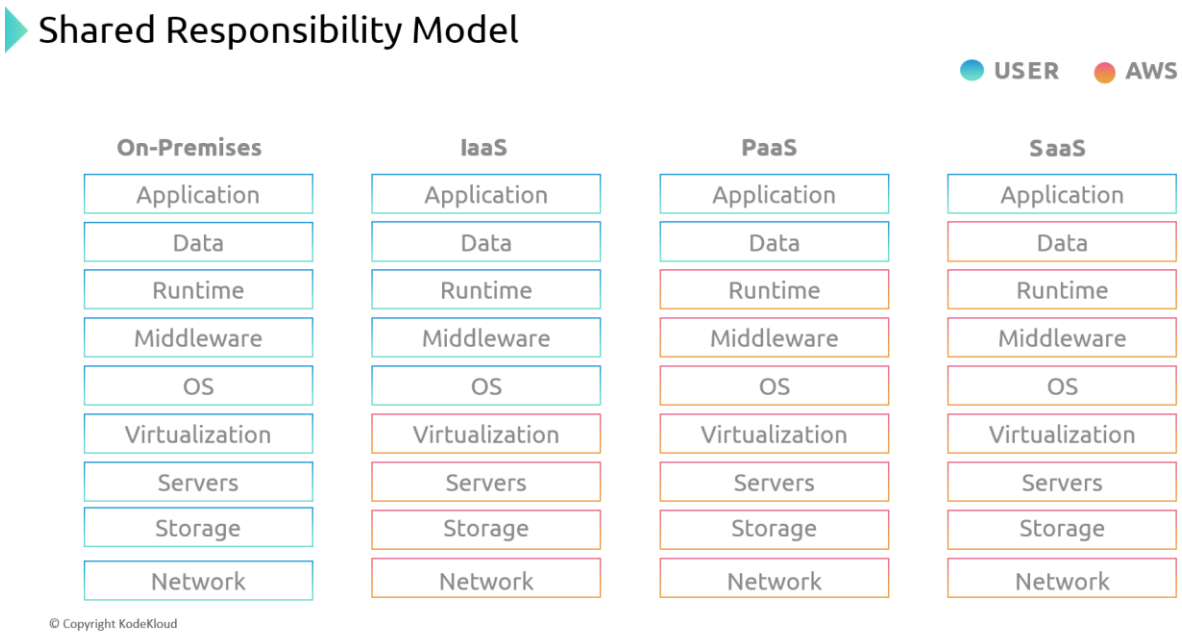
In traditional non-cloud-based deployments, all aspects of security are owned by you

1. Securing data center
2. Securing network and connectivity
3. Securing servers
4. Securing and patching operating systems
5. Securing application code so it isn't susceptible to exploits and vulnerabilities

Security in the cloud is a team effort between you and Amazon

The **Shared Responsibility model** outlines the role that AWS and you, as the customer, play when it comes to security. (Who is responsible for what?)





Note:

- ➢ Unmanaged services need to be secured by users.
- ➢ Managed services offload some of the security responsibilities of a service on AWS.

Compliance

1. Organizations in specific industries must adhere to certain rules and guidelines specific to that industry (**Finance, Health, Federal Government**).
2. Compliance and regulatory frameworks are sets of guidelines and best practices. Organizations follow these guidelines to meet regulatory requirements, improve processes, strengthen security, and achieve other business goals.
    - **Healthcare industry – HIPAA/HITECH**
    - **Payment card industry – PCI DSS**
3. Compliance is a **shared responsibility** between customers and AWS.
4. AWS undergoes certifications, reviews, and audits by various governing bodies.
5. These audit reports are made available to customers using **AWS Artifact.** Artifact allows customers to review and accept agreements to maintain compliance.

AWS Compliance Center

- AWS Compliance Center is a central location to research cloud-related regulatory requirements and assess their impact on your industry
    1. Identify regulatory requirements

2. Browse country-specific lAWS/requirements
3. Discover how companies in various industries solve compliance, governance challenges
4. Use AWS to answer key compliance questions
5. Get an audit and security checklist
6. Reference architectures with best practices

- AWS Audit Manager continuously collects data to prepare for audits and ensures that you are achieving compliance with regulatory standards. It helps build audit-ready reports.

- It tracks how the resource is configured and records the previous configuration states, so you can see how the configs for it have changed over time.

Identity Access Management (IAM) Users, Groups, and Roles

➢ New AWS accounts have a single user called "root user" created automatically.
- Grants full permissions to do anything in the account
- Not recommended to log in as the root user

➢ Identity Access Management (IAM) is responsible for managing access to AWS resources.
- Responsible for authenticating users and determining what they are authorized to do

➢ IAM has three types of identities
- Users
- Groups
- Roles

➢ An IAM user represents a person or application that needs access to AWS or a subset of services
- An employee that needs access to AWS will have a user created for them
- An application that needs to access or interact with AWS will have a user created on their behalf

➢ New users by default do not have access to anything in AWS
- Users have to be granted explicit permissions to access specific services/resources
- Users are implicitly denied all permissions by default

➢ To grant users access to resources, IAM policies need to be applied to the user, giving them permissions.

- ➢ Policies are documents that either grant or deny access to specific AWS services/resources.
- ➢ IAM policies define what resources a user/group/role can access and what actions they can perform on them.
- ➢ Policies can be assigned to multiple users.
- ➢ Users can have multiple policies assigned to them.
- ➢ It is a best practice to follow least-privilege permissions.
    - Identities/users should only be granted the minimum permissions necessary for them to perform their job.
    - No extra permissions should be given to them.
    - If a user should only be able to stop and start EC2 instances,
        a) Allow only stopping/starting of EC2; don't allow them to create/delete/modify instances
        b) Don't allow them to have permission to do anything with any other services
- ➢ Groups are a collection of IAM users.
    - Policies can be added to groups
    - IAM users within a group automatically inherit all the policies from the group
    - Example – having a separate group for each department
- ➢ Roles allow users, applications, or services to assume temporary permissions.
    - Roles are assigned permissions/policies just like users and groups
    - When someone assumes an IAM role, they inherit the permission of the role temporarily and return to their original permission when done
- ➢ IAM roles are recommended when there is a need to grant temporary access to a service.
- ➢ Multi-Factor Authentication (MFA) requires users to provide an extra security code from an MFA device/app to be able to log in to their account
    - MFA is enabled on a per-user basis
    - It is a best practice to enable MFA for all users

Organizations

- ➢ Organizations help manage multiple AWS accounts
- ➢ Organizational units (OUs) allow you to group accounts with similar business or security requirements
- ➢ Service Control Policies (SCPs) restrict what an account can do
    - SCPs can be applied to individual accounts or OUs
    - When applied to OUs, all AWS accounts within the OU inherit the policies

Security Resources

Web Application Firewall (WAF)

Monitors HTTP requests and prevents a variety of attacks like SQL injection and XSS attacks

Protects the following AWS resources

1. CloudFront distributions
2. API Gateway
3. Application Load Balancer
4. AWS AppSync GraphQL API
5. Amazon Cognito user pool
6. AWS App Runner service
7. AWS Verified access instance

Rules are defined as web access control lists (web ACL)

1. The resource you want to monitor
2. Inspection criteria (IP address, country of origin, size of request, malicious code)
3. Action (Allow, Block, Count, Captcha)

AWS Shield

AWS Shield detects and mitigates sophisticated distributed denial of service (DDoS) attacks.

A DDoS attack is when a perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting the services of a host connected to a network.

- Denial of service is typically accomplished by flooding the targeted machine or resource with fake requests to overload systems and prevent some or all legitimate requests from being fulfilled.

AWS Network Firewall

It is a stateful managed firewall and intrusion detection and prevention service for VPCs. It monitors traffic going into and out of a VPC.

AWS Inspector

AWS inspector scans workloads running on AWS for software vulnerabilities and undesired network exposure.

- It automatically discovers and scans EC2 instances, container images in ECR, and Lambda function.

- It continues to assess your environment throughout the lifecycle of your resources by automatically rescanning resources in response to changes that could introduce new vulnerabilities.
    1. Installing new package
    2. Installing a patch
    3. New common vulnerabilities and Exposures (CVE) are released

- If vulnerabilities are identified, AWS Inspector produces a finding that you can investigate.

Amazon Detective

Amazon Detective helps you quickly analyze and investigate security events across one or more AWS accounts.

1. It ingests data from VPC flow logs, CloudTrail logs, and GuardDuty findings.
2. It uses machine learning and statistical analysis to create advanced visualizations that show resource behaviors and interactions over time.

Amazon Detective simplifies the process of investigating security incidents by providing interactive dashboards and visualizations.

CloudTrail

CloudTrail tracks user activity with an AWS account.

- It tells us who is doing what.
- Anytime a user performs an action within AWS, an event is logged in CloudTrail, such as:
    1. Logging in
    2. Modifying security policies
    3. Accessing/creating/deleting a service
- It logs actions taken in the console, command line, and AWS SDK/APIs.

AWS Security Hub

It automates security checks and brings security alerts into a central location:

- Ingests data from other services like AWS Config, GuardDuty, Firewall Manager, Inspector, etc.
- Performs validation against AWS/Security best practices

Security Lake

It collects security logs and events from multiple sources including on-premises, AWS services, and third-party services. It transforms the data into storage and query-efficient Parquet format.

Firewall Manager

It helps manage security services (WAF, security groups, network firewall) across multiple AWS accounts. It configures rules just once and has them available across all accounts.

Resource Access Manager

It helps you securely share resources across accounts, organizations, and OUs. So, you can create a resource once and have the Resource Access Manager make that resource usable by other accounts.

AWS Cognito

AWS Cognito helps implement customer identity and access management for mobile and web applications.

- It manages all user credentials.
- It makes adding signup/login functionality easy without social authentication on any app.

IAM Identity Center

It simplifies managing users across multiple AWS accounts. It also allows you to manage sign-in security for your users centrally and grant them access across all AWS accounts and resources.

Secrets Manager

The Secrets Manager helps retrieve and rotate secrets and other sensitive data like credentials and Auth tokens.

- No need to hard code credentials in application source code
- Makes it more difficult to leak or compromise sensitive credentials
- Can configure automatic rotation schedule for your secrets
- Having short-term secrets decreases the risk of compromise

AWS Certificate Manager (ACM)

It handles the complexity of creating, storing, and renewing public and private SSL/TLS certificates and keys that protect AWS websites and applications.

- If users need to communicate with an ELB or API gateway, you now no longer need to purchase a paid certificate from a third party.
- ACM can generate them for you, and end-user traffic will be encrypted.
- You can issue certificates directly from ACM or import third-party certificates.

AWS Private Certificate Authority

It provides users with a private CA that is managed by AWS.

- Issues certificates for authenticating internal users, computers, and applications

- Certificates issued by a private CA are trusted only within your organization and not on the internet
- Saves the hassle of having to set up, configure, and maintain your own internal certificate authority

Key Management Service (KMS)

KMS helps create and manage cryptographic keys used for encrypting/decrypting data.

- It allows you to provide granular control over who has access to the keys
- Key rotation can be configured

CloudHSM

CloudHSM provides customers with a hardware security module in the cloud.

- All cryptographic keys are securely stored on the HSM, and they never exit the device
- Data is sent from the clients to the HSM for encryption/decryption
- Because all keys are securely stored on the HSM in a central location, it helps minimize the risk of keys getting leaked or stolen

Billing

General Billing

There are three main drivers of billing:

- **Compute** that was used/requested
  - ➔ I want a server with 4 CPUs and 32 GB of RAM
  - ➔ That server was run for 8 hours
- **Storage** that was used/requested
  - ➔ I want 50 GB of fast disk storage
- **Network** that was used/requested (only in the outbound direction)
  - ➔ I transferred 20 GB of data out to my office

- Ensure to understand all the aspects of pricing for a service
- Choose the correct sizing of service to optimize costs
- Make use of AWS' Optimize and Save tools when possible

- Scale up only when needed and make sure to scale back down
    - ➔ Utilize auto-scaling when possible

AWS Pricing Models

Free Tier	On-Demand	Reserved	Savings plan	Spot
- Over 100 services available for free - 12 months of free service - Some services are always free	- Pay for what you use or the size you request	- If you know you will be using a service for a long time, you can reserve it ahead of time (for 1-3 years) to save on cost	- Offers discounted prices on services in exchange for a commitment to spend a certain period of time	- If Amazon has spare capacity, they'll offer it at a discounted rate

EC2 Billing

EC2 pricing factors in the following:

What is the size of the virtual machine (How many vCPUs and how much memory?)
Are the charges per second or per hour? (How long did it run?)
EC2 licensing type
What features are turned on?
Is the machine running or stopped? (It will still incur a small cost when stopped.)

On-Demand	Pay for the compute capacity by hour or second; no long-term commitments.
Savings plan	For a commitment to spending a certain amount over a 1-3 year period, AWS will offer a discounted rate.
Spot	When AWS has spare compute capacity, they offer it at a discounted rate.
Reservation	Reserve capacity for EC2 instances in advance at a discounted rate.
Dedicated host	Physical server dedicated exclusively to your account. EC2 instances will always be launched on the same physical host.
Dedicated Instance	EC2 instances will always run on dedicated hardware reserved for your account, but the same server may not be running every time.

RDS Billing

RDS has several cost factors:

Which flavor of RDS is used? (Aurora? Aurora Serverless?)
What type of SQL engine is used? (Oracle, MSSQL, MariaDB, MySQL, or PostgreSQL?)
What is the memory size of the database?
What Storage Disk type is used? (General purpose or provisioned IOPS?)
What additional features are enabled, for example, Multi-AZ or backup retention?
RDS also has Reserved instances like EC2 but does not have the spot, dedicated, or savings plan.

VPC Billing

VPC costs include the following:

Charge per VPC and their base components
Data transfer charges
Add-ons to VPCs

- Data transfer charges are only outbound, not inbound
  - Includes leaving one region to go to another region
  - Does not include data moving from an EC2 instance to an S3 bucket in the same region
- Traffic going to **different regions or AZ or public IP** means you will get charged
- **NAT gateways** charge you for the existence of the object
- **No charges** for subnets, Network ACLs (NACLs), Security Groups, or IP ranges (network constructs incur no charges)

Lambda Billing

Lambda pricing is based on

Size
Duration
Frequency

- The more often your functions run, the more you pay
- The longer it runs, the more you pay
- The more memory it uses, the more you pay
- Any additional features like hot provisioning will cost extra

Other Services

EBC pricing factors:

Volume – size and type over time
Snapshots – size over time
EBS Fast Snapshot Restores

EBS Direct APIs for Snapshots

S3 pricing factors:

Type of storage class
Number and size of objects stored
Type of requests made to S3
Charged for outbound data
Other backup and management features

DynamoDB cost factors:

Reading, writing, or storing data
Optional features
Charges based on read request units and write request units

CloudFront cost factors:

Charges based on amount of data taken from CloudFront
HTTP/HTTPS requests
Invalidation requests

Macie cost factors:

Amount of data that needs to be scanned
S3 charges like reading objects and listing buckets as Macie scans

Kinesis cost factors:

For how long is data stored?
How much data is stored?

Billing Account Structure

Single Account	Multiple Accounts
Receives a **single bill** for all resources in this account If you have a **second account**, the billing for that account is separate	One account acts as the **"payer"** account ➔ Receive a **single bill** from AWS for all accounts ➔ View detailed billing from each account ➔ **Apply** Reservation and Savings Plans across **all accounts**

Multiple Accounts Within an AWS Organization	Multiple Accounts in Control Tower

Works identical to consolidated billing ➔ One account acts as the “payer” account	Works identical to consolidated billing ➔ One account acts as the “payer” account

Tools for Billing

Billing Dashboard	Cost Explorer	Budgets	Cost and Usage Report	AWS Calculator (calculator.AWS)
- Gives a general overview of your AWS spending in the console - Can view bills and itemized costs for the bill	- Provides visualization for your billing data with charts and graphs - Can also forecast if historical data exists	- Allow you to set soft and hard limits on bills	- Provides the most detailed report - Breaks down costs by dimensions you can control - Gets published to an S3 bucket	- Used for areas that don’t have historical data or if you need an estimate

Technology

Deployment Methods

AWS Console	Web GUI used to manage AWS resources ❖ Great for people who want to visually see their infrastructure ❖ Ideal for monitoring logs, alerts, and metrics in nicely presented graphs ❖ Lots of menus, so provisioning resources will involve a lot of clicking and navigating
AWS CLI	Command line utility (CLI) for managing AWS resources ❖ Engineers naturally prefer working on command line ❖ Very easy to manage resources and commands can be copied and pasted ❖ Some settings/knobs on resources can only be toggled through the CLI
AWS SDK	Provides APIs in most programming languages to manage and interact with AWS ❖ Allows applications to create resources in AWS

Global Infrastructure

Regions are locations across the globe to which services can be deployed.

- ❖ All services are not available in all regions
- ❖ Pricing can be different between regions

Availability Zones (AZ) are isolated and independent data centers inside regions.

Edge Locations are smaller points of presence across the globe.

- ❖ Allows you to get services closer to end-users to minimize latency
- ❖ Limited services available; mainly used for CDN
- ❖ Services available – CloudFront, Route 53, AWS WAF

Local Zones are extensions of AWS regions located near users in select metropolitan areas.

- ❖ Have isolated infrastructure but are connected to parent AWS regions through high-bandwidth network links
- ❖ Provide a subset of services like EC2 and EBS

Networking

VPC - VPC or Virtual Private Cloud is a secure, isolated network segment hosted within AWS.

- Isolates computing resources within the cloud
  - ➔ Acts as a network boundary
- Gives customer full control of networking in the cloud
  - ➔ Subnetting (IP address)
  - ➔ Routing
  - ➔ Firewalls
  - ➔ Gateways
- Specific to a single region
- CIDR block – A range of IP addresses that resources in the VPC can use

Subnets - These are a group of IP addresses inside your VPC.

- Subnets reside within a single availability zone.
- The range of IP addresses must be within the parent VPCs' CIDR block.
- Subnets can be public or private to allow external access to resources within them.

Gateways

- **Internet gateway** allows subnets in a VPC to communicate with the internet and vice versa. Internet gateways determine whether a subnet is public or private.
- **NAT gateways** provide access to the internet for resources.
  - ➔ In NAT gateways, connections must be initiated from within the VPC
- **Virtual private gateways** enable secure access to private resources over the internet.
- **Direct Connect (DX)** is a direct connection into an AWS region that provides low latency + high speeds.

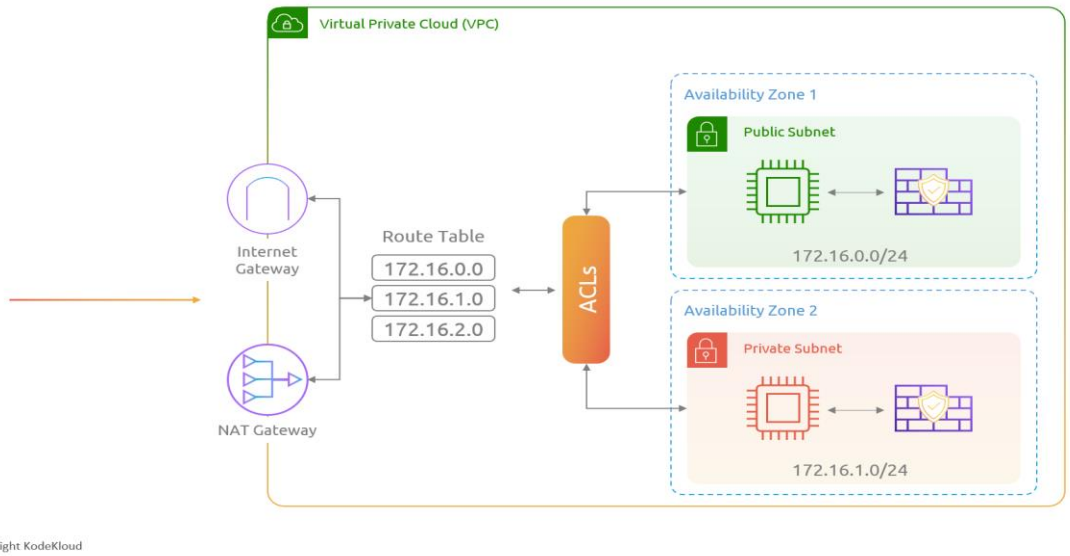
Default Networking

- Every region has a Default VPC with default subnets, security groups, and Network ACLs (NACLs).
- The CIDR block for Default VPC is 172.31.0.0/16.
- One default subnet exists in each Availability Zone (AZ).
- Default VPC and its subnets have outbound access to the internet by default.

- Security groups allow outbound traffic, and the NACLs are open in both inbound and outbound directions.
- Default subnets have access to an Internet Gateway for Internet connectivity.

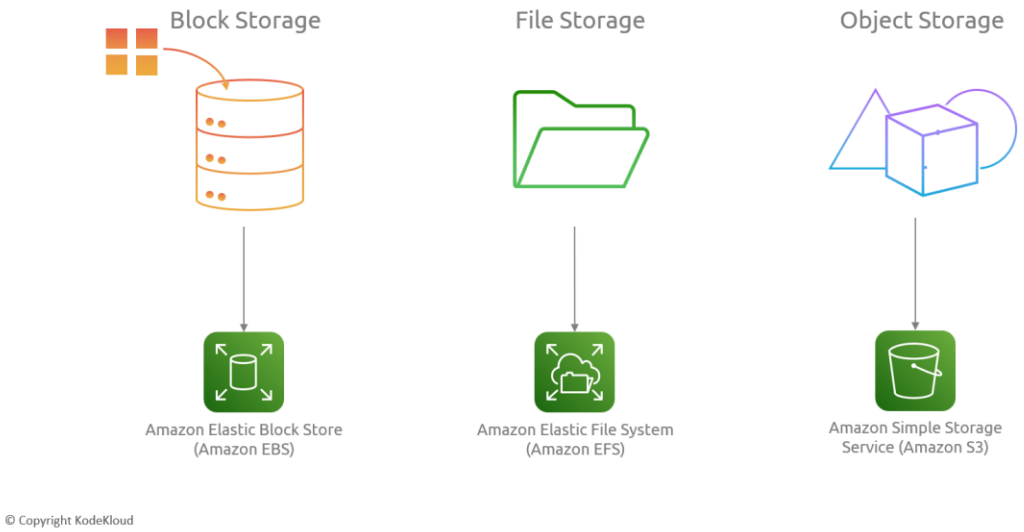
Firewalls

- Stateless firewalls require traffic to be explicitly permitted inbound and outbound.
- Stateful firewalls are intelligent firewalls that track requests and automatically allow the response.
- Network ACLs (NACLs) filter traffic entering and leaving a subnet.
- NACLs are stateless firewalls.
- Security groups are firewalls for individual resources such as EC2 instances, NICs, and other network objects.
- Security groups are stateful firewalls.



Storage

Types of Storage



Instance stores are temporary block-level storage, for instance.

- Data is lost if the instance is stopped and started

Block Storage (EBS) breaks up data into blocks and presents a collection of blocks as a volume or hard drive to the operating system.

- Can be mounted and booted (OS can be installed on it)
- EBS is Availability Zone specific
- Servers in one AZ cannot attach an EBS volume in a different AZ

File storage (EFS) stores data in a hierarchical structure of files and folders.

- Filesystem that is accessible remotely
- Multiple machines can connect to EFS volume at once
- Cannot be used as a boot volume (can't install OS)

Object storage

- Object storage (S3) stores objects (files) in a flat file structure
    - ➔ No sub-directories
    - ➔ Cannot boot or mount object storage
- Great for storing static websites, media files, logs, traces, and audit reports
- Storage classes impact accessibility, resiliency, and cost
    - ➔ **S3 Standard** – Default storage class and most expensive
    - ➔ **Standard-IA** – Same reliability as S3 standard but has a retrieval fee
    - ➔ **One Zone-IA** – Same as Standard-IA but only hosted in one AZ

- ➔ **Glacier instant** – Meant for files that are almost never used, but when they are needed, they should be retrieved instantly
- ➔ **Glacier flexible** – Cheaper than Glacier Instant, but data retrieval is time consuming
    - Bulk: 5-12 hours
    - Standard: 3-5 hours
    - Expedited: 1-5 minutes

- ➔ **Glacier Deep Archive** – Cheapest storage option; has the longest wait time for retrieval

Compute

Compute – EC2

- EC2 allows you to provision a server in AWS within minutes
- AMIs are templates for deploying EC2 instances
- AWS has a variety of different instance types
    - ➢ General Purpose – Good balance of compute, memory, and networking resources. It is the most versatile one.
    - ➢ Compute Optimized – Best for compute-heavy workloads
        - Contains high-performance CPUs
        - Great for batch processing workloads, media transcoding, and machine learning
    - ➢ Memory Optimized – Optimized for memory-intensive workloads like databases
    - ➢ Storage Optimized – Optimized for workloads that need high I/O operations per second (IOPS)
    - ➢ Accelerated Computing – Utilizes hardware accelerators (like GPUs) to perform expensive calculations
        - Great for graphics processing and data pattern matching
- Supports a wide variety of operating systems from RHEL, SUSE, Ubuntu, Amazon Linux, and Windows
- AWS also offers a variety of processors from ARM to AMD to Intel

Compute – Lambda

- AWS Lambda is a compute service that lets you run code without having to provision or manage servers.

- Lambda is AWS' serverless offering
    - ➢ Servers are still required to run the code, but you don't see them at all
    - ➢ Servers are managed completely by AWS

- ➢ AWS manages the server maintenance, scaling, capacity provisioning, and logging
- ➢ All you need to do is upload the code; AWS handles the rest

- Lambda use cases include
  - ➢ File processing
  - ➢ Stream processing
  - ➢ Web application
  - ➢ Mobile/Web backend

Lambda service includes three components

- Function – A traditional function in any programming language
- Trigger – An event that causes the function to run
  - ➢ File gets uploaded to S3
  - ➢ HTTP request to API gateway
  - ➢ CronJob
  - ➢ DynamoDB update
- Event Info – Information about the event that triggered the function
  - ➢ Gets passed to the function

Lambda Benefits

- ➢ No servers to manage
- ➢ Auto scale to handle traffic
- ➢ Pay only for what you use

Lambda Downsides

- ➢ No local state
- ➢ Can only run for 15 minutes; not good for long-running tasks
- ➢ Impacted by cold starts

Lambda Pricing Factors in

- ➢ Number of times function ran
- ➢ How long function ran for
- ➢ How much memory/CPU used

Compute – Containers

- Containers are a tool that allows you to package an application and all of the necessary files, libraries, and dependencies the application needs to run.

- Container orchestrators are the brains of a containerized environment.
  - ➢ Deploying containers across all available servers
  - ➢ Sending load-balancing requests to containers

- ➢ Providing container-to-container connectivity
- ➢ Restarting failed containers
- ➢ Moving containers when hosts fail

- ECS is a fully managed container orchestration service that helps manage and scale containerized applications.
  - ➢ AWS manages ECS, which handles all the orchestration
  - ➢ Containers run on EC2 instances or on Fargate

- ECS is specific to AWS only (vendor lock-in).

- Kubernetes is an open-source container orchestrator.

- Kubernetes cluster has two types of nodes
  - ➢ Control-plane nodes – Managers of the cluster; watch over the cluster and ensure the cluster is kept in a working state
  - ➢ Worker nodes – Responsible for actually running the containerized workloads

- EKS manages the control plane for you.

ECS	EKS
ECS is proprietary to AWS, so migrating to another cloud provider can be difficult.	EKS is Kubernetes, which is open-source and can be run on any platform. Remember, the more AWS services you configure your cluster to interact with, the harder it will be to move to another provider.
ECS has a simpler architecture, and a simpler API, and makes it easier to ramp up new team members.	Kubernetes is very complex and has a steep learning curve. With EKS, you'll have to learn Kubernetes and EKS-specific features. Kubernetes has a larger community. It has more support online. Offers more tooling like Helm/ Kustomize/ ArgoCD
ECS Pricing – No cost for control-plane, only pay for infrastructure running applications (EC2, EBS)	EKS Pricing – More expensive; you have to pay for control-plane and worker nodes.

Database

SQL Databases

➔ Ideal for structured data

- ➔ The shape (schema) of the data is predefined
- ➔ Follow ACID Properties (Atomicity, Consistency, Isolation, Durability), ensuring reliable processing of transactions
- ➔ Structured Query Language (SQL) is used for defining and manipulating data
- ➔ SQL databases are scaled vertically, which means increasing the horsepower (CPU, RAM, SSD) of the existing machine

NoSQL Databases

- ➔ Data is unstructured and schema-less
- ➔ Any type of data can be stored, and database entries can have a completely different shape from one another
- ➔ Base properties – Basically available, soft state, eventual consistency
- ➔ Horizontally scale – Adding more servers into the pool and distributing data and load

- Self-managed databases give you full control over the database.
- The burden of backing up the data, storing it in a highly available manner, and securing the data is solely your responsibility.
- It offers increased control and flexibility but has high operational overhead,
- Relational Database Service (RDS) is a managed database service for SQL databases (MySQL, MariaDB, PostgreSQL, Oracle, Microsoft SQL Server).
- AWS manages the database, the operating system, and the underlying EC2 instance running the database.
- AWS handles high availability (spreading across multiple AZs), replication, and backups.

Aurora is another managed database service (that only supports MySQL and PostgreSQL).

- ❖ Aurora is designed to run in the cloud, making it faster than RDS and more performant.
- ❖ Aurora also has a serverless option, which means it doesn't run on virtual machines (EC2).

RedShift is Amazon's fully managed data warehousing service.

- ❖ Database is designed for online analytics processing (OLAP)
- ❖ It is great for performing analysis and aggregation on large datasets (petabytes of data).
- ❖ Amazon manages the infrastructure of the database.
- ❖ RedShift has a serverless option as well.

DynamoDB – AWS flagship NoSQL system

- ❖ Great for fast writes and reads for loosely structured data

DocumentDB – MongoDB compatible NoSQL service

Keyspaces – Managed Cassandra database

Neptune – Graph Database

- ❖ Great for social networks and recommendation engines

ElastiCache – An in-memory cache that's compatible with Redis and Memcached

OpenSearch – AWS clone of ElasticSearch

Amazon Quantum Ledger Database (QLDB) – Fully managed ledger database that provides a transparent, immutable, and cryptographically verifiable transaction log

Timestream - Serverless time-series database

Application Integration

- **Application Integration** – Services in AWS that help another service, usually in relation to replicating messages or events or traffic
    - ❖ Commonly acts as a buffer or queue between different application components.

- **Simple Notification Service (SNS)** – Publish/Subscribe service
    - ❖ One component publishes a message, a topic, and any component listening for that topic will receive the message
    - ❖ Think of a topic as a news channel that people can tune into, to get update on events
    - ❖ SNS is used to duplicate message to multiple consumers
    - ❖ SNS does not persist messages – you can either listen for them or you miss out

- **Simple Queue Service (SQS)** – Messaging queue used to send, store, and receive messages between components
    - ❖ SQS can persist messages in the queue for days
    - ❖ FIFO or standard queues

- **Elastic Load Balancers (ELB)** – Used to direct traffic to multiple backend servers
    - ❖ Used to distribute workloads across multiple servers
    - ❖ Can also direct traffic to EC2, ECS, EKS, and Lambda workloads
    - ❖ Acts as a single point of contact and entry for application traffic
    - ❖ Send users to the load-balancer URL and let it do the rest

- **Auto Scaling** – Allows you to scale up or down a service depending on load
    - ❖ Usually have workload sit behind a load-balancer and have the workload autoscale
    - ❖ Many services support autoscaling like DynamoDB and EC2
- **AppFlow** – Used to integrate data without code
- **EventBridge** – Used to build event-driven applications
- **MQ** – Message broker service for ActiveMQ and RabbitMQ users
- **Step Functions** – For orchestrating Lambda functions

Management Services

Management services are services that help manage, provision, or optimize other services.

- **CloudFormation** – Infrastructure as code tool that allows you to create templates to provision services
- **OpsWorks** – Allows you to automate server configuration by providing a managed instance of Chef and Puppet
    - ❖ Chef/Puppet are automation platforms that allow you to generate server configs through code
- **Systems Manager** – A secure end-to-end management solution for resources (services) on AWS and on-premise environments
- **AWS Organizations** – Used to manage multiple AWS accounts
- **Service Catalog** – Allows you to provide CloudFormation and Terraform templates to customers who can use them to deploy resources they need
- **AWS Control Tower** – Helps you set up AWS Organizations in a secure best practice way, with auditing, logging, and compliance rules in place
- **CloudTrail** – A service that tracks and records all user and API activity in your AWS account

Migration Services

Management services are services that help with migrating services from on-prem to AWS

- **Migration Hub** – Allows you to centralize and see all migrations you have in place via AWS services
- **Snow Family** – Used to transfer data into AWS
    - Snowcone – most compact an portable device. Comes in SSD & HDD options

    - Snowball – Medium sized data transfer (80TB) comes in compute and storage optimized devices
    - Snowmobile – Exabyte-scale data migration device used to move large amounts of data to AWS
        - Migrate up to 100PB in a 45 foot long ruggedized shipping container
- **Online Data transfer**
    - FTP, SFTP, FTPS,
    - AS2 – send/receive messages into S3 backend
- **DataSync** – secure, online service that automates and acclerates moving data between on premises and AWS storage services
- **Application Discovery Service** - helps you plan cloud migration projects by gathering information about your on-premises data centers
- **Application Migation Service** – Does the actually moving of applications from on-premise into AWS
- **Database Migration Service** - is a managed migration and replication service that helps move your database and analytics workloads to AWS quickly, securely, and with minimal downtime and zero data loss
- **Elastic Disaster Recovery** - minimizes downtime and data loss with fast, reliable recovery of on-premises and cloud-based applications using affordable storage, minimal compute, and point-in-time recovery
- **Mainframe Modernization** – assists in migrating mainframe applications to the cloud

AWS has six methods/patterns for migration

- **Rehosting** – also known as “lift-and-shift" involves movign applications without changes
    - Allows companies to carryout migrations and scale quickly as possible
- **Replatforming** - make a few cloud (or other) optimizations in order to achieve some tangible benefit, but you aren’t otherwise changing the core architecture of the application
- **Refactoring** - involves reimagining how an application is architected and developed by using cloud-native features. Refactoring is driven by a strong business need to add features, scale, or performance that would otherwise be difficult to achieve in the application’s existing environment.
- **Repurchasing** - involves moving from a traditional license to a software-as-a-service model. For example, a business might choose to implement the repurchasing strategy by migrating from a customer relationship management (CRM) system to Salesforce.com
- **Retaining** - consists of keeping applications that are critical for the business in the source environment. This might include applications that require major refactoring before they can be migrated, or, work that can be postponed until a later time
- **Retiring** – When migrating to the cloud, companies may find certain components of their infrastructure are no longer needed, in which case they are retired and not moved to the cloud