Definition of the AWS Cloud.

AWS Amazon Web Services (AWS) is a secure cloud services platform, offering compute power, database storage, content delivery and other functionality to help businesses scale and grow. Millions of customers are currently leveraging AWS cloud products and solutions to build sophisticated applications with increased flexibility, scalability and reliability.

The AWS Cloud provides a broad set of infrastructure services (over 200 services are available), such as computing power, storage options, networking and databases, delivered as a utility: on-demand, available in seconds, with pay-as-you-go pricing. It is available with **81 Availability Zones** within **25 geographic regions** across the world - with announced plans for 21 more **Availability Zones** and 7 more **AWS Regions in Australia, India, Indonesia, Israel, Spain, Switzerland, and United Arab Emirates (UAE).**

AWS' broad security certification and accreditation, data encryption at rest and in-transit, hardware security modules and strong physical security all contribute to a more secure way to manage your business' IT infrastructure. With the AWS Cloud, capabilities for controlling, auditing and managing identity, configuration and usage come built into the platform helping you meet your compliance governance and regulatory requirements.

The AWS Value Proposition

Principle	**Concepts**
Agility	**Speed of implementation of services** Experimentation Innovation
Elasticity	Scale on demand Eliminate wasted capacity
Flexibility	Broad set of products Low to no cost to entry

Security	Amazon has acquired many certifications Shared responsibility model	Please consult the table below:

But...
Ultimately, the greatest value will be found when architecting for cloud native technologies, rather than a "lift-and-shift" of the existing data center

Six advantages of Cloud Computing
1. **Trade capital expense for variable expense** – Instead of having to invest heavily in data centers and servers before you know how you’re going to use them, you can pay only when you consume computing resources, and pay only for how much you consume.
2. **Benefit from massive economies of scale** – By using cloud computing, you can achieve a lower variable cost than you can get on your own. Because usage from hundreds of thousands of customers is aggregated in the cloud, providers such as AWS can achieve higher economies of scale, which translates into lower pay as-you-go prices.
3. **Stop guessing capacity** – Eliminate guessing on your infrastructure capacity needs. When you make a capacity decision prior to deploying an application, you often end up either sitting on expensive idle resources or dealing with limited capacity. With cloud computing, these problems go away. You can access as much or as little capacity as you need, and scale up and down as required with only a few minutes’ notice.
4. **Increase speed and agility** – In a cloud computing environment, new IT resources are only a click away, which means that you reduce the time to make those resources available to your developers from weeks to just minutes. This results in a dramatic increase in agility for the organization, since the cost and time it takes to experiment and develop is significantly lower.
5. **Stop spending money running and maintaining data centers** – Focus on projects that differentiate your business, not the infrastructure. Cloud computing lets you focus on your own customers, rather than on the heavy lifting of racking, stacking, and powering servers.
6. **Go global in minutes –** Easily deploy your application in multiple regions around the world with just a few clicks. This means you can provide lower latency and a better experience for your customers at minimal cost.

Definitions
1. **Shared responsibility model –** The shared responsibility model is an agreement where Amazon is responsible for the infrastructure whereas the customer of Amazon's Web Services is responsible for everything else. Amazon's infrastructure is defined as the hardware, software, networking, and facilities that run AWS.

1.2 Identify aspects of AWS cloud economics

Most likely, your organization is not in the business of running data centers, yet a significant amount of time and money is spent doing just that. Amazon Web Services provides a way to acquire and use infrastructure on-demand, so that you pay only for what you consume. This puts more money back into the business, so that you can innovate more, expand faster, and be better positioned to take advantage of new opportunities.

You can reduce your *Total-Cost-of-Ownership (TCO)* with:
Pay-As-You-Go Pricing

AWS does not require minimum spend commitments or long-term contracts. You replace large upfront expenses with low variable payments that only apply to what you use. With AWS you are not bound to multi-year agreements or complicated licensing models.

Tiered Pricing (Use More, Pay Less)

For storage and data transfer, AWS follows a tiered pricing model. The more storage and data transfer you use, the less you pay per gigabyte. In addition, volume discounts and custom pricing are available to customers for high volume projects with unique requirements.

Cost Optimization

Cost Explorer is a free tool that provides pre-configured reports for common AWS spend queries for current and historical periods, as well as forecasting. It also allows you to customize the reports to meet your specific needs or to download your billing information for use in your own tools.

Trusted Advisor

Trusted Advisor inspects your AWS environment to find opportunities that can save you money, improve your system performance, increase your application reliability, and help you implement security best practices. Since 2013, customers have viewed over 2.6 million best-practice recommendations and realized over $350 million in estimated cost reductions.

1.3 List the different cloud architecture design principles

The Well-Architected framework has been developed to help cloud architects build the most secure, high-performing, resilient, and efficient infrastructure possible for their applications. This framework provides a consistent approach for customers and partners to evaluate architectures, and provides guidance to help implement designs that will scale with your application needs over time.

The AWS Well-Architected Framework is based on five pillars—security, reliability, performance efficiency, cost optimization, and operational excellence:

Pillar Name	Description
Security	The ability to protect information, systems, and assets while delivering business value through risk assessments and mitigation strategies. The security pillar focuses on protecting information & systems. Key topics include confidentiality and integrity of data, identifying and managing who can do what with privilege management, protecting systems, and establishing controls to detect security events.
Reliability	The ability of a system to recover from infrastructure or service failures, dynamically acquire computing resources to meet demand, and mitigate disruptions such as misconfigurations or transient network issues. The reliability pillar focuses on the ability to prevent, and quickly recover from failures to meet business and customer demand. Key topics include foundational

	elements around setup, cross project requirements, recovery planning, and how we handle change.
Performance Efficiency	The ability to use computing resources efficiently to meet system requirements, and to maintain that efficiency as demand changes and technologies evolve. The performance efficiency pillar focuses on using IT and computing resources efficiently. Key topics include selecting the right resource types and sizes based on workload requirements, monitoring performance, and making informed decisions to maintain efficiency as business needs evolve.
Cost Optimization	The ability to avoid or eliminate unneeded cost or suboptimal resources. Cost Optimization focuses on avoiding unnecessary costs. Key topics include understanding and controlling where money is being spent, selecting the most appropriate and right number of resource types, analyzing spend over time, and scaling to meet business needs without overspending.
Operational Excellence	The ability to run and monitor systems to deliver business value and to continually improve supporting processes and procedures. The operational excellence pillar focuses on running and monitoring systems to deliver business value, and continually improving processes and procedures. Key topics include managing and automating changes, responding to events, and defining standards to successfully manage daily operations.

These principles will be described in greater detail below.

Core Principles

1. Scalability
2. Disposable Resources Instead of Fixed Servers
3. Automation

4. Loose Coupling
5. Services, Not Servers
6. Databases
7. Managing Increasing Volumes of Data
8. Removing Single Points of Failure
9. Optimize for Cost
10. Caching
11. Security

The below sections provide details:

Scalability

Elasticity and Scalability are two fundamental cloud architecture principles that guide the AWS Architecture.

Elasticity is the ability to use resources in a dynamic and efficient way so the traditional anti-pattern of over-provisioning of infrastructure resources to cope with capacity requirements is avoided. Significantly, elasticity avoids the costs of these over-provisioned resources such as power, space, and maintenance. This is the AWS pay as you go/pay for what you use model.

Scalability is the ability to scale without changing the design. With AWS, scalability is achieved by scaling-out. Infrastructure and application components are designed with the premise that they will fail, instead of a just being designed around High Availability. The technology components are commodities that can be thrown out when they fail and grown by adding more when demanded. A guiding principle is to have a consistent approach to architecture and growth.

There are two types of scaling:

- Horizontal Scaling - an increase in the number of resources. Autoscaling and Bootstrapping are used for horizontal scaling. Autoscaling allows you to automatically horizontally scale to accommodate load. Bootstrapping allows you automatically setup your servers after they boot. (Using components such as Amazon Machine Images (AMI’s) and CloudFormation to automate).
- Vertical Scaling - an increase in the capabilities of the resource (e.g. faster CPU, more memory, more storage).

Disposable Resources

- Resources need to be treated as temporary disposable resources rather than fixed permanent on-premises resources before.
- AWS focuses on the concept of **Immutable infrastructure -** a server once launched, is never updated throughout its lifetime. Updates can be performed on a new server with the latest configuration. This ensures resources are always in a consistent (and tested) state and easier rollbacks.
- AWS provides multiple ways to instantiate compute resources in an automated and repeatable way:
  - **Bootstrapping -** scripts to configure and setup *for e.g. using data scripts and cloud-init to install software or copy resources and code*
  - **Golden Images -** a snapshot of a particular state of that resource. Allows faster start times and removes dependencies to configuration services or third-party repositories
  - **Containers -** AWS support for docker images through Elastic Beanstalk and ECS. Docker allows packaging a piece of software in a Docker Image, which is a standardized unit for software development, containing everything the software needs to run: code, runtime, system tools, system libraries, etc.
  - **Infrastructure as Code -** AWS assets are programmable. Techniques, practices, and tools from software development can be applied to make the whole infrastructure reusable, maintainable, extensible, and testable. AWS provides services such as CloudFormation and OpsWorks for codifying deployment

Automation

Unlike traditional IT infrastructure, Cloud enables automation of a number of events, improving both your system’s stability and the efficiency of your organization. Some of the AWS resources you can use for automation are:

- AWS Elastic Beanstalk: This resource is the fastest and simplest way to get an application up and running on AWS. You can simply upload their application code and the service automatically handles all the details, such as resource provisioning, load balancing, autoscaling, and monitoring.

- Amazon EC2 Auto recovery: You can create an Amazon CloudWatch alarm that monitors an Amazon EC2 instance and automatically recovers it if it becomes impaired. But a word of caution – During instance recovery, the instance is migrated through an instance reboot, and any data that is in-memory is lost.
- Auto Scaling: With Auto Scaling, you can maintain application availability and scale your Amazon EC2 capacity up or down automatically according to conditions you define.
- Amazon CloudWatch Alarms: You can create a CloudWatch alarm that sends an Amazon Simple Notification Service (Amazon SNS) message when a particular metric goes beyond a specified threshold for a specified number of periods.
- Amazon CloudWatch Events: The CloudWatch service delivers a near real-time stream of system events that describe changes in AWS resources. Using simple rules that you can set up in a couple of minutes, you can easily route each type of event to one or more targets: AWS Lambda functions, Amazon Kinesis streams, Amazon SNS topics, etc.
- AWS OpsWorks Lifecycle events: AWS OpsWorks supports continuous configuration through lifecycle events that automatically update your instances' configuration to adapt to environment changes. These events can be used to trigger Chef recipes on each instance to perform specific configuration tasks.
- AWS Lambda Scheduled events: These events allow you to create a Lambda function and direct AWS Lambda to execute it on a regular schedule.

As an architect for the AWS Cloud, these automation resources are a great advantage to work with.

Loose Coupling

Loosely coupled architectures reduce interdependencies, so that a change or failure in a component does not cascade to other components:

- **Asynchronous Integration**
  - does not involve direct point-to-point interaction but usually through an intermediate durable storage layer *for e.g. SQS, Kinesis.*

        - decouples the components and introduces additional resiliency.
        - suitable for any interaction that doesn’t need an immediate response and where an ack that a request has been registered will suffice.
    - **Service Discovery**
        - allows new resources to be launched or terminated at any point in time and discovered as well *for e.g. using ELB as a single point of contact with hiding the underlying instance details or Route 53 zones to abstract load balancer’s endpoint.*
    - **Well-Defined Interfaces**
        - allows various components to interact with each other through specific, technology-agnostic interfaces *for e.g. RESTful APIs with API Gateway.*

Services, Not Servers

A wide variety of underlying technology components are required to develop manage and operate applications. Your AWS cloud architecture should leverage a broad set of compute, storage, database, analytics, application, and deployment services. On AWS, there are two ways to do that. The first is through managed services that include databases, machine learning, analytics, queuing, search, email, notifications, and more. For example, with the Amazon Simple Queue Service (Amazon SQS) you can offload the administrative burden of operating and scaling a highly available messaging cluster, while paying a low price for only what you use. Not only that, Amazon SQS is inherently scalable.

The second way is to reduce the operational complexity of running applications through server-less architectures. It is possible to build both event-driven and synchronous services for mobile, web, analytics, and the Internet of Things (IoT) without managing any server infrastructure.

Databases

On AWS, managed database services help remove constraints that come with licensing costs and the ability to support diverse database engines that were a problem with the traditional IT infrastructure. You need to keep in mind that access to the information stored on these databases is the main purpose of cloud computing.

There are three different categories of databases to keep in mind while architecting:

- **Relational databases** – Data here is normalized into tables and also provided with a powerful query language (SQL), flexible indexing capabilities, strong integrity controls, and the ability to combine data from multiple tables in a fast and efficient manner. They can be scaled vertically and are highly available during failovers (designed for graceful failures). RDS allows vertical scalability by increasing resources and horizontal scalability using Read Replicas for increasing read capacity and sharding or data partitioning for improving write capacity. RDS also provides High Availability using Multi-AZ deployment, where data is synchronously replicated. Furthermore, the RDS service can be set up across a hybrid environment (i.e. distributed across a company's data center and an AWS VPC).
- **NoSQL databases (DynamoDB)** – These databases trade some of the query and transaction capabilities of relational databases for a more flexible data model that seamlessly scales horizontally. NoSQL databases utilize a variety of data models, including graphs, key-value pairs, and JSON documents. NoSQL databases are widely recognized for ease of development, scalable performance, high availability, and resilience.
- **Data warehouse (Redshift)** – A specialized type of relational database, optimized for analysis and reporting of large amounts of data. It can be used to combine transactional data from disparate sources making them available for analysis and decision-making. Redshift achieves efficient storage and optimum query performance through a combination of massively parallel processing (MPP), columnar data storage, and targeted data compression encoding schemes. Redshift's MPP architecture enables increasing performance by increasing the number of nodes in the data warehouse cluster.

*In a database world **horizontal-scaling** is often based on the partitioning of the data i.e. each node contains only part of the data, in **vertical-scaling** the data resides on a single node and scaling is done through multi-core i.e. spreading the load between the CPU and RAM resources of that machine.*

Remove single points of failure

A system is highly available when it can withstand the failure of an individual or multiple components (e.g., hard disks, servers, network links etc.). You can think about ways to automate recovery and reduce disruption at every layer of your AWS cloud architecture. This can be done with the following processes:

- Introduce redundancy to remove single points of failure, by having multiple resources for the same task. Redundancy can be implemented in either standby mode (functionality is recovered through failover while the resource remains unavailable) or active mode (requests are distributed to multiple redundant compute resources, and when one of them fails, the rest can simply absorb a larger share of the workload).
- Detection and reaction to failure should both be automated as much as possible.
- It is crucial to have a durable data storage that protects both data availability and integrity. Redundant copies of data can be introduced either through synchronous, asynchronous or Quorum based replication.
- Automated Multi –Data Center resilience is practiced through Availability Zones across data centers that reduce the impact of failures.
- Fault isolation improvement can be made to traditional horizontal scaling by sharding (a method of grouping instances into groups called shards, instead of sending the traffic from all users to every node like in the traditional IT structure).

Optimize for cost

At the end of the day, it often boils down to cost. Your AWS cloud architecture should be designed for cost optimization by keeping in mind the following principles:

- You can reduce costs by selecting the right types, configurations and storage solutions to suit your needs. Implementing Auto Scaling so that you can scale horizontally when required or scale down when necessary can be done without any extra cost.
- Taking advantage of the variety of Instance Purchasing options (Reserved and spot instances) while buying EC2 instances will help reduce cost of computing capacity.

Caching

Caching improves application performance and increases the cost efficiency of an implementation

- **Application Data Caching**
    - provides services that help store and retrieve information from fast, managed, in-memory caches
    - ElastiCache is a web service that makes it easy to deploy, operate, and scale an in-memory cache in the cloud and supports two open-source in-memory caching engines: Memcached and Redis.
- **Edge Caching**
    - allows content to be served by infrastructure that is closer to viewers, lowering latency and giving high, sustained data transfer rates needed to deliver large popular objects to end users at scale.
    - CloudFront is Content Delivery Network (CDN) consisting of multiple edge locations, that allows copies of static and dynamic content to be cached.

Security

- AWS works on shared security responsibility model
    - AWS is responsible for the security of the underlying cloud infrastructure
    - you are responsible for securing the workloads you deploy in AWS
- AWS also provides ample security features
    - IAM to define a granular set of policies and assign them to users, groups, and AWS resources
    - IAM Roles to assign short term credentials to resources, which are automatically distributed and rotated
    - Amazon Cognito, for mobile applications, which allows client devices to get controlled access to AWS resources via temporary tokens.
    - VPC to isolate parts of infrastructure through the use of subnets, security groups, and routing controls
    - WAF to help protect web applications from SQL injection and other vulnerabilities in the application code
    - CloudWatch logs to collect logs centrally as the servers are temporary
    - CloudTrail for auditing AWS API calls, which delivers a log file to S3 bucket. Logs can then be stored in an immutable manner and automatically

processed to either notify or even take action on your behalf, protecting your organization from non-compliance AWS Config, Amazon Inspector, and AWS Trusted Advisor to continually monitor for compliance or vulnerabilities giving a clear overview of which IT resources are in compliance, and which are not

- For more details refer to AWS Security Whitepaper

These principles help when architecting for the AWS cloud.

2.1 Define the AWS shared responsibility model

AWS shares the responsibility for security with the customers as shown in their graphic below. AWS is responsible for the physical security of the facilities as well as the infrastructure that includes compute, database, storage and networking resources. The customer is responsible for software, data and access that sits on top of the infrastructure layer.

2.2 Define AWS cloud security and compliance concepts

Cloud security at AWS is the highest priority. As an AWS customer, you will benefit from a data center and network architecture built to meet the requirements of the most security-sensitive organizations.

An advantage of the AWS cloud is that it allows customers to scale and innovate, while maintaining a secure environment. Customers pay only for the services they use, meaning that you can have the security you need, but without the upfront expenses, and at a lower cost than in an on-premises environment.

Infrastructure Security

- Network firewalls built into Amazon VPC.
- In transit encryption using TLS across all services.
- Private or dedicated connections into your data center

Infrastructure Resilience

- Technologies built from the ground up for resilience in the face of DDoS attacks.
- Services can be used in combination to automatically scale for traffic load.
- Autoscaling, CloudFront, Route 53 can be used to prevent DDoS.

Data Encryption

- At rest encryption available in EBS, S3, Glacier, RDS (Oracle and SQL Server) and Redshift.
- Key management through AWS KMS - you can choose whether to control the keys or let AWS.
- Server side encryption of message queues in SQS.
- Dedicated hardware-based cryptographic key storage using AWS CloudHSM, allowing you to satisfy compliance requirements.
- APIs to integrate AWS security into any applications you create.

Standards and Best Practices

- A security assessment service, Amazon Inspector, that automatically assesses applications for vulnerabilities or deviations from best practices, including impacted networks, OS, and attached storage
- Deployment tools to manage the creation and decommissioning of AWS resources according to organizational standards
- Inventory and configuration management tools, like AWS Config, that identify AWS resources then track, and manage changes to those resources over time
- Template definition and management tools, including AWS CloudFormation to create standard, preconfigured environments

Monitoring and Logging

- Deep visibility into API calls through AWS CloudTrail, including who, what, when, and from where calls were made
- Log aggregation options, streamlining investigations and compliance reporting
- Alert notifications through Amazon CloudWatch when specific events occur or thresholds are exceeded

Identity and Access Control

- AWS Identity and Access Management (IAM) lets you define individual user accounts with permissions across AWS resources
- AWS Multi-Factor Authentication for privileged accounts, including options for hardware-based authenticators
- AWS Directory Service allows you to integrate and federate with corporate directories to reduce administrative overhead and improve end-user experience

Security Support

- Real-time insight through AWS Trusted Advisor

- Proactive support and advocacy with a Technical Account Manager (TAM)

2.3 Identify AWS access management capabilities

AWS access management is provided through AWS IAM - Identity and Access Management. AWS Identity and Access Management (IAM) enables you to securely control access to AWS services and resources for your users. Using IAM, you can create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources.

IAM is a feature of your AWS account offered at no additional charge. You will be charged only for use of other AWS services by your users.

IAM allows you to:

- Manage IAM users and their access – You can create users in IAM, assign them individual security credentials (in other words, access keys, passwords, and multi-factor authentication devices), or request temporary security credentials to provide users access to AWS services and resources. You can manage permissions in order to control which operations a user can perform.
- Manage IAM roles and their permissions – You can create roles in IAM and manage permissions to control which operations can be performed by the entity, or AWS service, that assumes the role. You can also define which entity is allowed to assume the role. In addition, you can use service-linked roles to delegate permissions to AWS services that create and manage AWS resources on your behalf.
- Manage federated users and their permissions – You can enable identity federation to allow existing identities (users, groups, and roles) in your enterprise to access the AWS Management Console, call AWS APIs, and access resources, without the need to create an IAM user for each identity.

Access and Federation

You can grant other people permission to administer and use resources in your AWS account without having to share your password or access key.

Also you can allow users who already have passwords elsewhere—for example, in your corporate Active Directory or with an Internet identity provider—to get access to your AWS account. You can use any identity management solution that supports SAML 2.0.

Granular Permissions

You can grant different permissions to different people for different resources. For example, you might allow some users complete access to Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple Storage Service (Amazon S3), Amazon DynamoDB, Amazon Redshift, and other AWS services. For other users, you can allow read-only access to just some S3 buckets, or permission to administer just some EC2 instances, or to access your billing information but nothing else.

IAM also enables you to add specific conditions such as time of day to control how a user can use AWS, their originating IP address, whether they are using SSL, or whether they have authenticated with a multi-factor authentication device.

Securing Application Access

You can use IAM features to securely give applications that run on EC2 instances the credentials that they need in order to access other AWS resources, like S3 buckets and RDS or DynamoDB databases.

Multi Factor Authentication

You can add two-factor authentication to your account and to individual users for extra security. With MFA you or your users must provide not only a password or access key to work with your account, but also a code from a specially configured device.

2.4 Identify resources for security support

AWS Support

AWS Support provides tools and resources to help you make sure your AWS environment is built and operated to be secure, highly available, efficient, and cost effective. With a focus on the security and operational health of your infrastructure, AWS Support provides:

- Real-time insight through AWS Trusted Advisor. Trusted Advisor is an online tool that serves as your customized cloud expert, and helps you provision your resources by following best practices. Trusted Advisor inspects your AWS environment and finds opportunities to save money, improve system performance and reliability, or help close security gaps.

- Proactive support and advocacy through Technical Account Manager (TAM). Your TAM is your single point of contact and advocate who provides technical expertise across the full range of AWS services. TAMs work closely with you to deliver proactive guidance and early awareness of new features and recommendations. And should any unplanned issues arise, your TAM ensures that they are resolved as swiftly as possible.

The full range of AWS Trusted Advisor checks is available to all customers with Business and Enterprise Level Support, and Technical Account Managers are included with Enterprise Support.

Professional Services

AWS Professional Services and the AWS Partner Network can help you create more comprehensive solutions to security problems, everything from strategic advice, deployments, to re-engineering your security processes. Example engagements include:

- Enterprise Security Architecture. Evaluate the nature of the workloads you are deploying in AWS, along with your security needs and define an architecture and set of security controls that will protect your data and workloads according to best practices.
- Policies and Controls-Mapping. Examine your requirements based upon your security policy and any third party or regulatory mandates and provide detailed recommendations on how to satisfy those requirements and demonstrate compliance.
- Security Operations Playbook. Define the right organizational structures and processes to ensure that security controls are working correctly, as well as detect and respond to security issues that arise within your AWS environment
- Business Unit Workshops. Work with IT and Business Leaders across your organization to understand their plans and strategies around Cloud adoption, educate them on the best way to satisfy their requirements while minimizing risks to the organization, and devise an organization wide security framework for deploying workloads on AWS.

3.1 Define methods of deploying and operating in the AWS cloud

Amazon Web Services offers multiple options for provisioning your IT infrastructure and the deployment of your applications. Whether it is a simple three-tier application or a complex set of workloads, the deployment model varies from customer to customer. But with the right techniques, AWS can help you pick the best strategy and tool set for deploying an infrastructure that can handle your workload. **The main principles to remember are AAA - Automate, Automate, Automate.**

AWS Elastic Beanstalk

Elastic Beanstalk is a high-level deployment tool that helps you get an app from your desktop to the web in a matter of minutes. Elastic Beanstalk handles the details of your hosting environment—capacity provisioning, load balancing, scaling, and application health monitoring—so you don't have to.

A platform configuration defines the infrastructure and software stack to be used for a given environment. When you deploy your app, Elastic Beanstalk provisions a set of AWS resources that can include Amazon EC2 instances, alarms, a load balancer, security groups, and more.

AWS CloudFormation

AWS CloudFormation is a service that helps you model and set up your Amazon Web Services resources so that you can spend less time managing those resources and more time focusing on your applications that run in AWS. You create a template that describes all the AWS resources that you want (like Amazon EC2 instances or Amazon RDS DB instances), and AWS CloudFormation takes care of provisioning and configuring those resources for you.

AWS OpsWorks

AWS OpsWorks is a configuration management service that helps you configure and operate applications in a cloud enterprise by using Chef. There are 2 variants: AWS OpsWorks Stacks and AWS OpsWorks for Chef Automate.

AWS OpsWorks Stacks

AWS OpsWorks Stacks, the original service, provides a simple and flexible way to create and manage stacks and applications. AWS OpsWorks Stacks lets you deploy and monitor applications in your stacks. Unlike AWS OpsWorks for Chef Automate, AWS OpsWorks Stacks does not require or create Chef servers; AWS OpsWorks Stacks performs some of the work of a Chef server for you. AWS OpsWorks Stacks monitors instance health, and provisions new instances for you, when necessary, by using Auto Healing and Auto Scaling.

AWS OpsWorks for Chef Automate

AWS OpsWorks for Chef Automate lets you create AWS-managed Chef servers that include Chef Automate premium features, and use the Chef DK and other Chef tooling to manage them. WS OpsWorks for Chef Automate manages both Chef Automate Server and Chef Server software on a single instance.

AWS CodeCommit

AWS CodeCommit is a fully-managed source control service that makes it easy for companies to host secure and highly scalable private Git repositories. CodeCommit integrates with AWS CodePipeline and AWS CodeDeploy to streamline your development and release process.

AWS CodePipeline

AWS CodePipeline is a continuous integration and continuous delivery service for fast and reliable application and infrastructure updates. CodePipeline builds, tests, and deploys your code every time there is a code change, based on the release process models you define.

AWS CodeDeploy

AWS CodeDeploy is a service that automates code deployments and software deployments to any instance, including Amazon EC2 instances and instances running on-premises. AWS CodeDeploy makes it easier for you to rapidly release new features, helps you avoid downtime during application deployment, and handles the complexity of updating your applications.

Amazon Elastic Container Service

Amazon Elastic Container Service (ECS) is a highly scalable, high performance container management service that
supports Docker containers and allows you to easily run applications on a managed cluster of Amazon EC2 instances. Amazon ECS eliminates the need for you to install, operate, and scale your own cluster management infrastructure.

Non-AWS Solutions

Infrastructure as Code

- Terraform
- Salt Stack
- Ansible

Configuration Management

- Chef

- Puppet
- Ansible

Continuous Integration

- Jenkins
- TeamCity

Hosted Version Control Repositories

- GitHub
- GitLab
- BitBucket

General Principles:

Good Practice

- Provision infrastructure from code
- Deploy artifacts automatically from version control
- Configuration managed from code and applied automatically
- Scale your infrastructure automatically
- Monitor every aspect of the pipeline and the infrastructure (CloudWatch)
- Logging for every action (CloudWatch Logs and CloudTrail)
- Instance profiles for embedding IAM roles in instances automatically
- Use variables, don't hard code values
- Tagging can be used with automation to provide more insights on what has been provisioned

Updating Your Stack

There are many ways to update your stack.

- You can update your AMIs and then deploy a new environment from them.
- You can use CI tools to deploy the code to existing environments.
- You can use the "Blue/Green" method to have one environment for production (blue) and one for the new version (green). When it is time to upgrade, simply redirect the traffic from blue to green.

3.2 Define the AWS global infrastructure

The AWS Cloud operates 80 Availability Zones within 25 geographic Regions around the world, with announced plans for 9 more Availability Zones and 3 more Regions in Cape Town, Jakarta, and Milan.

AWS Regions and Availability Zones

The AWS Cloud infrastructure is built around Regions and Availability Zones (AZs). A Region is a physical location in the world with multiple AZs.

Availability Zones consist of one or more discrete data centers, each with redundant power and networking, housed in separate facilities that are located on stable flood plains. These AZs offer the abilities to operate production applications and databases which are highly available, fault tolerant, and scalable than would be possible from a single data center. In total, the AWS Cloud operates 80 Availability Zones within 25 geographic Regions around the world.

Region & Number of Availability Zones

US East

N. Virginia (6), Ohio (3)

US West

N. California (3), Oregon (3)

Asia Pacific

Mumbai (2), Seoul (2), Singapore (2), Sydney (3), Tokyo (4), Bahrain

Canada

Central (2)

China

Beijing (2)

Europe

Frankfurt (3), Ireland (3), London (2)

South America

São Paulo (3)

AWS GovCloud (US-West) (2)

The components are:

- Availability Zones (AZs)
- Regions
- Edge Locations
- Regional Edge Caches

High Availability Through Multiple Availability Zones

Unlike virtually every other technology infrastructure provider, each AWS Region has multiple Availability Zones and data centers. As we've learned from running the leading cloud infrastructure technology platform since 2006, customers who care about the availability and performance of their applications want to deploy these applications across multiple Availability Zones in the same region for fault tolerance and low latency. Availability Zones are connected to each other with fast and private fiber-optic network, which enables applications to automatically fail-over between Availability Zones without interruption.

Further Improving Availability by Deploying in Multiple Regions

In addition to replicating applications and data across multiple data centers in the same Region using Availability Zones, you can also choose to further increase redundancy and fault tolerance by replicating data between geographic Regions. You can do so using both private and public network to provide an additional layer of business continuity, or to provide low latency access across the globe.

Meeting Compliance and Data Residency Requirements

You retain complete control and ownership over the region in which your data is physically located, making it easy to meet regional compliance and data residency requirements.

Geographic Expansion

The AWS Cloud has announced plans to expand with 9 more Availability Zones and 3 more Regions in Cape Town, Jakarta, and Milan.

Edge Locations

Edge Locations are AWS sites deployed in major cities and highly populated areas across the globe. They far outnumber the number of availability zones available.

While Edge Locations are not used to deploy your main infrastructures such as EC2 instances, EBS storage, VPCs, or RDS resources like AZs, they are used by AWS services such as AWS CloudFront and AWS Lambda@Edge (currently in Preview) to cache data and reduce latency for end user access by using the Edge Locations as a global Content Delivery Network (CDN).

As a result, Edge Locations are primarily used by end users who are accessing and using your services.

For example, you may have your website hosted on EC2 instances and S3 (your origin) within the Ohio region with a configured CloudFront distribution associated. When a user accesses your website from Europe, they would be re-directed to their closest Edge Location (in Europe) where cached data could be read on your website, significantly reducing latency.

3.3 Identify the core AWS services

Here's a list of the services you should definitely know. But, don't be surprised if you see questions about others as well:

- EC2
- VPC
- S3
- RDS
- Lambda
- Route 53
- SNS
- SQS
- ELB

The level of detail in each question depends on the service. More widely used services may require a bit more knowledge, and others will only require that you know what the service does. For example, EC2 is one of the most important AWS services, so you could be asked questions about different instance types for different scenarios. On the other hand, you may only be asked to choose the best description of a service like CloudFront.

In addition to the traditional services, the exam covers other AWS technology, including the command line interface (CLI) and software development kit (SDK). You may also see questions that overlap with other exam domains. For example, services like AWS Trusted Advisor and AWS Cost Calculator may fall into the technology domain as well as billing and pricing.

EC2

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers.

Instance Types

General Purpose

T2 - T2 instances are Burstable Performance Instances that provide a baseline level of CPU performance with the ability to burst above the baseline.

M4 - M4 instances are the latest generation of General Purpose Instances. This family provides a balance of compute, memory, and network resources, and it is a good choice for many applications.

Compute Optimised

C4 - C4 instances are the latest generation of Compute-optimized instances, featuring the highest performing processors and the lowest price/compute performance in EC2.

Memory Optimised

X1 - X1 Instances are optimized for large-scale, enterprise-class, in-memory applications and high-performance databases, and have the lowest price per GiB of RAM among Amazon EC2 instance types.

R4 - R4 instances are optimized for memory-intensive applications and offer better price per GiB of RAM than R3. The RAM sizes are a step below the X1s.

Accelerated Computing

P2 - P2 instances are intended for general-purpose GPU compute applications.

G3 - G3 instances are optimized for graphics-intensive applications. The GPU specs are a step below the P2s.

F1 - F1 instances offer customizable hardware acceleration with field programmable gate arrays (FPGAs).

Storage Optimised

I3 - High I/O instances. This family includes the High Storage Instances that provide Non-Volatile Memory Express (NVMe) SSD backed instance storage optimized for low latency, very high random I/O performance, high sequential read throughput and provide high IOPS at a low cost.

D2 - Dense-storage instances. D2 instances feature up to 48 TB of HDD-based local storage, deliver high disk throughput, and offer the lowest price per disk throughput performance on Amazon EC2.

Pricing

Amazon EC2 is free to try. There are four ways to pay for Amazon EC2 instances: On-Demand, Reserved Instances, and Spot Instances & Per-Second Billing. You can also pay for Dedicated Hosts which provide you with EC2 instance capacity on physical servers dedicated for your use.

On-Demand

With On-Demand instances, you pay for computing capacity by per hour or per second depending on which instances you run. No longer-term commitments or upfront payments are needed.

Spot Instances

Amazon EC2 Spot instances allow you to bid on spare Amazon EC2 computing capacity for up to 90% off the On-Demand price. Spot instances are recommended for applications that have flexible start and end times, applications that are only feasible at very low compute prices or users with urgent computing needs for large amounts of additional capacity.

Reserved Instances

Reserved Instances provide you with a significant discount (up to 75%) compared to On-Demand instance pricing. For applications that have steady state or predictable usage, require reserved capacity or can commit to using EC2 for a 1 or 3 year period, Reserved Instances can provide significant savings compared to using On-Demand instances.

Per-Second Billing

With per-second billing, you pay for only what you use. It takes cost of unused minutes and seconds in an hour off of the bill, so you can focus on improving your applications instead of maximising usage to the hour.

Security Groups

A *security group* acts as a virtual firewall that controls the traffic for one or more instances. When you launch an instance, you associate one or more security groups with the instance. You add rules to each security group that allow traffic to or from its associated instances. You can modify the rules for a security group at any time; the new rules are automatically applied to all instances that are associated with the security group. When we decide whether to allow traffic to reach an instance, we evaluate all the rules from all the security groups that are associated with the instance.

S3

Amazon S3 is object storage built to store and retrieve any amount of data from anywhere – web sites and mobile apps, corporate applications, and data from IoT sensors or devices. It is designed to deliver 99.999999999% durability, provides comprehensive security and compliance capabilities that meet even the most stringent regulatory requirements and gives customers flexibility in the way they manage data for cost optimization, access control, and compliance. Also, S3 is the only cloud storage solution with query-in-place functionality, allowing you to run powerful analytics directly on your data at rest in S3.

Storage Classes

Amazon S3 offers a range of storage classes designed for different use cases. Lifecycle transitions can be used to move data between classes, given certain events.

Amazon S3 Standard

Designed for general-purpose storage of frequently accessed data. Delivers low latency and high throughput, perfect for a wide variety of use cases. There is no retrieval fee, minimum object size or minimum storage duration.

Amazon S3 Standard - Infrequent Access

Designed for long-lived, but less frequently accessed data. For data that is accessed less frequently, but requires rapid access when needed. Standard - IA offers the high durability, throughput, and low latency of Amazon S3 Standard, with a low per GB storage price and per GB retrieval fee.

Amazon Glacier

Designed for long-term archive. Secure, durable, and extremely low-cost storage service for data archiving. You can reliably store any amount of data at costs that are competitive with or cheaper than on-premises solutions. Amazon Glacier provides three options for access to archives, from a few minutes to several hours.

RDS

Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficiency and resizable capacity while automating time-consuming administration tasks such as hardware provisioning, database setup, patching and backups.

Amazon RDS is available on several database instance types - optimized for memory, performance or I/O. RDS provides you with six familiar database engines to choose from. Amazon RDS supports encryption at rest and in transit, using keys managed through KMS. Backups are automated, user-initiated snapshots are available and database software is updated automatically.

Instance Types

- General Purpose
- Memory Optimized

Database Engines

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

Supporting Services

AWS Database Migration Service

AWS Database Migration Service can help you migrate databases to AWS easily and securely. The source database remains fully operational during the migration, minimizing downtime to applications that rely on the database. The AWS Database Migration Service can migrate your data to and from most widely used commercial and open-source databases. The service supports homogenous migrations such as Oracle to Oracle, as well as heterogeneous migrations between different database platforms, such as Oracle to Amazon Aurora or Microsoft SQL Server to MySQL.

It also allows you to stream data to Amazon Redshift, Amazon DynamoDB, and Amazon S3 from any of the supported sources including Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle, SAP ASE, SQL Server and MongoDB, enabling consolidation and easy analysis of data in the petabyte-scale data warehouse. AWS Database Migration Service can also be used for continuous data replication with high-availability.

AWS Schema Conversion Tool

The AWS Schema Conversion Tool makes heterogeneous database migrations predictable by automatically converting the source database schema and a majority of the database code objects, including views, stored procedures, and functions, to a format compatible with the target database. Any objects that cannot be automatically converted are clearly marked so that they can be manually converted to complete the migration. SCT can also scan your application source code for embedded SQL statements and convert them as part of a database schema conversion project.

Your source database can be on-premises, or in Amazon RDS or EC2 and the target database can be in either Amazon RDS or EC2. The AWS Schema Conversion Tool supports conversions from multiple RBMS providers to an equivalent database in RDS, or from multiple data warehouse providers to Amazon Redshift.

Lambda

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume - there is no charge when your code is not running. With Lambda, you can run code for virtually any type of application or backend service - all with zero administration.

AWS Lambda automatically scales your application by running code in response to each trigger. Your code runs in parallel and processes each trigger individually, scaling precisely with the size of the workload. With AWS Lambda, you are charged for every 100ms your code executes and the number of times your code is triggered. You don't pay anything when your code isn't running.

Route 53

Amazon Route 53 is a highly available and scalable cloud Domain Name System (DNS) web service. You can use Amazon Route 53 to configure DNS health checks to route traffic to healthy endpoints or to independently monitor the health of your application and its endpoints.

Amazon Route 53 Traffic Flow makes it easy for you to manage traffic globally through a variety of routing types, including Latency Based Routing, Geo DNS, Geoproximity, and Weighted Round Robin—all of which can be combined with DNS Failover in order to enable a variety of low-latency, fault-tolerant architectures. Using Amazon Route 53 Traffic Flow’s simple visual editor, you can easily manage how your end-users are routed to your application’s endpoints—whether in a single AWS region or distributed around the globe.

Amazon Route 53 also offers Domain Name Registration – you can purchase and manage domain names such as example.com and Amazon Route 53 will automatically configure DNS settings for your domains.

Amazon Route 53 is integrated with Elastic Load Balancing (ELB).

SNS

Amazon Simple Notification Service (SNS) is a Pub/Sub messaging and mobile notifications for microservices, distributed systems, and serverless applications. Amazon SNS Mobile Notifications makes it simple and cost effective to send push notifications to iOS, Android, Fire OS, Windows and Baidu-based devices. It supports HTTP/HTTPS, Email/Email-JSON, SMS or Amazon Simple Queue Service (SQS) queues, or AWS Lambda functions.

Amazon DevPay

Amazon DevPay is a simple-to-use online billing and account management service that makes it easy for businesses to sell applications that are built in, or run on top of, Amazon Web Services.

Amazon QuickSight

Amazon QuickSight is a fast business analytics service you can use to build visualizations, perform ad hoc analysis, and quickly get business insights from your data. You can access data from multiple sources – upload files or connect to AWS data sources or external databases.

SQS

1. Fully managed message queuing service
2. Lets you decouple and scale microservices, distributed systems, and serverless applications
3. Eliminates the complexity and overhead associated with managing and operating message oriented middleware
4. Send, store, and receive messages between software components at any volume, without losing messages or requiring other services to be available.
5. Two types of message queues
    - **Standard** queues offer maximum throughput, best-effort ordering, and at-least-once delivery.
    - **SQS FIFO** queues guarantee that messages are processed exactly once, in the exact order that they are sent.

3.4 Identify resources for technology support

The resources available to you for technical support depend on your support plan. They are:

- Developer Support
- Business Support
- Enterprise Support

AWS Support Ticket

If you are on one of these AWS Support plans you can raise a ticket. It's much faster to get support via the console and create a request.

Technical Account Manager

Your designated Technical Account Manager (TAM) is your primary point of contact who provides guidance, architectural review, and ongoing communication to keep you informed and well prepared as you plan, deploy, and proactively optimize your solutions.

- A dedicated voice within AWS to serve as your technical point of contact and advocate
- Proactive guidance and best practices to help optimize your AWS environment
- Orchestration and access to the breadth and depth of technical expertise across the full range of AWS

Trusted Advisor

AWS Trusted Advisor is an online resource that helps you provision your resources following best practices to help reduce cost, increase performance and fault tolerance, and improve security by optimizing your AWS environment. While the four core checks are available to all AWS customers, the full power of AWS Trusted Advisor is available with Business and Enterprise Support plans.

- Guidance on getting the optimal performance and availability based on your requirements
- Opportunities to reduce your monthly spend and retain or increase productivity
- Best practices to help increase security

AWS Whitepapers

AWS Whitepapers features a comprehensive list of technical AWS whitepapers, covering topics such as architecture, security, and economics. These whitepapers have been authored by the AWS Team, independent analysts, or the AWS Community (Customers or Partners).

AWS Service Health Dashboard
This is a general view of the health of all AWS services.

https://status.aws.amazon.com/
AWS Personal Health Dashboard
This is a personal view of the health of the AWS services that are used by you.

4.1 Compare and contrast the various pricing models for AWS

The pricing models for AWS are principally "Compute Resources" and "Data Storage and Transfer". However, there are also other pricing models such as request pricing, monthly targeted audience (MTA) pricing, events collected pricing, messages sent pricing, and more.

Below, we'll explain the differences between the two major pricing models, so before you decide to use a service, be sure to check which one it uses.

Compute Resources
When you use compute resources, you pay on an hourly basis from the time you launch a resource until the time you terminate it. You can also get volume discounts up to 10% when you reserve more. The following Amazon Web Services fall under the compute resources pricing model.

- AWS Lambda
- Amazon Lightsail
- Elastic Load Balancing
- Amazon EC2 Container Registry
- Amazon Virtual Private Cloud (VPC)
- Amazon Elastic Compute Cloud (EC2)

Data Storage and Transfer
When you use data storage and transfer resources, you pay on a per-gigabyte basis. This pricing model is also known as *tiered*. This means that several different aspects of the service differ in cost. For instance, Amazon S3 has different prices for storage, requests, and data transfer. However, the more you use, the less you pay per gigabyte! The following services fall under the data storage and transfer pricing model.

- AWS Snowball
- Amazon Glacier
- AWS Snowmobile
- AWS Snowball Edge

- AWS Storage Gateway
- Amazon Elastic File System (EFS)
- Amazon Elastic Block Storage (EBS)
- Amazon Simple Storage Service (S3)

4.2 Recognize the various account structures in relation to AWS billing and pricing

Accounts act as the main billing entity for AWS Resources. Different billing options are available including invoicing, such as "consolidated billing" - letting one account pick up the bill for multiple ‘sub accounts’. With regards to billing we can set up billing alerts, AWS Budgets and automated bill reporting for better insights. We can also utilise tagging for better cost allocation.

AWS Account Structures

Business Unit (BU) Account Structure

This account structure can be beneficial for customers who want to align their AWS operational and billing controls with individual BUs. It offers individual units operational autonomy while providing a company with a consolidated bill and combined view of all AWS charges, separated by group, OU, or cost center.

Environment Lifecycle Account Structure

This account structure can be beneficial for customers who want to align their AWS operational and billing controls with their application development lifecycle. It offers development-lifecycle operational autonomy while providing a company with a consolidated bill and combined view of all AWS charges, separated by development environment.

Project-Based Account Structure

This account structure can be beneficial for customers who want to align their AWS operational and billing controls by product, application workload, or program. It offers project or workload operational autonomy while providing the company with a consolidated bill and combined view of all AWS charges, separated by project. This structure also simplifies the ability to trigger cost alerts based on project, application workload, or program consumption of AWS resources.

Hybrid AWS Account Structures

The basic account structures described previously work for most small companies, however some larger AWS customers find it advantageous to create hybrid combinations that group accounts by multiple dimensions.

Other Billing and Pricing Considerations

Consolidated Billing for Organizations

You can use the consolidated billing feature in AWS Organizations to consolidate payment for multiple AWS accounts. With consolidated billing, you can see a combined view of AWS charges incurred by all of your accounts. You also can get a cost report for each member account that is associated with your master account. Consolidated billing is offered at no additional charge.

Resource Tagging

Resource tagging can help track expenses throughout the model. Some of the tags that can be useful for billing purposes include:

- Owner – Used to identify who is responsible for the resource
- Cost Center/Business Unit – Used to identify the cost center or business unit associated with a resource; typically for cost allocation and tracking
- Customer – Used to identify a specific client that a particular group of resources serves
- Project – Used to identify the project(s) the resource supports

Customers can activate an AWS-generated *createdBy* tag that is automatically applied for cost allocation purposes, to help account for resources that might otherwise go uncategorized. The *createdBy* tag is available for supported AWS services and resources only, and its value contains data associated with specific API or console events.

4.3 Identify resources available for billing support

The quickest way to find answers to questions about your bill might be to start with the AWS Knowledge Center.

In addition, all AWS account owners have access to account and billing support free of charge. Only personalized technical support requires a support plan. For more information, visit the AWS Support web site.

It is also worthwhile to know that the AWS Cost Calculator can be used to provide an estimate for your monthly bill. Also Trusted Advisor can help with reducing costs, by suggesting changes to your existing infrastructure.

This section guides you through contacting AWS Support and opening a support case for your billing inquiry, which is the fastest and most direct method for communicating with AWS Support. AWS Support does not publish a direct phone number for reaching a support representative.

Contacting AWS Support

1. Sign in and navigate to the AWS Support Center. If prompted, type the email address and password for your account.
2. Choose **Open a new case**.
3. On the **Open a new case** page, select **Account and Billing Support** and fill in the required fields on the form.

After you complete the form, you can choose **Web** for an email response, or **Phone** to request a telephone call from an AWS Support representative. Instant messaging support is not available for billing inquiries.

Support Concierge

Included as part of the Enterprise Support plan, the Support Concierge is a billing and account expert who provides quick and efficient analysis and assistance. Your assigned Concierge addresses all non-technical billing and account level inquiries - freeing up your time to run your business.

- A primary contact to help manage AWS billing and account-level services
- Proactive guidance and best practices for billing allocation, consolidation of accounts, and root-level account security
- Direct access to a billing advocate for payment inquiries, cost reports, service limits, and bulk purchases