KnoDAX

AWS Certified Cloud Practitioner Exam Guide

SK Singh
AWS, Kafka, Hadoop, Unix, Oracle, Java Certified
Founder (of KnoDAX), Software, Cloud, Data Engineer

ISBN:

Table of Contents

Introduction

The purpose of the book is the help you pass the AWS Cloud Certified Practitioner exam. The book covers topics for all domains. Knowledge gained from this book will help you in other AWS certification exams as well.

In the second part (Practice Tests), there are five practice test sets with answers – each of them contains 65 exam-like questions. These questions will help you apply your learning so that you can better prepare for the exam and clear the exam on the first attempt.

Chapter 1. Becoming an AWS Certified Cloud Practitioner

You will learn the following in this chapter:

- Why AWS Certified Cloud Practitioner (CCP) is a valuable cloud certification
- How to get AWS CCP certification
- High level overview of all domains covered in the exam

Whether we use Facebook, Twitter, Gmail, Zoom (for meetings), or Citrix server (to connect to the workplace remotely), cloud computing is everywhere around us. If we pay little attention, there are countless examples of cloud computing in our day-to-day life. The cloud computing has created titanic shift happening in how organizations used to have their IT infrastructure in 90s and compare to how the new trend which is: adoption to cloud. The change in how cloud is being adopted in IT infrastructure, causing another type of change which is need for cloud applications of different types. In other words, now we need engineers who can build application that can be much easily deployed and managed on cloud. Developing cloud applications require different type of architect, design and programming skills than the skills needed for traditional classic client-server or monolithic applications. Not only architecting, designing, and developing, new trend is occurring – it's already here -- how system administrator and technical operations engineers used to work – which is mainly dealt with OS and network level operations and scripting. Now the new DevOps engineers not only have the role of developer but also of system admin operations which is build and deploy the application. What it means software industry needs now more DevOps engineers as well.

Based on the above discussion we can see that how software (also called by many "IT") industry needs cloud engineers of different types who can architect, design, build, deploy and manage cloud applications and cloud infrastructure.

Why Get Certified in AWS?

Though there are many cloud providers but AWS has been leader in the cloud computing industry based on the Gartner Magic Quadrant for Cloud Infrastructure & Platform Services AWS has over million customers in more that 200 countries.

As AWS is used by many types of organizations all over the world, this is creating more employment opportunities for people having AWS skills. The question is when hiring how employer can have some kind of proof that the engineer who they are interviewing has the AWS skills or not. In other

words, having AWS certifications increases likelihood of not only getting interviews but also getting hired. In addition, because AWS exams are hard, having AWS certification, generates lots of self-confidence in making difference in your team and organizations with your AWS skills.

How to Get AWS CCP Certification?

Though there is no pre-condition for the certification exam, the target candidate should have 6 months of active engagement with AWS cloud platform with exposure of AWS Cloud design, implementation and/or operations. That's the reason, it is recommended or rather extremely important to have hands-on experience with AWS platform before taking the exam.

The candidate should have knowledge of AWS Cloud concepts, security and compliance with AWS Cloud, understanding of AWS core services, understanding of AWS economics of the AWS cloud.

The exam specifically mentions what are items are considered out of the scope for the exam. These are coding, designing cloud architecture, troubleshooting, implementation, migration, load and performance testing, and business application such as Amazon Arora, Amazon Chime, Amazon WorkMail.

The exam contains two types of questions: multiple choice in which there will be one correct answer, and multiple response in which there will be 2 or more correct answers. The unanswered questions are marked as incorrect and there is no negative marking for wrong answers. The exam contains 50 questions, which affects your score. In addition, there will be additional 15 questions that will not affect scores. AWS uses response of unanswered questions to make its quiz more better in future. Exam report is scaled and the minimum passing score is 700.

AWS CCP Exam Domains

The AWS CCP exam asks questions from the four domains and percentage of questions asked in each domain varies. The table below represents name of each domain and percentage of questions that are asked from each domain.

Domain	% of Exam
Domain 1: Cloud Concepts	26%
Domain 2: Security and Compliance	25%
Domain 3: Technology	33%
Domain 4: Billing and Pricing	16%
Total	100%

Let's see what the different topics are covered in each domain with respect to exam.

Domain 1: Cloud Concepts

This domain includes introduction to AWS and what value proposition AWS cloud provide to its customers which includes not only business value but also quality attributes such as security, reliability, high availability etc. to the deployed applications. Besides what value AWS provides, this domain also includes aspects of AWS cloud economics such as role of operational expenses (OpEx), role of capital expenses (CapEx), costs associated with on-premises data center operations, impact of cost of software licensing when moving to the cloud, and which operations costs will be reduced because of moving to cloud. Finally, this domain covers design and architecture aspect of cloud. It

includes different design principles related to cloud architecture such as design for failure, decoupled vs monolithic architecture, implementation of elasticity, parallelism.

Domain 2: Security and Compliance

This domain includes AWS shared responsibility model which is about what the customer's responsibilities are and what the responsibilities are of AWS. For example, AWS provides virtual machine but who will be responsible for applying patches or who will be responsible for maintenance of launched virtual machine for example backup of data or maintenance of software installed. This domain also covers how AWS manages security of applications deployed on AWS cloud and what types of compliance certification AWS cloud has received. In addition, this domain also covers how users, and their identities are managed on the AWS platform. For example, how to manage users, groups and their permissions using Identity and Access Management (IAM), how to secure AWS account using concept called multi-factor authentication (MFA), what is AWS root account and protection of AWS root account. Additionally, this domain includes different documentation such as best practices, whitepapers, official documents, and such and how to find them. Finally, it covers how to secure resources using different network security capabilities (Network ACLs, AWS WAF) and 3rd party security products form the AWS Marketplace, and AWS Trusted Advisor.

Domain 3: Technology

This domain covers technology aspect of AWS platform such as different methods of deploying and operating in the AWS Cloud, understanding AWS global infrastructure, core AWS services, and resources for technology related support. With regards to deploying and operating in the AWS Cloud, it includes different deployment models, connectivity options; and different ways of provisioning and operating such as programmatic access, AWS API, AWS SDK, AWS Management Console, and AWS CLI. With regards to AWS cloud infrastructure, it includes understanding and how to use AWS regions, availability zones and other related concepts depending on the use cases. Finally, with regards to AWS core services, it includes compute (EC2, Lambda, ECS, Autoscaling), storage (S3, EBS, Glacier, Snowball, EFS, Storage Gateway), network (VPC, Route 53, VPN, Direct Connect), and database (RDS, RedShift, DynamoDB) related services.

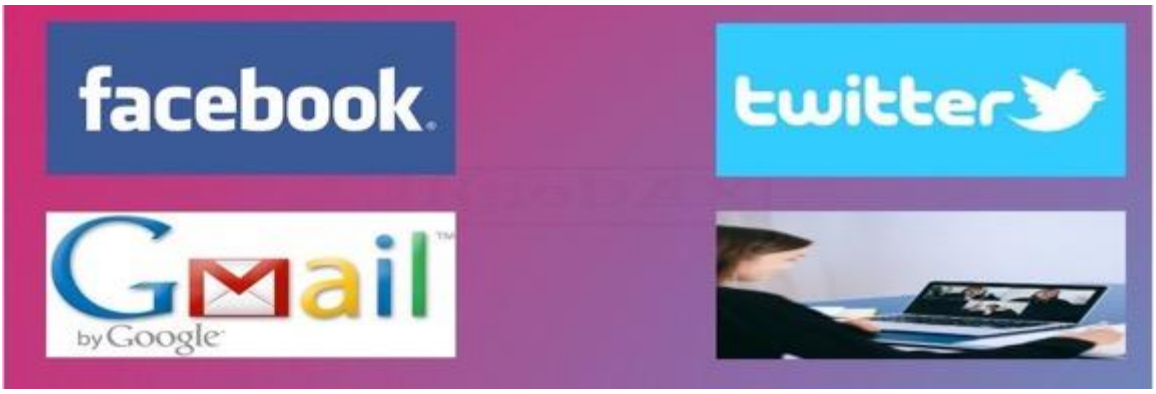
Domain 4: Billing and Pricing

This domain includes compare and contrast of various pricing models for AWS, various account structure (consolidated and multiple accounts), and billing support related resources. With respect to pricing models, it includes On-Demand Instance pricing, Reserved-Instance pricing, and Spot Instance pricing. With respect to billing support, it includes different ways to get billing and support information such as Cost Explorer, AWS Cost and Usage Report, opening billing support case, and the role of the Concierge for AWS Enterprise Support Plan customers. It also includes where to find pricing information on AWS services such as AWS Simple Monthly Calculator, AWS Services product pages, AWS Pricing API. And finally, it includes alarms and alerts notification and how to use tags in cost allocation.

To summarize, this chapter we covered motivation and value of getting AWS CCP certification, and then we discussed about to how to get AWS CCP certification including number of questions and passing scores. Then we discussed different domains how much percentage of questions are asked from each domain. And then finally, we talked in detail about topics included in each domain. You can find detail further detail about the AWS CCP exam at https://d1.awsstatic.com/training-and-certification/docs-cloud-practitioner/AWS-Certified-Cloud-Practitioner_Exam-Guide.pdf.

References:

- https://d1.awsstatic.com/training-and-certification/docs-cloud-practitioner/AWS-Certified-Cloud-Practitioner_Exam-Guide.pdf
- https://www.zdnet.com/article/the-top-cloud-providers-of-2021-aws-microsoft-azure-google-cloud-hybrid-saas/
- https://aws.amazon.com/blogs/aws/aws-named-as-a-leader-for-the-11th-consecutive-year-in-2021-gartner-magic-quadrant-for-cloud-infrastructure-platform-services-cips/
- https://www.yahoo.com/video/15-biggest-companies-aws-011011152.html



Chapter 2. Cloud Computing Introduction

Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:

- Traditional IT infrastructure
- What is cloud computing
- Cloud computing related terms
- Cloud computing benefits / key features
- Software quality attributes in cloud computing

AWS is a leading cloud provider -- according to the 2021 Gartner Magic Quadrant for Cloud Infrastructure & Platform Services -- with over a million customers of different types in around 200 countries. Moreover, AWS or any cloud provider's underpinning architecture is based on cloud computing. Therefore, the first and most important learning is a solid foundational and conceptual understanding of cloud computing as a cloud practitioner. But before cloud computing, we will discuss some background, mainly traditional IT infrastructure, the reasoning, and motivation for the emergence of cloud computing.

Traditional IT infrastructure

In the late 90s, with the dot com boom, we saw so many startups. Some of them have become big names now, such as Amazon Google. However, most of those startups have started from the so-called garage. First, they started with a few servers. Then, as their user base started increasing, they needed more machines to scale up their business.

Then, to handle the scalability issue, or in other words, to maintain system performance with the matching workload on the system, they moved their server infrastructure from garage to office, where they set up their servers in a so-called computer room or server room. That helped them overcome network bandwidth, power supply, and AC challenges when running the business with more servers.

When the user base increased further, they needed to scale further again. To manage the scalability issue this time, they moved their servers or IT infrastructure to data centers. These data centers have more computing resources, power, air conditioning, security, and other related things that run 24x7 operations of 100s or 1000s servers.

But still, there are challenges and issues with data centers, and what are those? And is there a better solution for this? Let's talk about them.

Depending on how much space you require for your servers, it costs a lot. And there are reasons for the cost as data centers provide 24 x 7 power supply, AC, maintenance, and security. So, it's obvious there will be a cost to all these services.

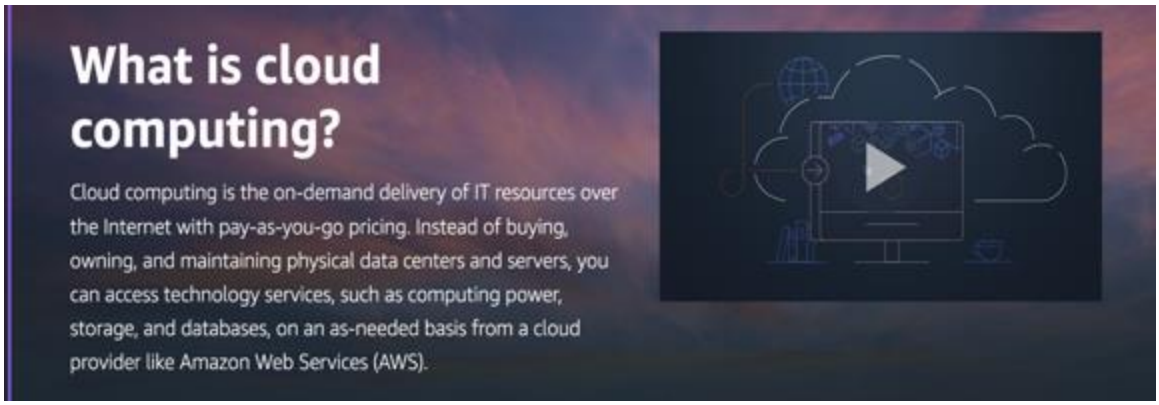
There is limited space – each data centers have some limited capacity. Even though data centers have a vast area, the space is limited. If you need to upgrade servers or do some maintenance, you will have to go to the data center (in many cases) to have the part replaced or do an upgrade, etc. You also need to manage and maintain servers 24x7. There is a single point of failure. What if any natural disaster happens?

So, the bigger general question is -- do we have a solution for all these challenges? Is there any other solution besides leveraging data centers for IT infrastructure? And the answer is: Cloud Computing. So, let's start with cloud computing.

What is Cloud Computing?

Whether we use Facebook, Twitter, Gmail, Zoom (for meetings), or Citrix server (to connect to the workplace remotely), cloud computing is everywhere around us. If we pay little attention, there are countless examples of cloud computing in our day-to-day life. That being the case, you might have an obvious question: what cloud computing is.

Before understanding the term **cloud computing**, it is important to know about the word **"cloud"** as this is an interesting word in this term. Interestingly, the word **"cloud"** in the term cloud computing is not related to the literal "cloud"-- at all. **Instead, the word "cloud" in cloud computing is a metaphor for the Internet.** Thus, cloud (as a metaphor for Internet) computing refers to Internet-based computing in which IT resources are delivered on-demand with pay-as-you-go pricing model.



Screenshot Reference: https://aws.amazon.com/what-is-cloud-computing/

Screenshot Reference: https://aws.amazon.com/what-is-cloud-computing/

Let's talk about a formal definition of cloud computing. According to Special Publication SP 800 – 145 [Sept 2011, Peter Mell (NIST), Tim Grance (NIST)] from The National Institute of Standards and Technology (NIST) of the United States.

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.

There are some keywords to notice in the NIST definition of cloud computing. These are: *on-demand network access*, *shared pool of configurable computing resources*, *rapidly provisioned and released*. On the other hand, in the traditional classic on-premises data center, the computing, storage, and network resources are bought, set up, and permanently configured by the customers in maximum capacity regardless of how much the actual need for help is. Depending on the business season, this resource allocation may be less. In that case, resources are wasted. However, there is also a possibility that the resources cannot meet demand. In that case, there is the chance of reducing service quality and the risk of losing customers because of quality concerns. There is no demand concept, sharing of the resource pool, and rapid on-demand provision in a classic on-premises data center. Another important point to keep in mind is that cloud computing is predicated upon the idea of purchasing "services" based on the needs of customers -- on-demand -- and stop, close the service, or terminate when you are done with the usage.

Using cloud computing, organizations (cloud computing providers) offer services such as virtual machines (compute resource that uses software instead of a physical computer), virtual storage (storage pool formed by combining multiple network storage devices), and many other types of software applications (or services) over the Internet. So, for example, if you would like to set up a

Linux virtual machine, and if you have an account with a cloud provider, you can launch it within a few minutes – just by using the web browser. And start using the Linux VM as you would use any regular physical Linux machine, for example, setting up a web server, database, or any other regular use of Linux machine you do.

In addition to virtual servers, cloud computing providers can also offer virtual storage. For example, if you need extra storage to store large collection of media files, you can use cloud computing provider's storage service to store them – very fast. You just need an account with the cloud provider and a web browser -- no need to shop around to buy the storage and waste additional time to set up the device, such as installing a driver before using the storage. On the other hand, using cloud computing, cloud computing users such as organizations can develop and offer software applications (for example, Gmail, Office365, Facebook) or other related services.

In the above discussion, we learned about the term **cloud computing, cloud computing providers,** and **cloud computing users.**

Based on the above discussion, we can see that to launch a virtual machine or get virtual storage, we only need an account with the cloud provider and a web browser. In other words, cloud computing offerings (the common term is services) are provided over the Internet. Nonetheless, in general, there is nothing special about hardware. Cloud computing's underpinning hardware is the same type of physical server, storage, and network used in on-prem datacenters.

Then, the question comes how cloud computing differs from classic (non-cloud) computing. The main difference is that cloud computing uses cloud architecture. The cloud architecture enables technology components to combine to help build a cloud that can perform resource pooling through virtualization – running virtual machines as an abstraction layer over a physical machine. In other words, cloud architecture helps organize and consolidate massive hardware such as computing resources, storage, and network -- to form resource pooling – and make it available over the Internet.

You may be thinking why there is so much talk about cloud computing. As you have noticed in the above discussion, cloud computing has many advantages. One aspect of the advantages is, though organizations have been developing, delivering, and managing software for many decades, cloud computing has made this process of developing, delivering, and managing software to end users -- globally --much faster and relatively cheaper (cheaper may not always be true). The reason is that hardware infrastructure, software tools, and other whole hosts of things required for software

development, testing, and deployment can be easily and quickly acquired and set up very fast. Additionally, it could be less expensive --- cloud providers nowadays offer various pricing models.

Though cloud computing has many advantages, it may not be appropriate for all use cases. Therefore, you will still need to do your homework if cloud computing is advantageous for your use case or organization.

Cloud Computing Related Terms

Before talking about cloud computing features, service models, deployment models, let's talk about some cloud other computing terms and roles as these terms and roles may be used later. It's better to be equipped with the knowledge of these cloud computing terms and cloud computing roles as these terms are used commonly in cloud computing talks and books.

Cloud Computing Platform

Now we know about the term cloud computing. There is another related term, cloud computing platform. The back-end system providing services is called a cloud computing platform.

Cloud Services (Web Services)

Another related term is cloud services -- also commonly called web services. Services provided by the cloud computing platform are called cloud services, for example, Gmail, Office365.

Cloud Services definition which is based on ISO/IEC 17788, "Cloud Computing – Overview and Vocabulary":

One or more capabilities offered via **cloud computing** invoked using a defined interface.

Cloud Computing Platform Provider

We know the terms cloud computing, cloud computing platform, and cloud services (web services). Another related term to know is cloud computing platform provider. Cloud providers such as AWS, Google, Microsoft, IBM, Oracle, Salesforce, SAP, and others that provide cloud services from their cloud computing platform are called cloud computing platform providers (also commonly called cloud services providers or cloud providers). AWS, Google, Microsoft are the leading cloud computing platform providers.

As a side note, sometimes you will notice that the word "computing" may be missing in some casual or informal discussion of cloud computing. For example, you might hear cloud service(s) as opposed to cloud computing service(s), cloud provider(s) as opposed to cloud computing provider(s), or cloud platform(s) as opposed to cloud computing platform(s). But that doesn't change their semantics.

Below some other cloud computing related terms and their definitions are given. These definitions are based on ISO/IEC 17788, "Cloud Computing – Overview and Vocabulary."

Availability
Property of being accessible and usable upon demand by an authorized entity.

Confidentiality
Property that information is not made available or disclosed to unauthorized individuals, entities, or processes

Integrity
Property of accuracy and completeness.

Information Security
Preservation of confidentiality, integrity, and availability of information. In addition, other properties, such as authenticity, accountability, non-repudiation, and reliability can also be involved.

Service Level Agreement (SLA)
Documented agreement between the service provider and customer that identifies services and service targets. A service level agreement can also be established between the service provider and a supplier, an internal group or a customer acting as a supplier. A service level agreement can be included in a contract or another type of documented agreement.

Cloud Application
An application that does not reside or run on a user's device, but rather is accessible via a network.

Cloud Service Provider
Party (Natural person or legal person, whether or not incorporated, or a group of either) which makes cloud services available.

Cloud Service Customer
Party (Natural person or legal person, whether or not incorporated, or a group of either) which is in a business relationship for the purpose of using cloud services

Cloud Service User
Natural person, or entity acting on their behalf, associated with a cloud service customer that uses cloud services. Examples of such entities include devices and applications.

Measured service
Metered delivery of cloud services such that usage can be monitored, controlled, reported and billed.

Tenant
One or more cloud customers sharing access to a pool of resources.

Multi-tenancy
Allocation of physical or virtual resources such that multiple tenants and their computations and data are isolated from and inaccessible to one another.

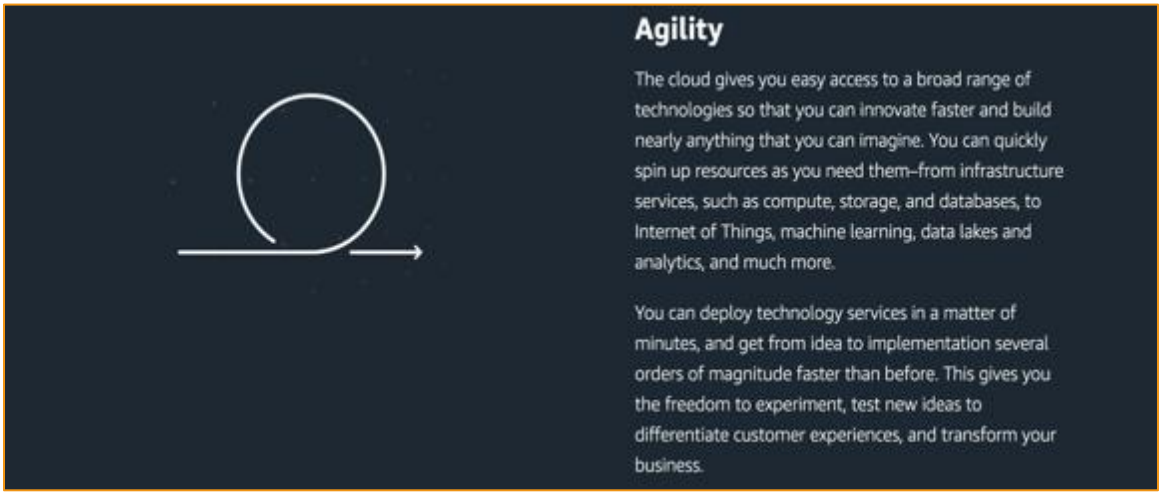
On-demand Self-service
Feature where a cloud service customer can provision computing capabilities, as needed, automatically or with minimal interaction with the cloud service provider.

Resource pooling
Aggregation of a cloud service provider's physical or virtual resources to serve one or more cloud service customers.

Cloud Computing Benefits / Key Features

Let's talk about key features of a cloud computing platform. Let's discuss each of these features in a little detail.

Agility



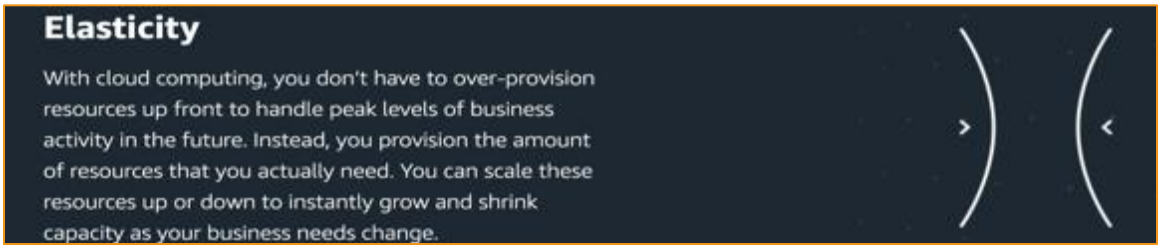
Screenshot Reference: https://aws.amazon.com/what-is-cloud-computing/

Offering broad range of technologies on-demand by cloud providers to its customers is one of the key features of cloud computing. The most common mechanism to offer on-demand service is Web UI. This is not the only way; the other ways are APIs, Command Line Interface (CLI), and programmatic way such as using SDK. If you take example of AWS, AWS users can launch services from using Web UI, AWS API, AWS CLI, and AWS SDK.

Network Access
Network access, more specifically Internet access, is another key feature of cloud computing. In another word, cloud provider must offer services over the Internet to be called as true cloud provider.

Resource Pooling
Another key feature of cloud computing is resource pooling. In fact, main driver for innovation of cloud computing was how to efficiently utilize vast pool of idle resources and generate business model of it. Cloud providers create resource pool of compute, storage, and network resource of difference types, shapes, and sizes and provide suffering offerings from the resource pool based on what has been requested. The beauty of this when the resource is released from the customer, it goes back to the resource pool and ready to be served to another customer.

Elasticity



Screenshot Reference: https://aws.amazon.com/what-is-cloud-computing/

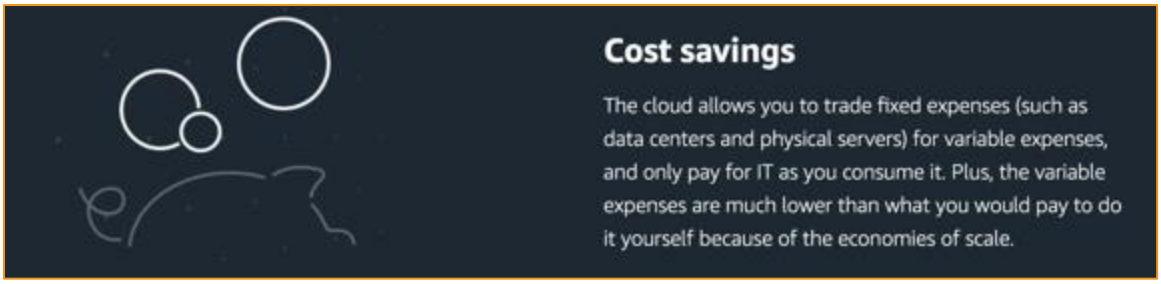
In cloud computing, you will come across the term *elastic* a lot. For example, AWS has many services that include the term *elastic* in its name, for example, Elastic Compute Cloud, Elastic Load Balancer, Elastic Block Storage, Elastic MapReduce.

The term *elastic* in cloud computing is sort of analogous to an elastic band. You can stretch an elastic band size beyond its rest state; and when you let it go, it will come back to its resting size. This elastic concept -- going back to its resting size when we let it go from its stretched state -- is extremely useful in cloud computing.

Let's take an example to understand the term elastic as it relates to cloud computing on AWS. The hypothetical use case is related to setting up a scalable web server. We set up the webserver with a minimum of 3, and a maximum of 6 EC2 instances. Each EC2 instance will be launched using a custom AMI to launch Apache webserver. We have also configured AWS Elastic Load Balancer to launch additional EC2 instances if CPU utilization for an EC2 instance reaches above 70% on AWS Cloud Watch – maximum up to six instances. And terminate the EC2 instance when the CPU utilization comes down to less than 70% -- minimum up to three instances. As you can see, we have set up a scalable and -- elastic -- web server.

Metered Service

Metered Service is very important feature of cloud computing. This concept is very similar to what we have already experienced such as paying for gas utility bill based on number of units of energy consumed or parking meter bill based on vehicle parking time. There are many examples of metered service in our daily life. The metered service of cloud computing is one of key drivers that cloud computing is so popular today in technology – you pay what you use, how much you use.



Screenshot Reference: https://aws.amazon.com/what-is-cloud-computing/

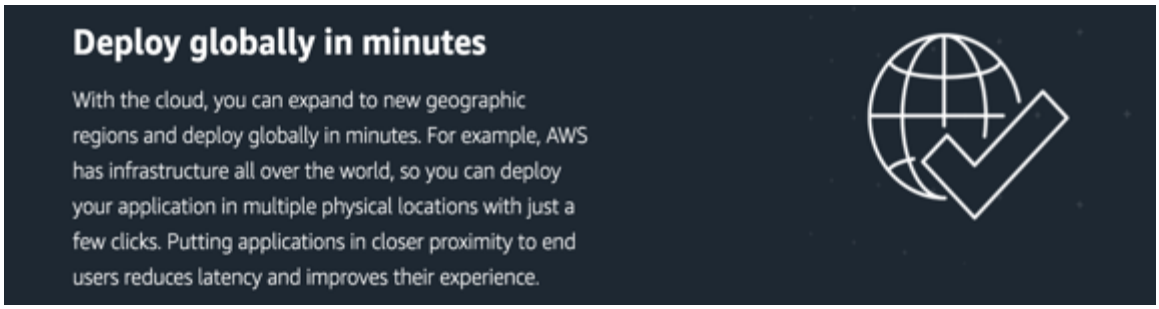
Let's try to understand with a concrete example of metered service in cloud computing. Say you need around 3TB storage to store some videos for a month – may be to share with your friends or relatives

or may by trying out some new business model. Let's first see how we can do it without leveraging a cloud computing solution. We would try to find out some machine which has 3TB space available and then upload videos on this server and then we need to also make sure to have this machine be available on Internet. Sounds easy, technically – right? How much would it cost? It depends how much is your rate and billable hour. Of course, in real life you are not going to calculate it – but there is hidden cost to doing all this setup. In some cases, you may not have the hardware, or have the hardware but not have the storage space. There is also possibility that you may have Internet connectopy issue. You get the idea that to get just 3TB storage space to share your videos with your friends and relatives sounds a bit of involved work technically.

Now let's see how easy how easy to setup this use case on cloud. We will consider AWS storage service S3 (you will learn later about S3). So, the question is how you would set up this use case. You register with AWS if you have. Once you have an account with AWS, you find S3 service, create a bucket and upload all your videos on the bucket. You will get an URL which you can share with your friends. You pay based on how long you have videos on the S3 and what was the size and what was AWS pricing – usually it is in per GB / month.

With metered service feature you pay for the cloud services based on your usage and pricing for the usage. This is very attractive feature of cloud as it saves lots cost in many use cases for almost all type of organizations and individuals who are trying out or learning cloud.

Deploy Globally in Minutes



Screenshot Reference: https://aws.amazon.com/what-is-cloud-computing/

Another advantage of cloud computing is the ability to make applications available worldwide within a few minutes.

Organizations can now quickly deploy their applications to multiple locations around the world with just a few clicks. This allows organizations to provide redundancy across the globe and reduce the application's latency.

Having redundant deployment across the globe helps increase applications availability. In addition, reducing network latency helps improve applications' network performance. The availability and reduced network latency translate the organization's customers' footprint and applications' degree of usability. And the critical point is that the organizations using cloud computing get these at a minimal cost and time.

There is one important point to bring to attention, which is the level playing field. Going global was very expensive in terms of cost and process. Only the organizations having deep pockets could afford to do. But cloud computing makes the deployment of applications globally – a level playing field. With cloud computing, any organization can deploy its applications globally at minimal cost and time.

Chapter 3. Cloud Computing Service Categories

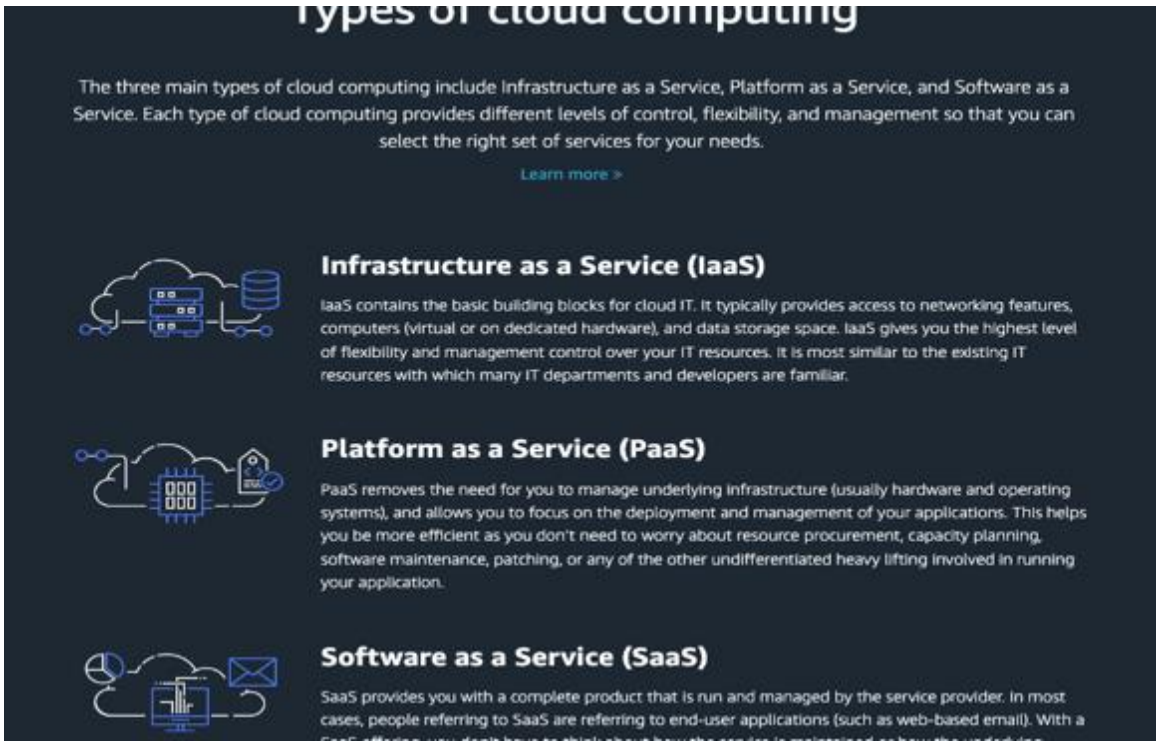
Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:

- Infrastructure-as-a-Service (IaaS)
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)
- Function-as-a-Service (FaaS)
- Scope of Responsibility – how the responsibilities are shared by the cloud provider and the customer in various cloud computing types

There are many names to this term -- cloud computing service categories, cloud service categories, cloud computing delivery models, cloud delivery models, cloud computing platform types, cloud platform types, cloud computing types -- and it's difficult to say which is more or most common. In this book, you may find any of these terms, but they all mean the same. With this note, let's start understanding about this topic.

A cloud computing platform is a back-end system that provides services over the Internet. The question is what kind of services the platform provides. For example, does the cloud computing platform provide Gmail, Office 365? Does it provide virtual servers, virtual storage? Or, does the cloud computing platform offer database services over the internet? Depending on the cloud computing platform's kind of service, it has been categorized into a type. This categorization is called cloud computing platform types or cloud computing types.



Screenshot reference: https://aws.amazon.com/what-is-cloud-computing/

Continuing further on our discussion about cloud computing platforms: cloud computing platform, the back-end system providing services, is a general term. To be more specific, there are three main types of cloud computing platforms: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS).

In addition to these main ones, other modern cloud computing platform types have emerged recently, such as Data-as-a-Service (DaaS), Desktop-as-a-Service, and Function-as-a-Service (FaaS). These modern cloud computing platform types, which provide more fine-grained kinds of services, are getting popular very fast. We will talk Function-as-a-Service (FaaS) in this chapter and ignore the other two as they are out of the scope with respect to the CCP exam.

Infrastructure-as-a-Service (IaaS)

One of the main types of cloud computing platforms is Infrastructure-as-a-Service, which is also called IaaS, in short. IaaS provides foundational types of services also called technology infrastructure that can be provisioned, managed, and maintained over the Internet. In other words, IaaS provides technology infrastructure components.

For example, IaaS offers virtual servers, virtual storage, and virtual network as a service. As you can notice in the given picture, in the cloud computing pyramid diagram, IaaS is at the foundation. What it means is that IaaS acts as the foundation for the cloud computing platform.

The following is the definition of IaaS from the NIST:

The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of selected networking components (e.g., host firewalls).

Let's understand virtual servers in IaaS with a use case. Suppose we need three Linux machines to work on some proof-of-concept (POC) type of work, for example, to build your home-grown load balancer. And we know that once our POC is complete, we will not need those machines further, and we are students with a tight budget. In this situation, IaaS is one of the best options. We could use the IaaS offering from a cloud provider to launch virtual servers and work on our POC. Once the POC is complete, we can terminate the servers and not be charged by the providers anymore.

Like virtual servers, we can also utilize the virtual storage feature of IaaS in situations where, for example, it would be a cheaper or more viable option to use virtual storage than buying physical storage. Let's try to understand it with a use case. Suppose we need temporary storage of around 5TB to store some media files for about a month to share with our friends, and we don't want to use the other video hosting services because of our own decisions. In this situation, as you can realize, utilizing virtual storage from cloud providers would be a better feasible choice than buying physical storage. Once we

decide that we don't need storage anymore, we can delete the media files, and the provider will not charge us.

As we discussed that IaaS provides technology infrastructure components. Some concrete examples of IaaS are AWS EC2 (Elastic Compute Cloud) for virtual servers, AWS EBS (Elastic Block Store) for virtual storage, and AWS Internet gateway for the virtual network.

For virtual servers, the AWS EC2 service is an example of an IaaS type of service. In other words, using the EC2 service, you can launch (AWS term of running virtual servers) Linux, Windows, or Mac virtual servers on AWS. For virtual storage, AWS EBS is an excellent example of an IaaS virtual storage service. EBS is an IaaS type of service as the service provides storage as a service.

We looked up virtual servers, virtual storage examples. With regards to the virtual network, AWS Internet gateway is an excellent example of a virtual network. AWS Internet gateway manages Internet access for the servers launched on AWS.

Key Features and Advantages of IaaS

We got an understanding about what IaaS is. Let's try to understand IaaS fruitfulness. Some of the advantages listed here are common across all cloud computing delivery models.

Eliminate On-premises Data Center's Expense

One of the advantages of the IaaS type of cloud computing platform is that it could eliminate an on-premises data center's buying, setup, and maintenance expense. It's an obvious advantage. When using IaaS -- to procure servers, storage, and network – depending on how much infrastructure you procure, you would be able to cut down huge on your on-premises data-center expenses.

Since IaaS helps reduce data-center expenses significantly, IaaS could be an excellent choice for smaller companies and startups that don't have the resources or time to set up their technology infrastructure. Not only does IaaS help reduce the setup cost of technology infrastructure, but IaaS also takes away the operational expense and the burden of day-to-day managing of computing infrastructure. For example, you can outsource day-to-day tasks such as taking backup, applying patches, ensuring that the system is secured (not a security risk) to the IaaS provider.

Metered Billing

This is a common benefit which customers get in all cloud computing delivery models. In IaaS, cloud customers get bills based on what IaaS type of resources has been used and what was the duration usage – this is very typical how metered billing generally works. With this are many good practical advantages of metered billing, just to mention one here is there is no need to buy and maintain some special high-end servers only for some special need, day, or hour. If you need any servers and there is very likely hood that AWS will have that server available. You launch it and pay only for the usage duration.

Choice of Server Hardware
With IaaS it's much easier get different types of servers. For example, AWS offers not only traditional Intel-based processers, but also offers AMD, GPU, and ARM processor options.

Scalability
IaaS helps in making system scalable. IaaS system can easily provision additional resources to meet the unexpected or planned demand. The reason is IaaS utilizes very large pool of resources.

High Availability
High availability is a general feature of cloud irrespective of a delivery model. With respect to IaaS, since IaaS utilizes very large pool of resources along with redundancy, it is easier to manage high availability by quickly provisioning additional resources or just failover to other available resources.

Security
Cloud providers take responsibility of physical security of servers. In addition, there is network security. In IaaS default no inbound traffic is allowed, and additionally there is security at user level as well, for example, the user has access to the resource or not.

Platform-as-a-Service (PaaS)
Platform-as-a-Service (PaaS) is another primary type of cloud computing platform or cloud computing type. PaaS provides platform technology infrastructure-related services, for example, databases, web servers, messaging, to build, test, and deploy software. In other words, PaaS offers complete development and deployment environment in the cloud.

The following is the definition of PaaS from the NIST:

The capability provided to the customer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider. The customer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

To understand how PaaS relates to IaaS, let's visit the cloud computing pyramid diagram as you can notice that PaaS is above IaaS. What it means is that PaaS can utilize IaaS for its infrastructure-related needs. In other words, PaaS can fulfill its virtual servers, storage, and network-related needs from IaaS.

Now we got the understanding that PaaS offers development and deployment-related services on the cloud. Let's try to understand PaaS with a use case example. Suppose you are VP engineering of a startup with a global team, and even for the local team members, you would like to have the flexibility of remote work. You have heard that platform-as-a-service is a cloud computing type where you can get complete development and deployment environment. So, the question is, what those tools, services, or platforms are that you could consider procuring for your team's software development needs using PaaS.

Depending on your needs, in PaaS computing type, you can find almost anything you require to set up a classic application software development environment. For example, PaaS can offer you IDE (Integrated Development Environment), source code management tools, and build tools. Moreover, in PaaS, you can also get databases, integration tools, web servers, ETL (Extract-Transform-Load) tools, analytic tools, and many more on the cloud platform like AWS. To summarize, PaaS can help you get complete development and deployment-related services on the cloud.

AWS has many PaaS services, such as AWS RDS (Relational Database Service), EMR (Elastic Map Reduce), to name a few. Google App Engine is also an excellent example of PaaS.

Let's discuss further PaaS examples, mainly what we can get on AWS. For IDE, AWS has Cloud9, a cloud-based integrated development environment that lets you write, run, and debug your code with just a browser. For the source code management system, you can use AWS CodeCommit, which is a secure, highly scalable, managed service that hosts private Git repositories. To build a data pipeline and schedule ETL jobs, you can use AWS Glue. Finally, to develop and manage your applications' Docker images, you can use AWS ECS. Elastic Beantalk, which is an easy-to-service to deploy web applications, qualifies as AWS offering of PaaS.

There are many services on AWS that qualify for PaaS. The above ones are just examples to give you an overall understanding of PaaS.

Key Features and Advantages of PaaS

Following are the key features and the advantages of PaaS. You may find that some of the features are overlapping with the cloud service models, that is because all cloud service models inherit general features of cloud computing, in-addition to having specific features for the particular service model.

Multiple Environments

PaaS makes it easier to setup and test applications on multiple environments as it if much easier and quicker to provision different types of environments. Let's take an example, if you were test to an application on Windows, Linux and MacOS operating system, you can launch these environments much quickly and you pay only based on metered billing. You can also create Sandbox environment and provide access to a few developers to try out services to build proof of concept. Sandbox environments are very good to test out applications which are build using AWS services in a separate environment before deploying to production.

Ease of upgrades

Since cloud providers take care of upgrades of PaaS, it becomes much easier of customer. Let's say, for example, you are a data engineer on the AWS platform, and if you are using Jupyter Notebook, or EMR (Elastic Map Reduce) to build data analytic application, if any of these need to be upgraded, it will be incumbent upon the cloud provider.

Cost effective

Since PaaS or any cloud service model used metered billing, customers only pay for the services they are using. If you ran EMR service to run some analytic job for 2 hours, you only for using EMR for the duration you used the service.

Licensing
Licensing is another feature of PaaS. In cloud environment, cloud providers are assumed to be handling and managing licensing of operating systems and platforms as opposed to the organization taking care of this important task to maintain compliance and security of system. Within PaaS cloud service model, the licensing cost are assumed to a part of the metered cost. What it means the cloud provider is responsible for coordinating with vendors to manage licensing aspect of PaaS – not the customer.

To summarize, PaaS provides complete development and deployment-related tools and services in the cloud. Moreover, these services can be accessed anytime on-demand from anywhere over the Internet. Thus, it eliminates in-house buying and setup of databases, web servers, development, and deployment-related tools and services. PaaS can help in setting up multiple environments quickly, upgrade and licensing of PaaS incumbent upon cloud provider, and cost effective.

Software-as-a-Service (SaaS)

Have you used Gmail? Did you happen to watch movies on Netflix? Do you have a Facebook account? Does your workplace use Zoom for meetings? Have you used Microsoft Office 365? If your answer is "Yes" to any of these questions, essentially, you are using software-as-a-service (SaaS).

As IaaS and PaaS, SaaS is another main cloud computing platform or cloud computing type. If IaaS is about infrastructure, PaaS is about the platform – then SaaS is about software. In SaaS, essentially, software solutions are delivered as a service over the Internet. Therefore, SaaS software is mostly executed directly within a web browser. This feature of SaaS eliminates the need to install or download software to execute it.

For example, Gmail, Netflix, Facebook, Zoom, Microsoft Office 365 are some common examples of SaaS. There are countless examples of SaaS, but I'll limit it to a few to keep it simple. To illustrate further, to use Gmail, you don't need to install Gmail on your local computer. You just open a web browser, type the Gmail Web URL in the address bar, and start using Gmail. On the same token, to watch a movie on Netflix, since Netflix is a SaaS solution or SaaS software, you don't need to install Netflix software on your local computer. Just open a web browser, type the Netflix web URL in the address bar, and you're ready to start watching movies on Netflix. As you can notice with these illustrations, that in SaaS, you don't need to install or download software to execute it.

The following is the definition of SaaS from NIST.

The capability provided to the customer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application settings.

SaaS is a very well-known type of cloud computing platform. The reason is SaaS is very visible to the common public usage wise. For example, as we know SaaS software such as Facebook, Netflix, Zoom, Microsoft Office 365 are very popular and have a global reach to millions of users. Additionally, with cloud computing, building SaaS applications have become relatively much faster, which further helps increase its popularity in the developer community.

Comparing SaaS with other main cloud computing types, if you look at the cloud computing pyramid diagram, SaaS is at the top. It means that if you are building SaaS solutions, you can use PaaS for platform-related needs and IaaS for infrastructure-related needs.

Key Features and Advantages of SaaS

The following are the key features and advantages of SaaS. As you have earlier for IaaS and PaaS, some features are general in nature inherited from cloud computing, however some of them are unique to SaaS.

Support Costs and Efforts

In SaaS, we do not need to install any special software. SaaS software can be up and running quickly and can scale as needed. There is a substantial cost benefits for smaller or startup organizations in using SaaS.

Licensing

With respect to cost, SaaS software is typically licensed on a subscription basis. SaaS providers manage all the aspects of software, such as delivery and management, ensuring that service level agreement (SLA) is maintained. Thus, the software is available whenever or wherever the customer needs it, and it performs as per the service level agreement.

Ease of Use

Since SaaS is delivered over the Internet, we don't need to deploy or install any software on your local computer -- we can start using SaaS software as soon as we establish the connection using a web URL. In other words, it can be quickly up and running.

Scalability

Furthermore, SaaS software can be easily scaled as needed. What it means, we wouldn't notice any performance degradation if traffic or the number of users increases.

Standardization

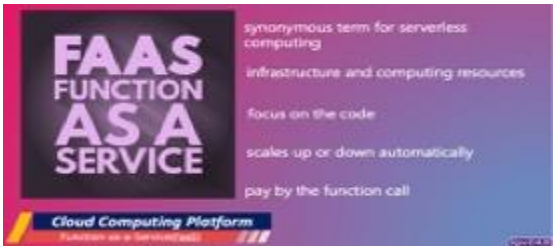
This is very important feature if we compare it with software of 90s. In other words, software which are not of SaaS type. Compare SaaS with the software of 90s where engineers would have visit to different geographical location for the installation of new release or a patch for large software that were installed on multiple locations such as ERP, Manufacturing, and Financials. Since software is deployed centrally, all users get the same screen, same version; and new feature release and patch management are much easier because of centralization. In other words, SaaS helps in having software more standardized in many key aspects.

In the late 90s, before cloud computing dominance, buying, and setting up enterprise software such as ERP, CRM, HR was very expensive. However, SaaS has made a significant difference in pricing, particularly for smaller or startup organizations, which could not afford to buy and set up expensive software such as ERP, CRM, HR, and many. In other words, the subscription-based pricing model of

SaaS has made it much easier for smaller or startup organizations to use or subscribe to costly SaaS software to help grow their business.

Function-as-a-Service (FaaS)

Function-as-a-Service (FaaS), synonymous with serverless computing, is another type of modern cloud computing platform or cloud computing type.

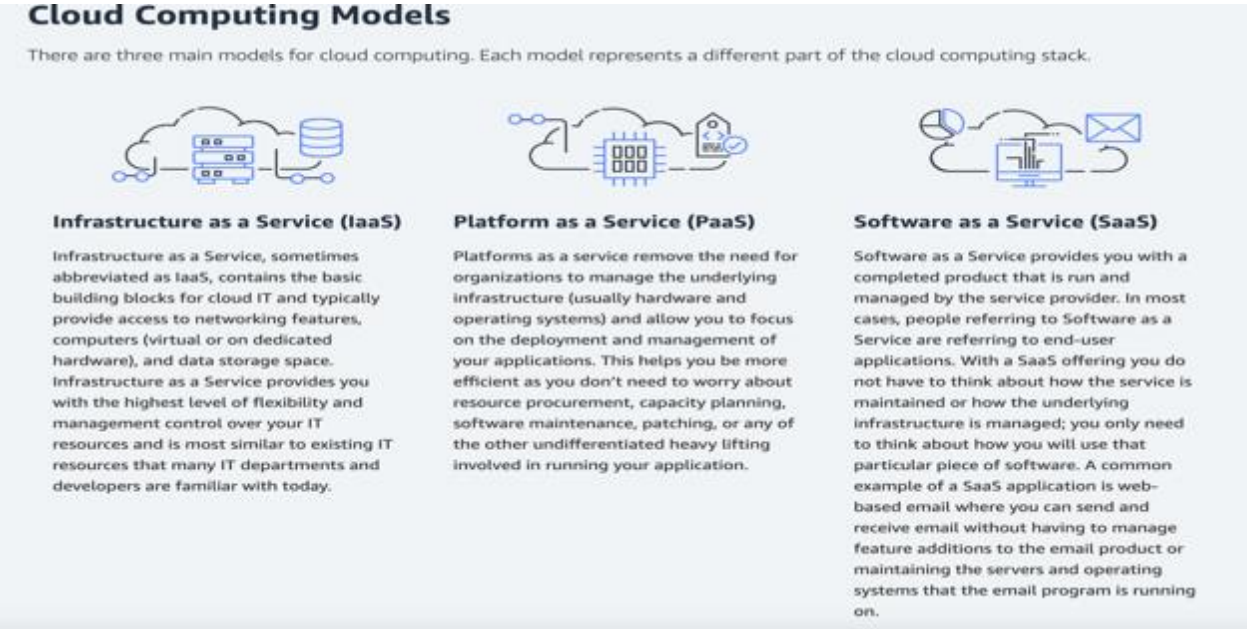


So what does FaaS do? Essentially, in FaaS, users only need to focus on the code (write a Java class, for example) – not on the infrastructure (no need to set up JVM, for example). Users deploy the code having a function (Java class, for example), and the FaaS provider executes the code. The runtime environment is not only provided by the providers but managed as well.

FaaS providers provide infrastructure and computing resources to functions without users setting up the infrastructure and computing resources to execute the process. Additionally, the execution environment scales up or down automatically. Because of the automatic theoretical unlimited scalability feature of FaaS, it is an excellent solution choice for method or function calls, which have a dynamic workload that fluctuates a lot. Moreover, one distinct advantage of FaaS is that we only pay for the computing resources used by function calls – essentially a pay-as-you-go-pricing model.

One of the main drawbacks of function-as-a-service is the execution time. Since process needs to have resource provisioned each time they run, there is a possibility of some performance lag.

One of the examples of function-as-service is AWS Lambda. For example, say we have an image processing function for generating thumbnail images. We can write the process in the language choices, such as Java, Python, and other supported languages by the cloud provider, and let AWS Lambda execute the function. In this use case, we only need to write a function for image processing and configure the computing resource requirement on AWS Lambda. Then, AWS Lambda will take care of the allocation of computing resources and run the image processing function.



Screenshot Reference: https://aws.amazon.com/types-of-cloud-computing/

AWS in general offers more services in IaaS and PaaS category. However, it is important to have good knowledge SaaS cloud service or cloud delivery model as well, particularly its feature and benefits. As you know many popular application software such Salesforce, Gmail, Microsoft 365, Netflix, Facebook, and there many others which are SaaS based.

Chapter 7: CapEx and OpEx

Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:

- Understand the role of operational expenses (OpEx)
- Understand the role of capital expenses (CapEx)

If you have ever been involved in any discussion about cloud computing and you make an argument that cloud computing reduces the cost of IT infrastructure more than the traditional data center-based computing. It is very unlikely that you would be challenged. However, when it comes to ROI on cloud computing, comparing capital expense (CapEx) of running IT infrastructure in a traditional data center with operating expense (OpEx) in the cloud reveals that the cloud is an effective way to reduce CapEx cost.

When it comes to IT spending for businesses, typically, there are two financial models to choose from: capital expense (CapEx) and operational expense (CapEx). So let's explore these two financial models of IT expenditure.

Capital Expenses (CapEx)

CapEx is defined as business expenses in order to gain long-term benefits by buying assets such as buildings or equipment. With respect to IT infrastructure, CapEx examples are an office building, data center, physical servers, desktops, laptops, storage, printers, scanners, generators, software, and other related assets. In this CapEx type of business expense, you invest once, and you can reap the benefits of that business expense many years in the future.

The question is, what **about the** maintenance of these assets? The cost related to the maintenance of these assets also falls under CapEx as the cost extends the lifetime and usefulness of these assets.

From a business accounting perspective, CapEx expenditures are long-term investments for many years. In accounting books, CapEx expenditures are spread over many years in the form of amortization or depreciation. Thus, for accounting purposes, the CapEx form of investment is not a

straightforward deduction for tax purposes. It is also a little challenging exercise to find out the actual direct and indirect value proposition to the company from the investment over the years.

Operational Expenses (OpEx)

Operational expenses are defined as expenses to run day-to-day business operations like services and consumable items that get used up and you need them again to run the business. Examples are stationery supplies, printer cartridges, utility bills, monthly office rent, domain name registration and renewal, website maintenance, and alike. The main idea is that these things are needed to run the businesses, but they are not considered long-term investments like items in CapEx.

From the business accounting perspective, operational expenses can be deducted from the revenue, thus decreasing your tax liability and increasing your profit. Another advantage of OpEx is that if it is not working as intended, you can stop using it or replace it with another alternative, which is a little difficult in CapEx. One potential drawback of OpEx is that sometimes it is difficult to gauge value propositional because of the short-term nature of OpEx – which is typically a year from the business accounting perspective.

CapEx, OpEx and the Cloud

In the above discussion for OpEx and CapEx, it's clear there are steep differences in OpEx and CapEx both from the accounting perspective and the gauge of value proposition provided in both of these financial models of IT expenditure.

The question is what their use cases are. In other words, when to use one over the other.

If you are tight on capital expenditure or if you would want to start very quickly as what you are looking for is already available on the public cloud, the public cloud's pay-as-you-go model could be the right choice –that would be the OpEx financial model. On the other hand, if you would like to have full control of IT resources, you would rather prefer to build a private cloud -- this would be the CapEx financial model. In this private cloud case, your organization will be responsible for all the expenditures.

The current trend is using the public cloud as this is where there are true multi-facet advantages of the cloud. However, organizations having tight, unique security and regulatory requirements tend to think very critically about the adoption of the public cloud. In some cases, these organizations adopt a middle ground of hybrid cloud where they migrate those applications which are non-essential and not in the tight security and regulatory requirements realm. But keep critical and essential applications on private cloud or on-premises data center as it is.

CapEx Approach to Spending

CapEx approach to spending provides stability as expenses are known over the years in the form of amortized value and assets depreciation. The part which usually is grey is about the value proposition. In the technology field, the degree of uncertainty is more as technology is changing very fast. Let's discuss what the risks in the CapEx financial modeling are.

- Buying resources for future
  One of the challenges of the CapEx model is an upfront investment for the future. Technology is changing very fast. And this fast change poses a significant risk in investing today in building something or training staff for something that may be needed in the future.

- Redundant or obsolete resources
  This is another type of risk in CapEx. If the technology changes, those resources such as staff, software, or machine could become obsolete or redundant. Another example is suppose a business builds a private cloud, but the need for the cloud services doesn't grow as the business expected, then the business risk losing money on the investment in building the private cloud.

- Long term business contract
  This is another type of risk in CapEx, where you set up a long-term contract with a vendor but technology or requirement changes, but you are stuck with this vendor contract.

- Maintenance staff salary
  This is another type of risk in CapEx, in which an organization pays the maintenance staff to maintain servers or equipment which are redundant or not utilized.

- Longer buying and setup time
  Some projects take much longer to complete, and they have a risk of not bringing fruitful results as, by the time the project completes, the technology or business has changed. Thus, causing the entire project to be a waste of CapEx investment.

OpEx Approach to Spending

OpEx approach to financial modelling has less risks to expenditure compared to CapEx.

- Purchase IT Resources and Services
  This is the most important feature or sweet spot of the OpEx approach to spending. The OpEx approach to expenditure makes each purchase temporary and helps reduce the risk associated with it. If a vendor fails to make its commitment, or if technology changes, or if your business requirements change to changing market demand, or if your IT budget changes, you aren't locked into one cloud vendor and IT infrastructure.

- Overseeing Over Maintenance
  With the OpEx approach, migrating to a cloud provider for IT infrastructure or IT resources needs, your maintenance staff will have to oversee the resources instead of maintaining them. That staff time can be utilized for other meaningful work to meet the organization's objective.

- Improved Lead Times or Quickly Availability of Resources
  Straight from one of the advantages of cloud computing, the OpEx approach helps improve lead times as there is a high possibility that you can quickly find available resources on cloud providers and start working on your project. Thus, a faster time-to-time market, which many times is the top priority for businesses.

- Agility and Flexibility
  The OpEx approach to IT spending provides agility and flexibility. Trying out something new is much faster and less risky. If the POC (Proof of Concepts) isn't working as you intended, you can try something else. The reason is recourses can not only be readily available, but they can be quickly terminated as well. And with the pay-as-you-model of OpEx, you will only pay for using the services.

Reference:
https://www.10thmagnitude.com/opex-vs-capex-the-real-cloud-computing-cost-advantage/

Chapter 8: Total Cost of Ownership (TCO)

Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:

- What is Total Cost of Ownership (TCO)?
- Understand labor costs associated with on-premises operations
- Understand the impact of software licensing costs when moving to the cloud

---

The cloud computing has made a difference in all types of organizations. However, the most noticeable is that startups generally speaking have taken the most advantage of cloud computing compared with the organizations of the 90s or earlier. The reason is that -- in general -- startups don't have their internal IT setup. Instead, they leverage cloud provider services for their IT infrastructure needs as it saves on CapEx.

What is Total Cost of Ownership (TCO)?

The question that comes to everyone's minds is whether the cost of cloud computing is really very low? To get an answer to this question, we need to understand the concept of Total Cost of Ownership (TCO). Because of high flexibility and low cost, many companies overlook and do not analyze their decisions carefully, which could rise to risk factors such as hidden costs and vendor lock-in. In fact, the Total Cost Ownership (TCO) approach is all about considering and doing analysis all types of direct and indirect costs to mitigate risk factors.

The Total Cost of Ownership (TCO) is the sum of all costs involved in the purchase, operation, and maintenance of a given asset during its lifetime.

The Total Cost of Ownership approach includes the majority of possible costs and cost categories. The TCO approach uses a pre-defined scheme to analyze all possible costs component of an IT artifact. The Total Cost of Ownership (TCO) includes the purchase price, plus operating costs of an IT artifact, over the asset's lifespan. TCO is a way of assessing the long-term value of a purchase of an IT artifact in cloud computing.

In Total Cost of Ownership (TCO), the purchase price of an IT artifact is called CapEx, and the operation costs are called OpEx.

The TCO is intended to help buyers and owners determine the direct and indirect costs of a product or service to improve customer-supplier communication with regard to the entire lifecycle of an IT artifact from a cost perspective.

Businesses, in general, calculating a TCO, overlook many hidden and intangible aspects from both sides – TCO of on-prem IT infrastructure and TCO of cloud. They often apples-to-apples comparisons of the total cost of running servers on-prem vs. the total cost of running servers in the cloud. For example, if a business has 500 servers in operations in its on-premises IT infrastructure. They would look at what would be the rack rate of 500 compute instances of the same CPU, memory, networking bandwidth, and storage space.

This is a good place to start thinking about the cost of migrating to the cloud. However, that should be the final basis of making the final decision-making point. There are many hidden costs such as decreased total time or faster time-to-market, increased productivity, and better handling of elasticity of demand.

Though the goal of TCO is to have a mathematical model of the "real world," certain assumptions are included to reduce the complexity of the real world. Some examples of assumptions are:

- Availability of internal infrastructure for the internet communication and client PCs
- The internal server infrastructure of the company is not included in TCO. The reason is internal server infrastructure needed is not required for the cloud computing services.
- Service provider change is considered to be a new cloud computing deployment. The reason is whenever the service provider is changed, and the effort is the same in the initial deployment.

Cost Types Included in TCO

These are some examples of cost types included in TCO.

- The cost related to the strategic decision on sourcing a cloud computing service which includes as-is analysis of IT infrastructure and business applications, analysis of performance indicators, selection of cloud computing services (IaaS, PaaS, SaaS), and cloud types (public, private, or hybrid cloud)
- The cost related to evaluation and selection of service provider, for example, if the provider provides the services defined in the requirements, evaluations of the provider's offerings with respect to SLA, and identification of the best alternative.
- Implementation of cloud services includes user and group creation, access control, and configuration.
- Support-related costs such as phone, email, ticket, or support via chat and messaging
- The training-related cost which includes internal training by own employees, training by an external vendor for using and operating cloud computing services
- Maintenance and modification-related costs which include testing of service operability, performance, monitoring, and reporting

- System failure related costs which include lost working time, possible cost penalty for non-delivery of services, loss of reputation

Pricing Scheme

Now let's see the general pricing scheme for IaaS, PaaS, and SaaS in TCO. Since this book is for AWS for the brevity and focus, and understanding of concepts, we have included the AWS pricing scheme in TCO.

AWS IaaS Pricing Scheme for EC2

- Price per hour for on-demand instances depends on the virtual machine's RAM, CPU cloud speed and storage, and platform (32-bit / 64-bit)
- Price of data transfers outside of AWS
- Price of reserved instances

AWS IaaS Pricing Scheme for S3

- Price per GB
- Price per transferred GB (outbound); inbound data transfer is free within the same region
- Price per 1000 queries for PUT, POST, COPY, or LIST operations

Most providers charge hourly – usage dependent. Some may charge less than others, but then there is room for an increase in the cost for inbound / outbound data transfer or even charging for internal data transfer.

PaaS Pricing Scheme

There are three types of pricing schemes that are used in PaaS: free of charge, complete package, usage dependent pricing. On AWS, most of the pricing for PaaS is usage dependent if you look at EMR, AWS Glue, or other PaaS type of services such as AWS Elastic Beanstalk.

SaaS Pricing Scheme

SaaS pricing schema is also used-based but simple compared to IaaS and PaaS. You may find it "free of charge" with non-binding and obligatory registration, or monthly charge based on the scope of services, the number of API calls, and the number of users.

Calculating TCO in Cloud Computing

Calculate Current IT Infrastructure Costs

The first step in calculating TCO is to calculate current IT infrastructure costs. This includes all current direct and indirect costs of running and maintaining IT infrastructure. In addition, the business also needs to find out the current workload on its servers, databases, and network bandwidth.

Estimate Total Cost of Cloud (TCO)

When migrating to the cloud, there are two things to keep in mind: many cost components present on on-premises will be off-loaded to the cloud provider. Also, cloud services are not inherently cheaper, and you need to know how to handle cloud costs well. Otherwise, your cost of operating the cloud

spirals out quickly as developers will be launching servers and running available cloud services. Understanding major cost areas in the cloud are key to optimizing your cloud cost.

Two major cost areas are migration cost and the monthly cost of the selected cloud services.

The migration cost depends on how the migration is done. There are different ways to migrate the application to the cloud. According to Gartner, there are five ways to move applications on the cloud: rehosting the application without making any changes, refactoring, and running the application on the cloud provider's infrastructure, revising the application means extending the application, rebuilding or rearchitecting the entire application for the cloud, and last one in replacing the application with commercial software delivered as SaaS. Each way adopted to migrate has its own cost implications.

Another cost area to consider in the cloud migration is the monthly usage cost of utilizing the cloud services. What it means is the monthly cost depends on the workloads and specific cloud services consumed and method of purchase, for example, reserved, spot, or on-demand instances. Since this cost area differs for each organization, leading cloud providers provide pricing calculators to estimate monthly cloud usage costs. For, AWS provides a pricing calculator to calculate estimation based on the AWS products and services selected.

Two major factors that can affect your cloud bills with respect to monthly usage are the types of services consumed and the consumption model. With regards to types of services consumed, commodity services are much cheaper than running machine learning or analytic services. With regards to the consumption model, though on-demand is very popular --but it could turn out to be expensive depending on your usage. Another pricing model is to use a savings plan or pre-paid instances, such as reserved instances.

Training and Consultation Costs

If the organization doesn't have sufficient or no resources, there could be the cost of training or consultant hired for the cloud migration effort.

Labor costs in On-premises Operations

On-premises operations are influenced heavily by labor costs, which include not only the physical security of the on-premises data center but also maintenance and operation of servers to make sure the IT infrastructure is operating as expected. Some examples of maintenance and operation tasks are applying patches, taking backups, setting up new servers, troubleshooting servers, and networking issues -- fixing if something is broken or not running as expected. When businesses migrate to the cloud, these direct labor costs are outsourced to the cloud provider.

Software Licensing Costs in Cloud Migration

Though cloud is, in general, cheaper than on-premises, there are so many moving parts of the costs that, if not managed properly, the costs of cloud migration could give headache.

Software licensing on the cloud is different on the cloud than on-premises. It is often pay-as-you-go and/or subscription-based for the number of users. The best example is salesforce.com, which licenses its usage to the number of users on a subscription basis. This model of licensing works well for a fully integrated software vendor or if you replace existing on-premises software with a cloud-based software solution.

Many big software vendors such as Oracle and IBM have license programs BYOSL – Bring Your Own Software License to the cloud. Under this program, organizations can use their existing on-premises license on AWS EC2. Microsoft also has a similar program called "License Mobility." This offer is limited to a few main popular products such as SQL Server, SharePoint Server, and Exchange Server.

However, as a number of businesses are adopting the cloud, many software vendors have created new licensing models or adjusted the existing ones to make them more flexible. For example, Microsoft is also a cloud service provider, and many of its applications, such as Office, SharePoint, and Exchange software as subscription basis. IBM is also a cloud provider and provides services as a pay-as-you-go model.

There are three factors to look into with respect to software licensing: the number of users accessing the software, the number of processors on the physical hardware on the cloud where the software will be migrated, and special software usage rights for using the software on the virtualized environment. These three factors affect the software licensing when migrating software to the cloud.

References:

- https://ieeexplore.ieee.org/document/6149074
- https://www.cloudzero.com/blog/cloud-tco
- https://www.wired.com/insights/2012/03/licensing-cloud/

Chapter 9: Cost-Benefit Analysis

Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:

- Cost-benefit analysis of a cloud computing platform
- Identify which operations will reduce costs by moving to the cloud
- Right-sized infrastructure
- Benefits of automation
- AWS managed services

You have gone through many chapters about cloud computing and AWS with a great deal of understanding of cloud computing. This chapter is about cost-benefit analysis. There is a saying nothing comes free in the world. Though AWS provides Free Tier, which is very extensive; however, AWS Free Tier cannot be used as a production load. There are many possible advantages to transferring to a cloud platform, depending on how large and old the organization is. Any organization migrating to a cloud or start-up thinking about adopting a cloud must undertake a cost-benefit analysis before taking the final decision.

The cloud may be a very cost-effective option. However, migrating to a cloud environment may not be the right choice if the organization is under tight regulatory constraints or if its existing applications are not ready for the cloud because of architectural coupling or are not designed to be easily portable to a cloud environment. There may be other options, in that case, to migrate to a cloud, such as a hybrid cloud or a private cloud.

Let's focus our discussion on the cost-benefit analysis of moving to a public cloud, as this is one of the main reasons the cloud computing paradigm has become so popular.

Cyclical or Seasoned Demand

Some organizations have cyclical demand for more high-end resources to manage load peaks. For example, organizations in the e-commence domain tend to have more traffic during the holiday season. On the same token, news or media organizations also have a similar scenario where organizations' websites get more traffic if some newsworthy events happen. Generally, to handle

seasonal or unexpected demand, organizations buy and maintain additional servers to make sure that their applications are scalable to distribute the load on these additional servers.

The organizations have two challenges here. The first is an extra upfront cost of buying IT resources such as servers and then maintaining them. And the next is that predicting future traffic is a bit challenging. In a growing business, they may have to keep buying and keeping additional servers each time to manage the new demand. Would that be enough? What if there are budget issues? What if the business scenario changed and traffic decreased? Then, the bought servers would be sitting idle, and invested capital expenditure -- also called CapEx-- in purchasing those servers would provide little benefit to the organization. Managing scalability in an on-premises environment is like chasing a dog's tail from the capital expenditure perspective. There is a strong possibility of resources sitting idle in regular or off-peak hours. This discussion was from a CapEx (Capital Expenditure) perspective. Let's discuss this from an operational expenditure (OpEx) aspect.
Buying additional servers increase maintenance and operation cost as well. System admin or IT admin staff will require to take the work of maintaining additional servers. Organizations may have to hire and add more system admin or operations staff to manage and support additional servers. This would lead to an increase in operating expenses or OpEx.

We got an understanding of CapEx and OpEx challenges in an on-premises environment. Let's see how organizations can address the scalability issue—handling cyclical or unexpected demand that can pop up anytime in the case of news or media organizations. Since the cloud platform provides unlimited resources, organizations can seamlessly drive unexpected or seasoned traffic and pay for operational expenses (OpEx). As you can see that the cloud platform is handy in handling scalability issues just by paying for operations (operational expenditure). In fact, on the cloud platform, there is no capital expenditure for the deployed applications, as cloud providers bear the cost of setting up, running, and maintaining servers on their cloud platform.

Change in Focus

Even though there is no direct cost-benefit of change in focus, however, it's essential to discuss this aspect with respect to migrating to the cloud. Let's see the situation or environment before moving to the cloud.

The transition to the cloud for the organization, which has had an on-premises environment for quite a long time, may not be that easy. Transitioning to the cloud may create uneasiness in employees who run and maintain systems. When the transition is made, their role will be more of overseeing the operation than the direct involvement. And other employees and the department may also have similar uneasiness towards the transition to the cloud. Migration to the cloud in a rush may cause unhealthy environments such as productivity disruption, internal fighting, and employees leaving the organization. On the other hand, many other employees may be very excited and interested in the transition as they may learn new technology and new skills.

The most significant change that happens with respect to change in focus is employees would be more creative in doing proof of concepts type projects or taking on new projects. Project managers would also find migration to the cloud much more helpful as the IT resource needed for the development is much faster -- in almost no time. All development engineers need is an AWS account and the necessary permission to access cloud resources.

Ownership and Control

Organizations in an on-prem data center scenario have complete control of everything -- hardware, operating system, software, and data. This control is very advantageous with the freedom to manage every aspect of IT. However, in the case of cloud migration, the ownership and control parts are shared with the cloud provider.

Nonetheless, cloud providers such as AWS provide lots of flexibility in controlling the system and handling baseline security, which is better than data centers – which is, for the most part, about just renting rack space for servers.

Cost Predictability

In an on-premises data center environment, not only does the organization has in control of hardware, software, and data, but the organization has a predictable cost – the cost of physical IT infrastructure cost of running and maintaining the IT infrastructure. This predictability is changed totally in the cloud environment.
In general, the cloud is known for variable pricing – metered cost. This is because organizations pay based on the usage of resources. This variable or unpredictably type of cost may not be a right for some finance or budget departments of some organizations. Nonetheless, cloud providers have different pricing options, such as reserved instances that can bring more predictability to the cloud expenditure cost (OpEx).

How Does Moving to Cloud Help Reduce Costs?

Right-Sized Infrastructure

In general, most of the on-premises are overprovisioned. According to one research (http://tsologic.com/resources/economics-of-cloud-migration-2017/)more than 80% of on-premises workloads are overprovisioned. There could be a reason behind that. Often, organizations buy server infrastructure to meet the demand of current workloads and the anticipated demands of future workloads. If the workloads need don't increase, the servers run as overprovisioned. This is not a good use of business capital, as you can realize.

When businesses move to the cloud, they can take advantage of its agility and flexibility features. The agility and flexibility lead to leaner delivery processes and thus help cut overall IT infrastructure costs. You can provision resources on-demand on the cloud – no need to overprovision. As the demand for workloads increases, you can provision more resources and de-provision if you don't need them. You can do it manually or automatically using auto-scaling and elasticity features to manage right-size provisioning automatically.

Essentially, a cloud platform can help run your workloads with the right-size infrastructure – no need to over-provision or under-provision the resources.

Utilizing Automation Strategies

Migrating to the cloud reduces the cost of setting up, running, and managing IT infrastructure by leveraging automation. For example, engineers can write scripts to automate backup, storage, code deployment, settings, and configurations. These automation tasks reduce the amount of human intervention needed and allow IT staff to focus on critical business priorities. This is one aspect of saving costs with automation – reducing human intervention. Another aspect of automation is

dynamically provisioning and de-provisioning needed resources at -- *right size (no over-provision or under-provision)*.

Automation is a significant cost-reducing factor in the cloud. AWS has a vast source of APIs to automate almost most of AW'S services without managing them using the AWS Management Console.

Reduce in Security and Compliance Scope

AWS has a concept of the Shared Responsibility Model. Regarding security and compliance, it means responsibilities are shared between AWS customers and AWS. When migrating to the cloud, this is excellent news as organizations' scope of managing security and compliance is reduced when moving to the cloud; for example, the physical security of IT resources is taken care of by AWS. However, in the Shared Responsibility Model, you still will have lots of control over customers' data and how it will be stored and encrypted -- at rest or in transit.

Managed services

AWS managed services -- from database to analytic and logging and monitoring -- provide significant cost savings when moving to the cloud. In addition, these services help reduce operational costs -- coupled with the pay-as-you-go model, the managed services offer flexibility to businesses.

Let's take an example, suppose that you need to store master lookup tables data in an Oracle database with a maximum of 20 tables. If you were to compare it with buying an Oracle license just to store data for master lookup tables vs. using the AWS RDS service – you would probably be inclined to use the RDS service. Cost savings from managed services are a significant component for many essential and ad-hoc type operations.

References:

- https://aws.amazon.com/blogs/enterprise-strategy/rightsizing-infrastructure-can-cut-costs-36/
- https://voleer.com/blog/2019/9/17/reduce-your-cloud-costs-with-these-5-strategies
- http://tsologic.com/resources/economics-of-cloud-migration-2017/

Chapter 10. Cloud Architecture Key Design Principles

Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:
Different cloud architecture design principles

In the previous chapter, we discussed six pillars of a well-architected AWS framework: operational excellence, security, reliability, performance efficiency, cost optimization, and sustainability. The idea behind a well-architected framework's pillars is to help cloud architects build operationally excellent, most secure, resilient, high-performance, and cost-effective IT infrastructure possible for their applications. The Well-architected AWS Framework provides a consistent approach for customers and partners to evaluate architectures. In addition, it guides to help implement design principles that will help scale your cloud applications as they need to grow over time.

Key Design Principles in Building Cloud Architecture

We mentioned design principles that help implement a well-architected AWS Framework. Now, look into crucial design principles that help build well-architected cloud solutions. The design principles are scalability and elasticity, automation, loose coupling, security, caching, cost optimization, think parallel, and design for failure. First, we will start with scalability and elasticity, as these two are one of the most compelling reasons for cloud adoption besides cost and other features.

Scalability

Scalability is the ability of a system to scale without changing the design as input or workload increases. Cloud infrastructure and applications are designed with the premise that the load on the application can grow. In this scenario, if proper mechanisms are not in place in the design, the system will suffer – either the system will stop functioning or underperform. We need to design the system to allow components to be added when demand increases on the system – without changing the design.

Additional components can be added to manage the extra load to drive seasonal traffic. However, automatic scalability is considered a much better design, where the additional system components are added automatically based on the runtime metrics such as CPU, memory, or storage utilization.

Design horizontally scalable cloud applications. There are two ways to manage scalability: horizontal and vertical.

A "vertical scalable" system is considered constrained on resources such as CPU, RAM, and storage, negatively impacting the overall system's performance. Therefore, to improve this system's implementation by the "vertical scalable" mechanism means adding more resources such as CPU, RAM, and storage. However, since there is still no addition of a machine or node, making the system vertical scalable doesn't improve the fault tolerance of the overall design.

A "horizontally scalable" system increases its resource capacity by adding more nodes or machines to the system. If we compare a horizontally scalable system with a scalable vertical design, the horizontally scalable system is preferred over vertically scalable systems. The reason is that a horizontally scalable system helps increase the degree of fault tolerance of the overall strategy and helps improve performance by enabling parallel execution of the workload and distributing that workload across multiple machines. Horizontal scalability helps increase in making the system horizontally scalable. In a horizontally scalable system, since more machines are added to increase the pool of resources, thus if one machine goes down, the other machine is allocated to process the workload of the failed machine. Thus, helping to increase the degree of fault tolerance of fault the overall system.

Vertical Scalability is an old style in which the application is ported to a new server with more CPU, memory, or storage. It could lead to some downtime. The other one, horizontal Scalability, is more modern and a common approach to handling Scalability. In horizontally scalable systems, additional resources such as servers are added automatically to maintain the same performance as the load increases—design horizontally scalable cloud applications.

To summarize, scalable architecture is critical to take advantage of a scalable infrastructure. Increasing resources results in a proportional increase in the system's performance. A scalable service is capable of handling heterogeneity, operationally efficient, and resilient, and it becomes more cost-effective when it grows.

Elasticity

Let's talk about elasticity as a design principle in architecting cloud applications. Elasticity and scalability are generally considered together when architecting solutions on the cloud application. Elasticity is the ability of a system to use resources in a dynamic and efficient way to maintain the SLA as the workload on the system increases and release them as the workload on the system decreases. The deallocation or release of the resources dynamically when they are not needed is the key aspect of elasticity as it avoids the cost of over-provisioned resources such as server, power, space, and maintenance.

Don't assume that components will always be in good health. Don't assume the fixed location of components. Use designs to re-launch and bootstrap your instances. Enable dynamic configuration to help answer instances on boot question: Who am I & what is my role?

Automation

DevOps, in which automation is one of the key features, has become an essential role in many software engineering organizations. Automation is one of the key design principles for architecting applications on a cloud platform. The reason is it avoids human intervention – particularly if it relates to repetitive tasks, integrating systems, or batch jobs. Thus, many operations become more automated

and efficient, and organizations save time on staff – particularly maintenance staff. This frees up some staff time. Time saved from the automation could be utilized on some other high-priority tasks in line with the organization's business objectives. Moreover, with automation with thoroughly tested scripts, we not only automate start, stop, and terminate operations, but we also minimize failures by handling failures in codes. As the system throws an error, we look up the error and fix the script so that next time we don't need to handle it manually. These automated processes overtime make the system resilient -- running with very less human intervention.

AWS has an extensive set of APIs to automate its services. For example, you can easily write a Python script (AWS Python SDK is called boto 3) to automate launching EC2.

Loose Coupling

Enterprise systems have many modules or services (term used in modern micro-services architecture) encapsulating unique business features such as shopping cart service, checkout service, billing service, warehouse service, and support service. These modules are loosely coupled in well-architected systems – typically using web services or messaging frameworks (for example, JMS in Java).

Design everything as a Black Box.

Loose coupling is a key design principle in building any kind of system – even in monolith systems. Loose coupling becomes critically important in building distributed and cloud applications. The reasons are many. We can replace, modify, maintain, or test part of the application in isolation as a separate module or as a separate component by not taking down the entire application, as the price of taking down the entire system could be huge. Imagine if Google, CNN, BBC, Amazon, or any critical applications are going down even for a few seconds for maintenance, to add a new release feature, or fix some bugs.

Security

Security is paramount for any organization, startup, small, medium-sized, or large enterprise. It is even critical for organizations that are handling data related to public health and money. These organizations are also bound to many security and compliance regulations.

Design security in mind. Design security in every layer.

When designing systems, security must be thought of from the very beginning as opposed to thinking and implementing security in bits and pieces when the application is deployed on production. It could be catastrophic if any security-related incident happens.

Broadly speaking, we can divide security into three aspects: physical security, platform security on which application runs, such as operating system and web server, and application security. When it comes to designing the security of data, data should be secured at transit and at rest. In other words, data should be stored in encrypted form both at rest and at transit.

With the cloud, you lose some part of physical control but not your own. A few guidelines related to cloud security. Restrict external access to specified IP ranges. Encrypt data "at-rest" and encrypt data "at-transit" (SSL). Consider an encrypted file system. Rotate your credentials. When passing arguments, pass the arguments as encrypted. Use Multi-factor factor authentication.

Caching

Cloud computing's architecture basis is distributed computing. Distributed computing, on the one hand, by using the basic computing science technique divide-and-conquer, helps to improve processing workload and loose coupling improves modifiability, maintainability, and scalability. There are extremely important features for enterprise systems. However, distributed computing adds in some challenges, for example, more indirection and more layer to communicate to get the final result. This increases latency and thus impacts how fact an output will be retrieved by the end-user

Not all information in the system needs to fetch or calculated each time to process a request. There is any information that changes almost very less, for example, country names, city names, persons' demographic, and such – in fact, master data, or look up values in database terms. On the same token, many contents such as images, videos, or documents in general, don't frequently change in production systems.

Since this information is of mostly static nature, we can leverage caching design principle of cloud architecture to not only improve request processing time but also help reduce operational cost. Data movement from the bottom layer to the top layer is reduced by a few layers reducing data transfer cost. Also, computing resource usage is also reduced due to caching.

As we discussed above that, caching improves request processing time, saves cost on data transfer, and saves cost on computing resource utilization. Let's understand the type of caching. There are two types of caching: application data caching and edge caching.

In application data caching, essentially, data that is mostly of static nature, such as master data, are cached in the in-memory cache. There are many products you can leverage to manage application data caching, such as Amazon ElasticCache (managed Memcache), Redis (in-memory database), and Hibernate Ehcache. You can also implement your custom application data caching for problems smaller in scope.

The other type of caching mechanism which is, by and large, very common in cloud architecture is edge caching. Essentially, for content management, the common caching solution is edge caching. In cache caching, content is served by the infrastructure's edge node server's (AWS Edge Locations) which is closer to the user, thus improving latency and overall system performance. Amazon CloudFront is a typical example of edge caching.

Cost Optimization

Cost optimization is the most important design principle. The reason is cloud costs, to a large part – particularly in the public cloud -- are based on OpEx (operating expenditure) model. Cost optimization essentially becomes an extremely important consideration.

Some principles are common: utilizing the right services for the right duration. For example, if an EC2 medium size instance provides the required performance, then utilizing large or z-large will cost more. If services are being utilized, terminate them or stop them if you are using on-demand instances. You can also consider reserved instances and spot instances as opposed to on-demand instances for EC2 instances to optimize costs.

Auto-scaling is also a very good feature to optimize the cost. Using Auto-scaling, you can not only scale by adding more instances horizontally to maintain performance if the workload increases, but

you can also scale down to terminate the resources if they are needed automatically by adding some configuration using the CloudWatch service.

The main points here are: right service for the right job, and do not use more resources and for more time if you don't need it. Look into various cost options and their pros and cons (for example, on-demand instances, reserved instances, and spot instances) provided by the cloud provider, and select the best option for your use case to optimize the cost.

Think Parallel

Many software engineering problems can be solved in less time if the concept of parallel processing is used. For example, a data processing job can be divided into many parts, and each part can be processed parallel. Map-Reduce job is a good example of parallel processing. Extending on the parallel processing, when you are designing applications to run an on a cloud platform, parallel thinking becomes even more important and valuable as the cloud has massive resources. Parallel processing helps solve large problems in less time.

There are two main reasons for using parallel computing. Parallel computing saves time (wall clock time) and it helps solve large problems. Some guidelines for parallel thinking: experiment different architectures for multi-threading and concurrent requests. Run parallel MapReduce jobs. Use Elastic Load Balancing with Auto-Scaling to distribute loads across multiple machines.

Design for Failure

"Everything fails, all the time," Werner Vogels, CTO, Amazon.com. Design for failure, and nothing will really fail!

A few guidelines: ***Avoid single points of failure -- assume everything fails***. Design with a backward goal as applications should continue to function even if the underlying physical hardware fails, is removed, or replaced.

References:

https://aws.amazon.com/blogs/apn/the-6-pillars-of-the-aws-well-architected-framework/
https://www.botmetric.com/blog/aws-cloud-architecture-desig n-principles/
https://aws-certified-cloud-practitioner.fandom.com/wiki/1.3_List_the_different_cloud_architecture_design_principles

Chapter 11. AWS Well-Architected Framework

Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:

- Pillars of AWS Well-Architected Framework
- Well-Architected Framework General Design Principles
- AWS Well-Architected Tool

Software design is one level below software architecture. Before discussing the design principles, let's briefly discuss software architecture—engineering enterprise software solutions in many ways building civil engineering systems such as building bridges. If the foundation is not architected, designed, and built-in a proper engineering way, the structural building problem may undermine the integrity and function of the building. Or it may cause extension, modification, and repair to be expensive. Building software systems have two types of requirements: functional and non-functional. The software architecture addresses non-functional requirements, for example, performance, reliability, scalability, security, etc.

Cloud systems also have functional and non-functional requirements. What it means is that when building software systems on the cloud, we need to consider quality attributes to build well-architected software solutions. Now the question is what those quality attributes are that we need to consider when building software solutions on the cloud platform.

Architecting Software Solutions on the AWS Cloud

Screenshot from: https://aws.amazon.com/architecture/well-architected/

According to a blog on the AWS Partner Network (https://aws.amazon.com/blogs/apn/the-6-pillars-of-the-aws-well-architected-framework/), there are six pillars of the well-architected framework. These are operational excellence, security, reliability, performance efficiency, cost optimization, and sustainability. Therefore, architecting systems by focusing on these six pillars help produce efficient and stable systems. So, let's discuss these quality attributes for building well-architected solutions on the AWS cloud platform.

Operational Excellence

The Operational Excellence pillar of the AWS Well-Architected framework includes supporting the development team and effectively running workloads most efficiently. The support to the development team and effectively running workloads are critical to a successful cloud platform. The reason is that your engineers will be using the AWS cloud platform to do all their development work. In addition, depending on your business area, you will be running different workloads such as analytic jobs, machine learning-related models, and many other operations. The important points are you need to continuously gain insights to improve the processes and procedures to deliver business value.

There is a saying from the Greek philosopher Heraclitus: "The only constant in life changes." Things change: your customers' requirements and business context may change. Therefore, it is essential to design operations in such a way as they can evolve quickly with the change and incorporate changes from the insights.

Key design principles to consider for the Operational Excellence pillar:

- Perform operations as code:
- Make frequent, small, reversible changes
- Refine operations procedures frequently
- Anticipate failure

- Learn from all operational failures

Security

Security is another essential pillar to consider for well-architected solutions on the AWS platform. The Security pillar includes the ability to protect data and IT assets. You can leverage various AWS security-related services such as IAM, KMS, and other related services to provide security to your solutions. It would be best if you had proper procedures to manage any security incidents. Strong security and operations to handle security incidents help mitigate financial loss and help comply with regulatory obligations.

AWS has a concept of the Shared Responsibility Model. What it means is that the AWS platform protects your physical infrastructure. As a result, this helps you focus on using AWS services to achieve your business goals and not being concerned or responsible for the security of the physical infrastructure, such as servers and other components of a data center.

Key design principles to consider for the Security pillar:

- Implement a strong identity foundation
- Enable traceability
- Apply security at all layers
- Automate security best practices
- Protect data in transit and at rest
- Keep people away from data
- Prepare for security events

Reliability

Reliability is another pillar of well-architected solutions on the cloud. Reliability emphasizes the ability of a system to operate without any failure. Before running your workloads, testing what resources are required for compute, storage, and network helps run the workloads reliably in production. Cloud by design has theoretically unlimited resources. That means you should easily find the resources and services you need to build reliable solutions, for example, AWS Auto-Scaling service, to run workloads without any failure or outage.

To build a reliable system, you will need to anticipate changes such as a spike in workload or changes in the environment – what if the server running workload fails, and other related demands of resources such as extra resources needed when deploying new feature releases. And you will need to take steps such as fault isolation, automated failover to healthy resources, and a disaster recovery strategy to implement resiliency.

Keep these in mind to help you increase reliability:

- Automatically recover from failure
- Test recovery procedures
- Scale horizontally to increase aggregate workload availability
- Stop guessing capacity
- Manage change in automation

Performance Efficiency

The Performance Efficiency pillar includes the ability to use computing resources efficiently to manage the current demand of resources, including when there is a change in requirements – essentially maintaining SLA (Service Level Agreements) by utilizing compute resources efficiently.

The question is how we can ensure we are using resources efficiently. First, we can review our AWS solution to find out if we are using resources efficiently – logs and monitoring will be a good help. We can also review if there is any alternate way to use the system more efficiently. For example, we can tradeoff such as compression or caching to improve managing resources efficiently.

The following design principles can help you achieve and maintain efficient workloads in the cloud.

- Democratize advanced technologies
- Go global in minutes
- Use serverless architectures
- Experiment more often
- Consider mechanical sympathy

Cost Optimization

Cost Optimization pillars deal with the system's ability to deliver business value at the lowest cost. The point here is not to concede service level agreements to save costs. Instead, we must review our choices and if there are alternate ways where we can provide the same business value – go for it. That's the essence of the cost optimization pillar.

Some key design principles to manage cost optimization:

- Implement cloud financial management
- Adopt a consumption model
- Measure overall efficiency
- Stop spending money on undifferentiated heavy lifting
- Analyze and attribute expenditure

Sustainability

The Sustainability pillar addresses how in the long-term, architecture manages the change in business requirements, environment, or economic change.

The following are the key design principles when architecting your cloud workloads to maximize sustainability and minimize impact.

- Understand your impact
- Establish sustainability goals
- Maximize utilization
- Anticipate and adopt new, more efficient hardware and software offerings
- Use managed services
- Reduce the downstream impact of your cloud workloads

Please review key design principles of each of the pillar – there may be question(s) in the exam related to the key design principles of the AWS Well-Architected Framework pillars.

Well-Architected Framework General Design Principles

The Well-Architected Framework identifies a set of general design principles to facilitate good design in the cloud:

- **Stop guessing your capacity needs**: Before deploying an application, you often buy expensive idle resources or deal with limited capacity when you plan to make a capacity decision. With cloud computing, you can use and access the suitable capacity -- as much or as little capacity as you need. In addition, you can scale up and down very quickly as required. Cloud computing helps you stop guessing capacity.
- **Test systems at production scale**: In the cloud, you can create a production-scale test environment on-demand to help set up and perform the complete testing of your application. And then, you can release the resources. Simulating a live production environment is much cheaper because you only pay for what resources you used for the testing.
- **Automate to make architectural experimentation easier**: Automation saves time and money on repetitious tasks and avoids the expense of manual effort when you have to do the same thing next time. In addition, automation helps you track changes, audit the impact, and revert to previous parameters when necessary.
- **Allow for evolutionary architectures**: In traditional classic enterprise architecture, architectural decisions are often slow and implemented as static, one-time events, with a few major versions of a system during its lifetime -- again, a slow and sometimes a bureaucratic process. However, as a business and its context continue to evolve, these initial decisions might hinder the system's ability to deliver changing business requirements. The capability to automate and test on-demand lowers the risk of impact from design changes in the cloud. This allows systems to evolve so that businesses can take advantage of innovations as a standard practice.
- **Drive architectures using data**: In the cloud, you can log and collect data about how your architectural choices affect the behavior of your workload --cost, performance, etc. This helps you make more fact-based decisions on how to improve your architecture. Your cloud infrastructure is code, so you can use that data to inform your architecture choices and improvements over time.
- **Improve through game days**: Try simulating events in production by regularly scheduling game days. This will help you understand where improvements can be made and can also help develop organizational experience in dealing with different types of events.

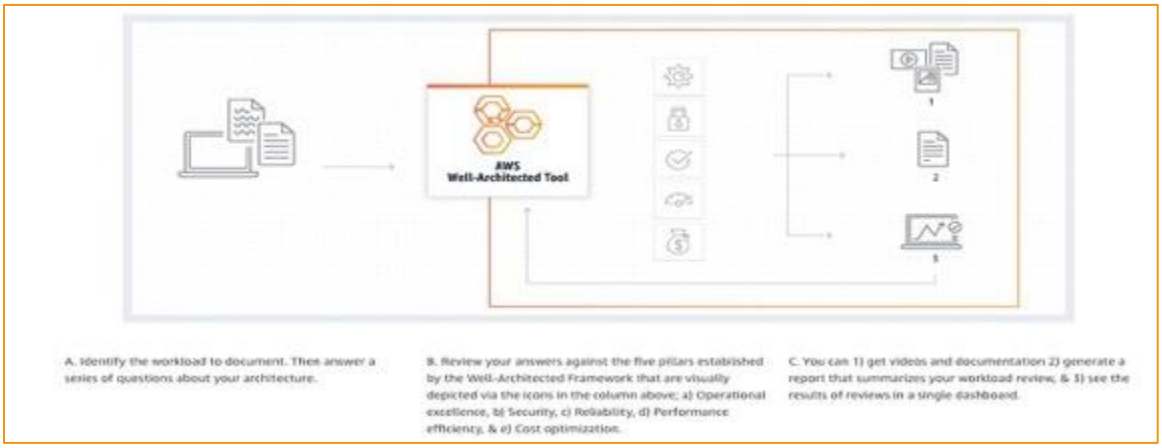
AWS Well-Architected Tool

The AWS Well-Architected Tool guides reviewing the workloads state and compares them to the architectural best practices of AWS. The AWS Well-Architected Tool using the AWS Well-Architected Framework is developed to help cloud architects build secure, high-performing, resilient, and efficient application infrastructure.

To use the AWS Well-Architected Tool, which is available in the AWS Management Console, first define your workload and then answer a set of questions regarding operational excellence, security, reliability, performance efficiency, and cost optimization. The AWS Well-Architected Tool then provides a plan on how to architect for the cloud using established best practices.

The AWS Well-Architected Tool gives you access to knowledge and best practices used by AWS architects whenever you need it. You answer a series of questions about your workload, and the tool delivers an action plan with step-by-step guidance on how to build better workloads for the cloud.

How the AWS Well-Architected Tool works:



References:

https://aws.amazon.com/blogs/apn/the-6-pillars-of-the-aws-well-architected-framework/
https://docs.aws.amazon.com/wellarchitected/latest/framework/sec-design.html
https://docs.aws.amazon.com/wellarchitected/latest/operational-excellence-pillar/design-principles.html
https://docs.aws.amazon.com/wellarchitected/latest/performance-efficiency-pillar/design-principles.html
https://docs.aws.amazon.com/wellarchitected/latest/reliability-pillar/design-principles.html
https://docs.aws.amazon.com/wellarchitected/latest/cost-optimization-pillar/design-principles.html

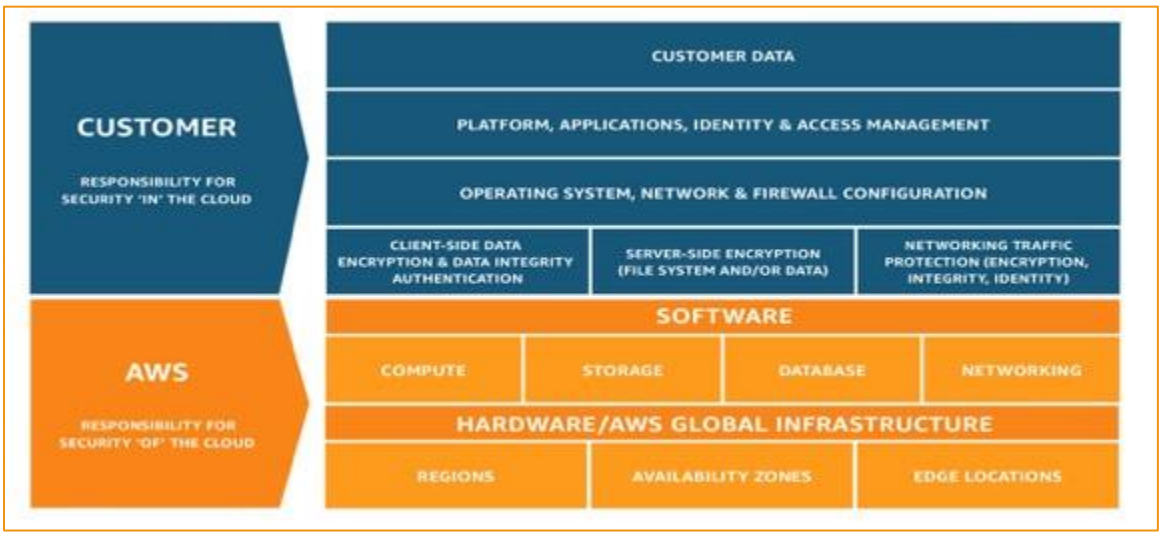
Chapter 21. AWS Shared Responsibility Model

Exam Domain(s): Cloud Concepts

You will learn the following in this chapter:

- AWS Shared Responsibility Model

Security and Compliance are a shared responsibility between AWS and the customer. This shared responsibility model can help reduce the customer's responsibility on the AWS Cloud. What it means is that AWS operates, manages, and controls the components in the host operating system, in the virtualization layer, and in the physical security of data centers. The customer assumes management responsibility of the guest operating system, including updates, security patches, and other associated application software; however, AWS provides a security group firewall. It is essential for AWS customers to carefully consider the services they choose as their responsibilities vary depending on the services used and applicable laws and regulations. This differentiation of responsibility between AWS and AWS Customers is commonly referred to as Security "of" the Cloud versus Security "in" the Cloud.



Reference: https://aws.amazon.com/compliance/shared-responsibility-model/

Security of the Cloud

AWS is responsible for "Security of the cloud." What it means is that AWS is responsible for the infrastructure that runs the Cloud. The infrastructure includes physical hardware, software, network, and physical facilities that host infrastructure and run Cloud services. Based on the AWS Responsibility Model, AWS is responsible for AWS's global infrastructure, which means the hardware and software of AWS Regions, AWS Availability Zones, and Edge Locations. AWS is responsible for computing, storage, databases, and networking infrastructure along with physical facilities hosting data centers for the AWS global infrastructure.

Security in the Cloud

"Security in the Cloud" is the responsibility of the customer. AWS Customer responsibilities depend on the AWS services. For example, the customer has more responsibility and control when the customer is using EC2. In the case of EC2, the customer is responsible for securing the instance by configuring Security Groups and Network ACLs, along with applying updates and security patches. "For abstracted services like Amazon S3, AWS operates the infrastructure layer, the operating system, and platforms" - For abstracted services, such as Amazon S3 and Amazon DynamoDB, AWS operates the infrastructure layer, the operating system, and platforms, and customers access the endpoints to store and retrieve data. It includes the disposal and the replacement of disk drives as well as data center security.

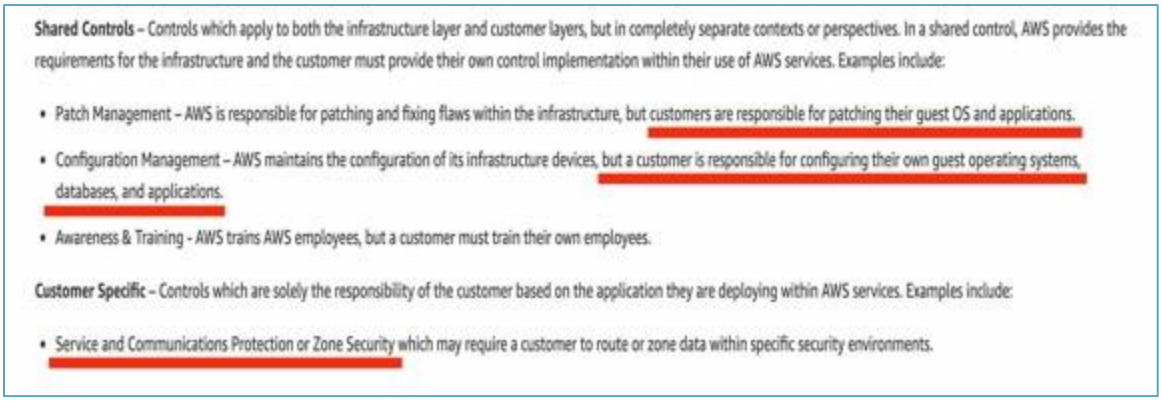
Inherited Controls

Physical and Environmental controls

Physical and Environmental controls are part of the inherited controls, and hence these are the responsibility of AWS. AWS is responsible for protecting its infrastructure, which is composed of the hardware, software, networking, and facilities that run AWS Cloud services. For example, replacing faulty hardware of Amazon EC2 instances comes under the infrastructure maintenance "of" the cloud. This is the responsibility of AWS.

Shared Controls

As we have discussed above, how AWS operations in an IT environment are shared between AWS and its customers. Likewise, management and verification of IT control on AWS are handled between AWS and the AWS customers. I have added a screenshot that shows examples of controls that are managed by AWS, AWS Customers, and/or both.



Screenshot reference:
https://aws.amazon.com/compliance/shared-responsibility-model/

Patch Management

The customers must provide their own control implementation within their use of AWS services. The customers are responsible for patching their guest OS as well as for configuring their applications. AWS is responsible for fixing flaws within the infrastructure.

Configuration Management

Configuration Management forms a part of shared controls - AWS maintains the configuration of its infrastructure devices. However, AWS customer is responsible for configuring their own guest operating systems, databases, and applications. Customers are responsible for the management of the guest operating system, which includes updates and security patches, any application software or utilities installed by the customer on the instances, and the configuration of the AWS-provided firewall (which is called Security Group) on each instance. For example, AWS services such as Amazon EC2 are categorized as IaaS, and as such, it requires that the customer performs all of the necessary security configuration and management tasks.

Training AWS And Customer Employees

Awareness & Training is also a shared responsibility. For example, AWS trains AWS employees, but a customer must train their employees.

OS Configuration
OS configuration as a whole is a shared responsibility but be careful: the host OS configuration is the responsibility of AWS, and the guest OS configuration is the customer's responsibility.

Data Security and Encryption
Under the shared model, customers are responsible for managing their data, including data encryption. AWS is responsible for keeping data on AWS Cloud Secure, Durable, Available, and Reliable. AWS is responsible for keeping the data safe from hardware and software failure.

Enabling Multi-Factor Authentication for AWS accounts in your organization is the AWS customer's responsibility. On the other hand, AWS is responsible for making sure that the user data created and their relationships and policies are stored on fail-proof infrastructure.

Creating bucket policies for Amazon S3 data access is the responsibility of the customer. The customer decides who gets access to the data he stores on S3 and will use AWS tools to implement these requirements. Creating user roles and policies is the responsibility of the customer. Customers will decide "which" resources get "what" access. In the Shared Responsibility Model, customers are responsible for managing their data (including encryption options), classifying their assets, and using IAM tools to apply for the appropriate permissions.

Customer Specific Responsibility
Customers are responsible for Service and Communications Protection or Zone Security which may require the customers to route or zone data within specific security environments.

- Customer is responsible for maintaining versions of a lambda function.
- Under the AWS Shared Responsibility Model, customers are responsible for enabling MFA on all accounts, analyzing access patterns, and reviewing permissions.

Reference:
https://aws.amazon.com/compliance/shared-responsibility-model/

Chapter 23. AWS ML/AI Services

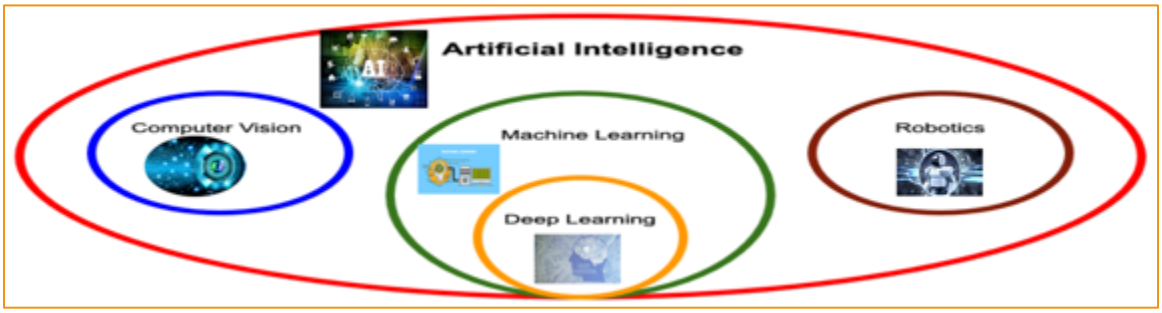
Exam Domain(s): Technology

You will learn the following in this chapter:

- Introduction to ML/AI
- AWS ML/AI Services

Introduction

Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL), though all three are related -- but there are differences.



Artificial Intelligence

Artificial Intelligence (AI) is science that empowers computers to mimic intelligence of human such as decision making, text processing, and visual perception. AI is a broader or umbrella area of where machines are empowered to mimic human intelligence. AI encompasses several subfields such as machine learning, computer vision, and robotics.

Machine learning

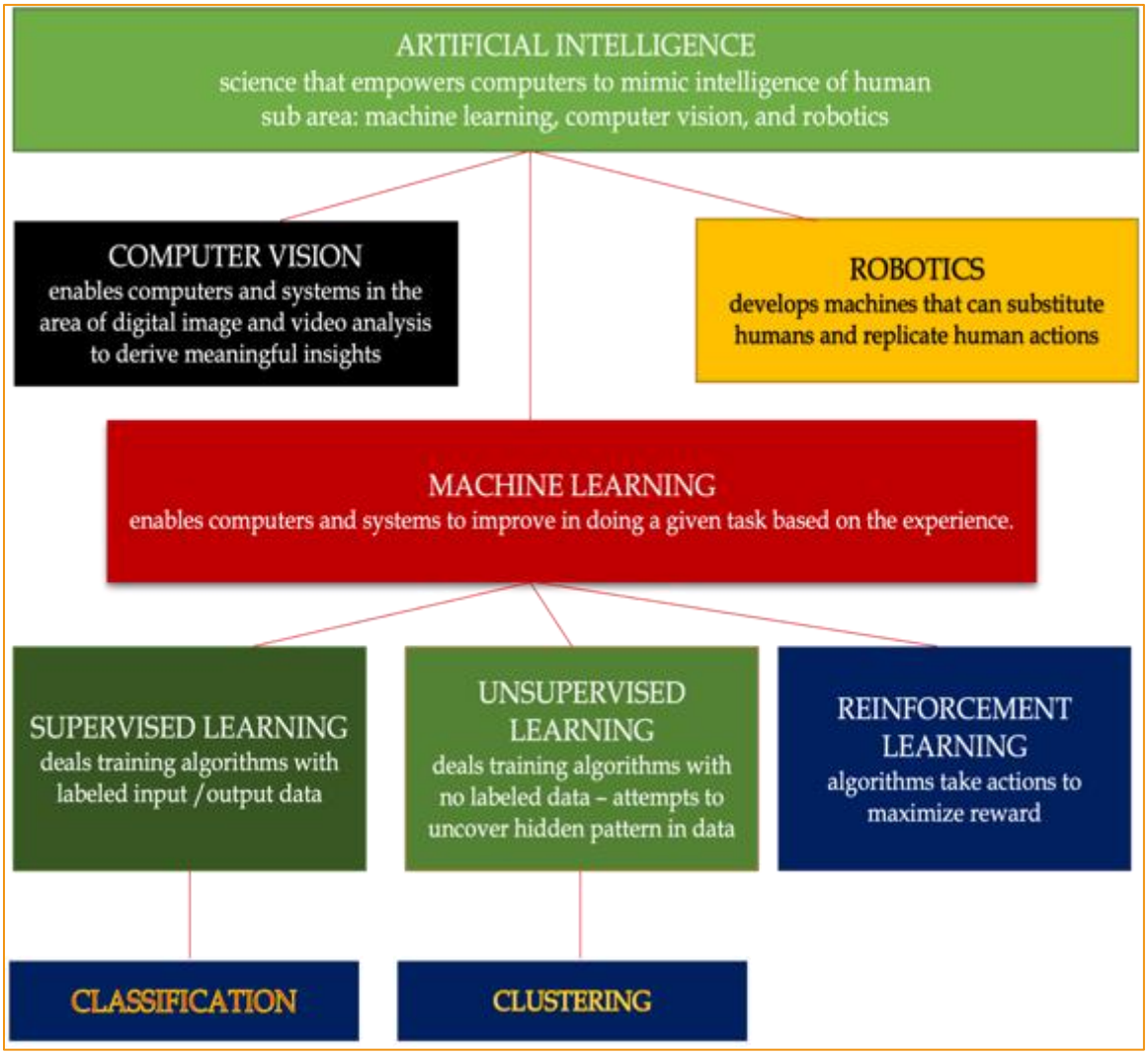
Machine learning is a subarea of AI that enables computers and systems to improve in doing a given task based on the experience. It is important to understand here is that all machine learning can be classified as AI ones. However, not all AI can be classified as machine learning -- there are some rule-

based engines could be classified as AI but not as ML. The reason is since they do not learn from experience, therefore they are not considered to be on ML.

Computer vision
Computer vision is subarea of AI that enables computers and systems in the area of digital image and video analysis to derive meaningful insights and take actions or make recommendations based on that information.

If AI enables computers and systems to think, computer vision enables them to observe and understand.

Robotics is a subarea of AI or an interdisciplinary branch of computer science and engineering. Robotics develops machines that can substitute humans and replicate human actions.
Robots can be used in many situations for many purposes particularly where it's very difficult or risky for humans such as manufacturing processes, or where humans cannot survive for example, in space, underwater, in high heat. Robots can take on any form, but some are made to resemble humans in appearance. (Reference: https://en.wikipedia.org/wiki/Robotics)

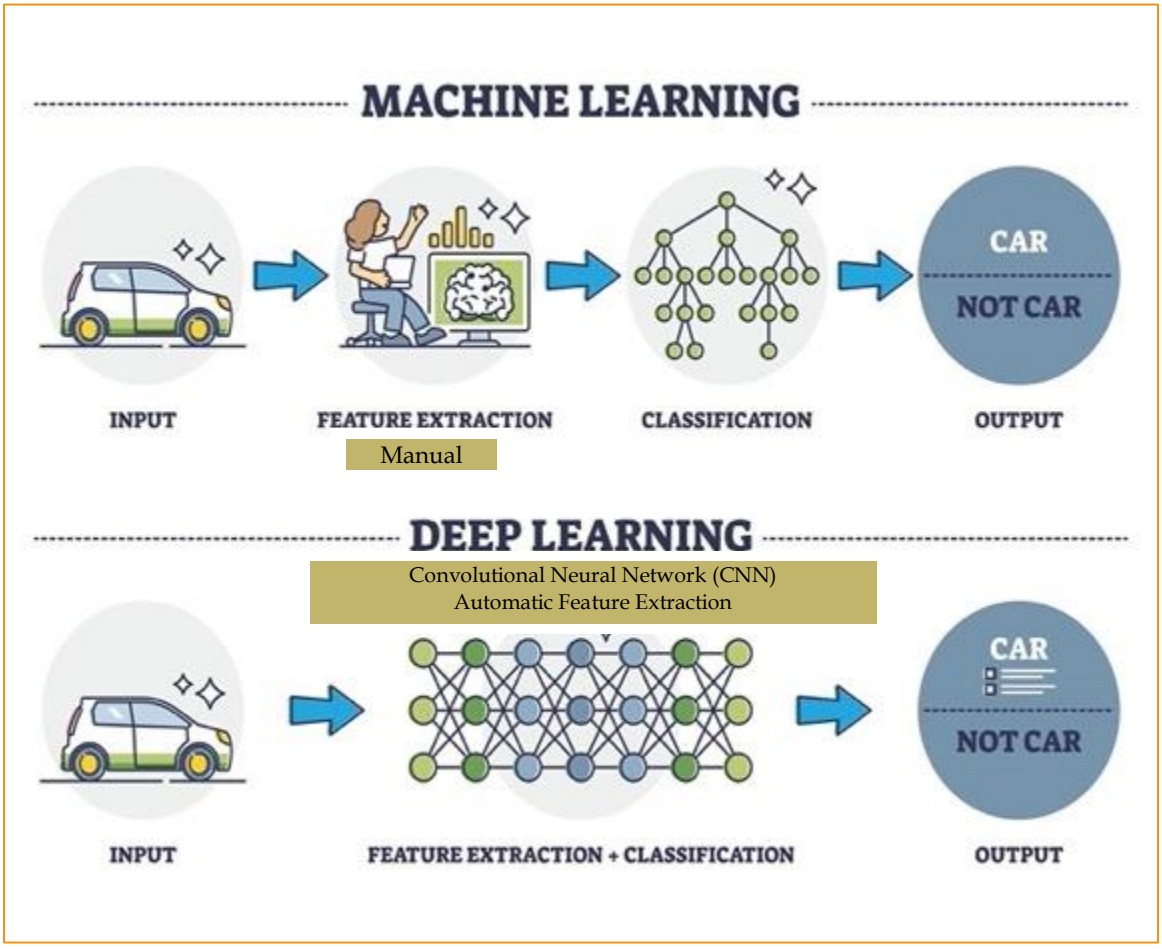


Deep Learning

Deep learning is specialized subarea of machine learning that is based on training of deep artificial neural network (ANNs) using a large dataset such as images or texts. The concept of ANNs is essentially inspired by human brain – how human brain processes information. The human brain consists of billions of neurons that communicate to each other using electrical and chemical signals that enables humans to see, observe, feel, and make decisions. ANNs are mathematically mimicked to work how human brain works – connecting multiple "artificial" neurons in a multilayered fashion. The beauty of ANNs is that adding more hidden layers to the network, makes the network deeper.

The questions what differentiates Deep Learning with Machine Learning.

- In Machine Learning the process is: (1) select the model to train and (2) features are extracting manually.
- In Deep Learning the process is: (1) select the architecture to network (2) features are extracted automatically by feeding in training data, such as image, along with the target class (label).



Amazon AI and ML Services

The Amazon AI and ML Services easily add intelligence to applications. Amazon has many Amazon AI and ML Services; I have listed some main ones below. As we know AWS is an evolving platform, new services are continuously getting added.

- Amazon Augmented AI
- Amazon CodeGuru
- Amazon Comprehend
- Amazon Comprehend Medical
- AWS DeepComposer
- AWS DeepLens
- AWS DeepRacer
- Amazon DevOps Guru
- Amazon Forecast
- Amazon Fraud Detector
- Amazon HealthLake
- Amazon Kendra
- Amazon Lex
- Amazon Lookout for Equipment
- Amazon Lookout for Metrics
- Amazon Lookout for Vision
- Amazon Monitron
- AWS Panorama
- Amazon Personalize
- Amazon Polly
- Amazon Rekognition
- Amazon SageMaker
- Amazon Texttract
- Amazon Transcribe
- Amazon Translate

Amazon Comprehend

Amazon Comprehend applies natural-language processing (NLP) to uncover valuable insights and relationships in unstructured text. Regarding the use cases of this service, it can be used to mine business and call center analytics, for example, to detect customer sentiment and analyze customer interactions to categorize inbound support requests automatically. In another use case, the service can index and search product reviews by focusing on context and sentiment, not just keywords. You can also use the Amazon Comprehend service to secure your documents by identifying and redacting Personally Identifiable Information (PII).

Amazon Lex

Amazon Lex is a service for building conversational interfaces into any application using voice and text. Amazon Lex provides the advanced deep learning functionalities of automatic speech recognition (ASR) for converting speech to text, and natural language understanding (NLU) to recognize the intent of the text, to enable you to build applications with highly engaging user experiences and lifelike conversational interactions.

Amazon Lex is a service for building conversational interfaces using voice and text. Powered by the same conversational engine as Alexa, Amazon Lex provides high-quality speech recognition and language understanding capabilities, enabling the addition of sophisticated, natural language 'chatbots' to new and existing applications.

Amazon CodeGuru

Amazon CodeGuru is a developer tool for improving code quality. It uses machine learning and automated reasoning to identify critical issues, security vulnerabilities, and hard-to-find bugs. You can integrate CodeGuru into your existing software development workflow to automate code reviews. You can use this service during application development and to continuously monitor the application's performance in production, provide recommendations and visual clues on how to improve code quality and application performance, and reduce overall cost. CodeGuru Profiler helps developers find an application's most expensive lines of code and remove code inefficiencies -- Thus, improving performance and significantly decreasing compute costs.

Amazon Forecast

Amazon Forecast uses statistical and machine learning algorithms to deliver highly accurate time-series forecasts – without any machine learning experience. It is a fully managed service. Amazon Forecast provides automation by finding the optimal combination of machine learning algorithms for your datasets. In addition, it offers several filling methods to automatically handle missing values in your datasets.

You can use this service for use cases such as retail demand planning to predict product demand, allowing you to vary inventory and pricing more accurately at different store locations. It can also be used in supply chain planning to forecast the quantity of raw goods, services, or other inputs required by manufacturing. Another use case is a resource planning to predict staffing, advertising, energy consumption, and server capacity requirements. And finally, Amazon Forecast can be used in operational planning to predict levels of web traffic, AWS usage, and IoT sensor usage.

You can use the APIs, AWS Command Line Interface (AWS CLI), Python Software Development Kit (SDK), and Amazon Forecast Console to import time-series datasets, train predictors, and generate forecasts.

Amazon Textract

Amazon Textract service enables you to add document text detection and analysis to your applications easily. Using Amazon Textract, customers can automatically extract text and data from millions of scanned documents in just hours. Amazon Textract has many use cases. For example, you can use Amazon Textract to detect typed and handwritten text in various documents. In another use case, using the Amazon Textract Document Analysis API, you can extract text, forms, and tables from structured data documents. You can process invoices and receipts with the AnalyzeExpense API in another use case. Finally, by using the AnalyzeID API, you can process ID documents such as driver's licenses and passports issued by the U.S. government.

Amazon Kendra

Amazon Kendra is fully managed intelligent search service that adds natural language search capabilities. Amazon Kendra reimagines enterprise search for websites and applications so that employees and customers can easily find the right answers to questions when that they need them. How Kendra does it -- Kendra does it by searching through troves of unstructured data to provide the right answer.

Amazon Fraud Detector

Amazon Fraud Detector is a fully managed service enabling customers to identify potentially fraudulent activities. For example, you can flag suspicious online payment transactions before processing payments and fulfilling orders. In another example, you can detect new account fraud. You can accurately distinguish between legitimate and high-risk account registrations, so that you can selectively introduce additional checks—such as phone or email verification.

Amazon Personalize

Amazon Personalize is a fully managed ML service for real-time personalized recommendations. For example, you can use this service for product recommendations, personalized product re-ranking, and customized direct marketing. Amazon Personalize provisions the infrastructure and manages the entire ML pipeline, including pre-processing, features extraction, applying the best algorithm. Additionally, it then trains, optimize, and deploy the model. You just need to call API endpoints for the deployed model. All data is encrypted, private, and secure, and is only used to create recommendations for your users.

Amazon Personalize supports the following key use cases:

- Personalized recommendations
- Similar items
- Personalized reranking i.e. rerank a list of items for a user
- Personalized promotions/notifications
- To recommend personalized products for users based on their previous purchases

Amazon Transcribe

Amazon Transcribe service helps you quickly add high-quality speech-to-text capabilities to your applications. For example, you can quickly extract actionable insights from customer conversations. In another use case, content producers can use this service to convert audio and video assets into fully searchable content automatically. For example, you can create subtitles to your broadcast content to increase accessibility and improve customer experience. Amazon Transcribe service can be used in the medical field as well. For example, medical doctors and practitioners can use Amazon Transcribe Medical to quickly document clinical conversations into electronic health record (EHR) systems for analysis.

Amazon Polly

Amazon Polly is a service that turns text into lifelike speech. Amazon Polly's Text-to-Speech (TTS) uses advanced deep learning techniques to synthesize natural-sounding human speech. This natural-sounding human speech helps developers build speech-enabled products -- applications that can talk.

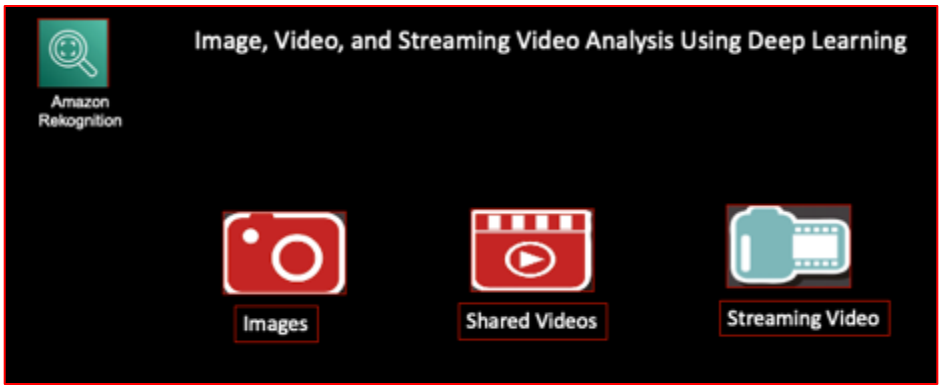
Using machine learning, Amazon Polly offers Neural Text-to-Speech (NTTS) voices, delivering advanced improvements in speech quality. Amazon Polly Brand Voice can also create a custom NTTS voice for your organization's exclusive use.

Amazon Translate

Amazon Translate is a neural machine translation that delivers fast, high-quality, affordable, and customizable language translation. Amazon Translate differs from traditional statistical and rule-based translation algorithms. Instead, it uses neural machine translation, which uses deep learning models to provide more accurate and natural-sounding translations. As a result, the Amazon Translate service can help you localize content such as websites and applications for your different types of diverse users. In addition, it can quickly translate large volumes of text for analysis and efficiently enable cross-lingual communication between users.

Amazon Rekognition

Extracting specific insights in images and videos is costly, time-intensive, prone to error, and hard to scale. Amazon Rekognition is a simple and easy service to quickly analyze pictures and videos stored on S3. It is a fully managed computer vision service that helps automate your image and video analysis, thus avoiding manual inspection. In addition, the service employs proven and highly scalable deep learning technology. Using Amazon Rekognition, you can easily add image and video analysis capability to your applications through simple API endpoints.



With Amazon Rekognition, you get highly accurate facial analysis. You can also perform analysis by matching objects, scenes, segments, and text detection in large numbers of images and videos.

You can automatically tag or label media content to make it searchable. Amazon Rekognition custom labels extend object detection capabilities further, allowing you to quickly train custom models by simply supplying labeled images unique to your business. For example, Amazon Rekognition can help you use custom labels to find images of related to your company, identify products and inventory on store shelves, or classify parts of your assembly line.

You can also flag any inappropriate content to enforce policies, comply with regulations, and protect your brand. You can identify and verify users such as customers and students by matching their faces with their identity document pictures, -- for example, picture ID -- using facial analysis.

Image Analysis

As discussed earlier, Amazon Rekognition has many features. For example, you can perform object and scene detection, facial analysis, face recognition, unsafe image detection, celebrity recognition, and extract text in an image.

Let's look into a little about each feature of Amazon Recognition. With Amazon Rekognition object and scene detection, you can identify keywords that describe the content in an image from object detection like vehicle, sidewalk, or boat to scene descriptions like sunrise, sunset, or beach.

Another feature of Amazon Rekognition is that it lets you perform facial analysis. You just feed Amazon Rekognition an image. It will return many attributes about the photo, such as facial landmarks, emotion analysis, and demographic data like gender and age.

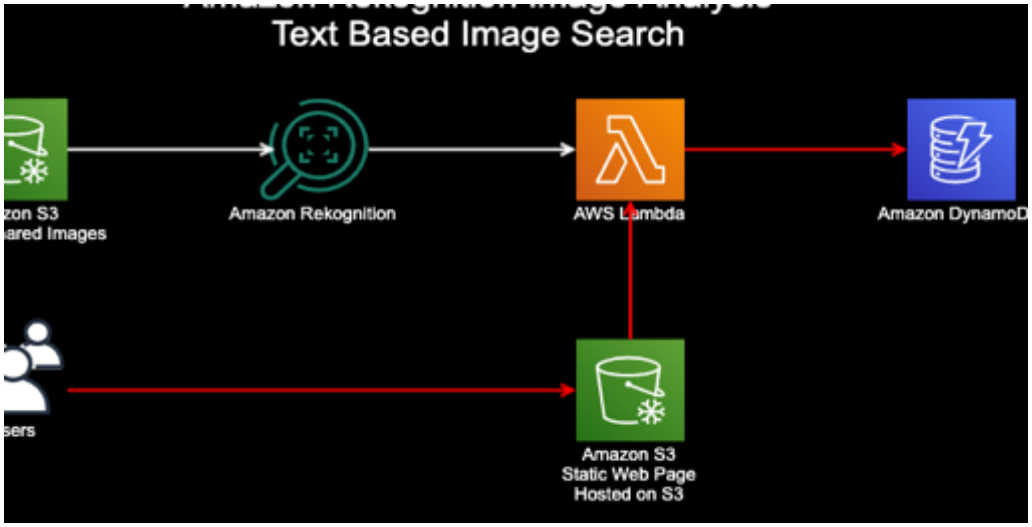
Another feature of Amazon Rekognition is face similarity detection and searching. Rekognition lets you match faces in an image against the index of images you've created. You can also find pictures of one face against faces in another photo, looking for the matches between the two. Rekognition also returns a similarity score which your application can use to determine the possibility of the correctness of the match.

Amazon Rekognition can recognize thousands of famous, noteworthy, and prominent individuals in their field. With each match, the service also provides relevant URLs for more information available.

If your application deals with user-contributed or user-generated content, you might want to ensure that images that your users share are flagged for explicit or suggestive content. Amazon Rekognition can help your image content moderation automatically. Each image moderation response includes a hierarchical list of moderation labels and confident scores, providing insight into the explicit or suggestive nature of the content in the image.

Amazon Rekognition makes it easy to locate and extract text found in a picture even when you're the picture contains real-world scenes like street signs, posts, or license plates. When analyzing the text in an image, Rekognition returns all the detected pieces of text, both individually and in a group, along with a confidence score for every detection.

The typical application architecture when building an application using Amazon Rekognition is for image search using custom labels is as follows:



- The most interesting part is how we use Rekognition to extract relevant information from images so that we can store it in DynamoDB. To get this, we set up a simple pipeline that takes pictures uploaded into an S3 bucket and fires an event that causes an AWS Lambda function to execute.
- This Lambda function calls Rekognition to hit each image's APIs, *DetectLabels* API to detect objects and scenes detection, and *DetectFaces* API to perform face analysis. You could also use other APIs like the DetectText API to find text inside images or *DetectModerationLabels* API to perform unsafe content moderation to extend the application capability further.
- The Lambda function processes the results from API calls, stores the results into DynamoDB, and the results are ready to be queried by our web front end.
- Build a static HTML page for searching and filtering the images hosted on S3
- The search queries call a Lambda function via API Gateway to lookup images tagged with results we obtained via Amazon Rekognition. That is stored inside the DynamoDB.

This is a serverless and event-driven image analysis pipeline. With this setup, it's easy to scale. Furthermore, it is extremely cost-effective because we are not paying for any idle server time because of the serverless nature of the application.

Non-storage and Storage operation API Operations

Looking at the Amazon Rekognition API, there are two types of API operations: non-storage operation and storage operation. The non-storage operation APIs take the input and return the results without persisting in any state. The *DetectLabels* and *DetectFaces* endpoints are both examples of non-storage operations.

On the other hand, storage API operations persist some information to AWS servers to let you make additional API calls that rely on this information. For example, let's say you have a use case in which you need to build an application that will authenticate users by matching pictures of their faces against a set of profile pictures. In this case, you will start by creating a collection on Rekognition to store all of the face data for your users. You do this using the API named *CreateCollection* endpoint. Then you want to analyze and remember the faces of all of your users. Using the IndexFaces endpoint, you can pass in images and the collection's name you wish to Rekogniton to store the face vector metadata. Once you've done this, you can use the *SearchFacesByImage* API to search for faces in the picture to find any match in your collection. The *CreateCollection* and *IndexFaces* APIs are both storage API operations.

Many AWS customers are already leveraging Amazon Rekognition in their products. For example. Pinterest Rekognition to detect and extract the text found in images at scale.

Video Analysis

We have seen how Amazon Rekognition analyzes images – what about video analysis? There are countless opportunities to innovate and find actionable insights from the video content, from personal entertainment to physical security to in-store customer behavioral analysis.
The old classic video analysis technique was to sample still frames from a video feed, performing analysis on these images one at a time. In some cases, this approach can work, but there is a lot of opportunity for improvement here – and Amazon Rekognition does an excellent job of performing video analysis. Amazon Rekognition can perform analysis on videos stored on S3 and can perform analysis of streaming video content as well.

Video is all about motion over time -- this builds up context. For example, the adjacent frames are related, and a good analysis should consider this. It's hard to leverage this temporal context if we just perform a simple frame-by-frame analysis looking at an image in isolation. It doesn't surprise you that Amazon Rekognition does this type of analysis, but it does it in a way that preserves this context. As a result, you can perform object and activity detection, person tracking, facial analysis, unsafe video detection, celebrity recognition, and face recognition for video files.

Let's see how it works. Unlike Rekognition APIs operations for images, which return results immediately -- analyzing videos works asynchronously. We start by storing files on Amazon S3. Then, we call one of Recognition's asynchronous APIs like StartFaceDetection or StartCelebrityRecognition, by inputting the information for the location of video files on S3 and a resource identifier or ARN for the Amazon SNS topic. After the analysis finished, Recognition published a notification on the topic we provided. We connect to this topic to invoke a Lambda function which then fetches the results from Rekognition using the next appropriate call like GetFaceDetection or GetCelebrityRecognition. When the Lambda function gets the results, it stores them in DynamoDB -- ready to be queried from a web app. The web app makes API Gateway calls to call the Lambda function to fetch the result from the DynamoDB.

Amazon Rekognition also can perform person tracking. The person tracking feature lets you follow a person throughout a video, learning their position in each frame and time stamp when they are seen in the video. It can also track people through frames when their faces are obscured or not facing the camera. The GetPersonTracking API returns this location information for each detected person and any detected facial landmarks for each person.

Rekognition also allows you to perform searches of detected faces against a collection of faces you've previously defined. For example, suppose you have defined a collection of faces with the CreateCollection API and populated it with results from the IndexFaces API. In that case, Rekognition can search your videos for matches against the collection using the FaceSearch APIs. The Amazon Recognition also lets you detect explicit or suggestive content in your videos, allowing you to flag or filter content based on the needs of your application.

Streaming Analysis

Amazon Rekognition can do video streaming analysis using Amazon Kinesis Video Streaming. Amazon Kinesis Video Stream is a secure and scalable service for streaming video from connected devices to AWS for analytics, machine learning, and other processing. It automatically scales the infrastructure needed to ingest streaming video data from millions of devices. It also stores, encrypts, and indexes your video data in your streams and allows you to access it through easy-to-use APIs.

To send streaming video data into a Kinesis video stream, you need Producer SDKs. The Producer SDKs -- currently, there are Producer SDKs available for Java and C++ -- make it easy to stream video data to AWS securely. Then, once your streaming data is into the Kinesis Video Stream, you can point to Rekognition for analysis.

Let's see how this works.

There are three parts of a stream when you want to do video analysis. First is an Amazon Kinesis Video Stream to send your video stream content. You would like to use Producer SDK to make it easy to capture the stream from your devices and stream it directly to Kinesis Video Stream. The Producer SDK will also handle stream creation, token rotation, and other actions for reliable streaming. You will also take note of the ARN of Kinesis Video Stream to start receiving the content.

Next, you will need to create Kinesis Data Stream to store the result as Kinesis analyzes streaming video content. You will want to note down this ARN also. That takes care of setting up Source and Destination Kinesis Stream. Now you just need to connect them.

Amazon Rekognition has a built-in stream processer that can receive a video stream, analyze it, and send the results out as it processes the frame. So, start with creating a Rekognition video stream processor providing the ARNs for the incoming video streams and the outgoing data streams you created. Then, Rekognition glues this process together and is ready to run, which triggers one of the Recognition video start APIs like StartPersonTracking or StartLabelDetection, for example. After completing all the parts needed, your application can now consume the data from Kinesis Data Stream.

Use Cases

If your application or app contains user-generated or submitted content, you probably want to know if that content is free, explicit, or suggested. You can set up an automatic moderated pipeline like this. If you upload your content on Amazon S3, you can trigger the Lambda function whenever new content is added to your storage bucket. In addition, the Lambda function can start Rekognition Detect Moderation Label's API. If it gets any label back, you can set the image to be reviewed by a human for selling to your users.

Another use case. Automatic sentiment analysis for a retail store. You can take a stream video feed from a retail store and have Rekognition perform real-time demographic and sentiment analysis on

the faces of the people it sees in the store. You can keep this metadata on S3, periodically load it into RedShift, and plug it into Amazon QuickSight to quickly analyze the content and visualize trends, demographics, and customer sentiment over time.

As you can see, Amazon Rekognition lets you add the power of AI to image and video analysis capability to your application in minutes with only a few lines of code. For example, automatically extract and scene description, recognize and compare faces, detect suggestive and explicit video content, collect demographic details, track people in image and video streams, perform sentiment analysis, etc.

It's easy to get started and cost-effective too.

Amazon Sumerian

Amazon Sumerian is a managed service that lets you create and run 3D, Augmented Reality (AR) and Virtual Reality (VR) applications. You can build immersive and interactive scenes that run on AR and VR, mobile devices, and your web browser. Whether you are non-technical, a web or mobile developer, or have years of 3D development experience, getting started with Sumerian is easy. You can design scenes directly from your browser and, because Sumerian is a web-based application, you can quickly add connections in your scenes to existing AWS services.

Amazon Sumerian leverages the power of AWS to create smarter and more engaging front-end experiences. Easily embed conversational interfaces into scenes using Amazon Lex and embed scenes in a web application using AWS Amplify. Amazon Sumerian embraces the latest WebGL and WebXR standards to create immersive experiences directly in a web browser, accessible via a simple URL in seconds, and able to run on major hardware platforms for AR/VR. Build your scene once and deploy it anywhere.

Reference:

- https://docs.aws.amazon.com/rekognition/latest/dg/recommendations-camera-streaming-video.html
- https://github.com/aws-samples
- https://www.youtube.com/watch?v=v662kWVBmdc

Chapter 29. AWS Networking

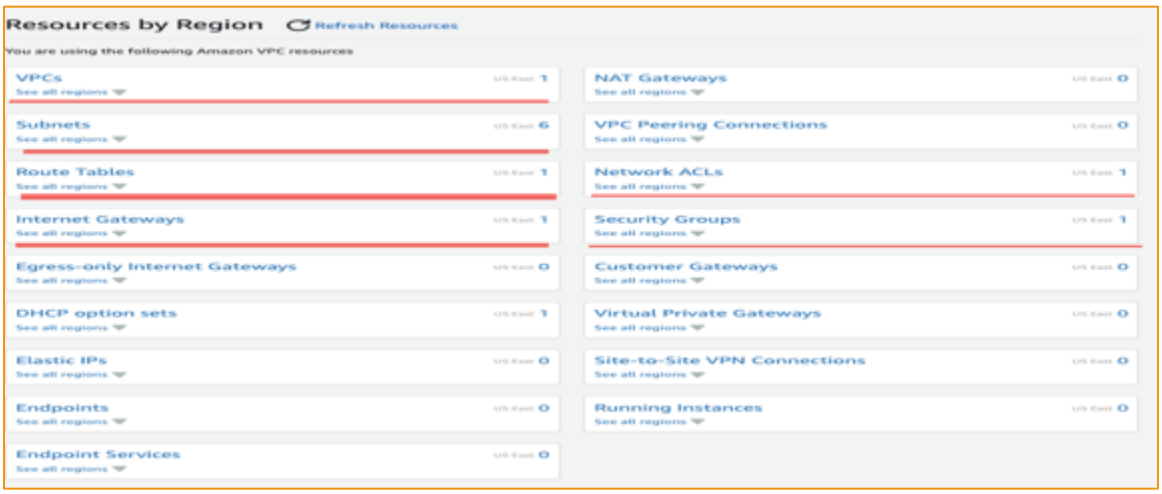
Exam Domain(s): Technology

You will learn the following in this chapter:

- VPC Concepts
- CIDR Block
- Subnets
- Public Subnet
- Private Subnet
- NAT Gateway
- Internet Gateway
- Routing in VPC
- Security Group
- NACLs
- FlowLogs
- VPC Peering
- Transit Gateway
- AWS Direct Connect
- AWS Site-to-Site VPN
- VPC Endpoints

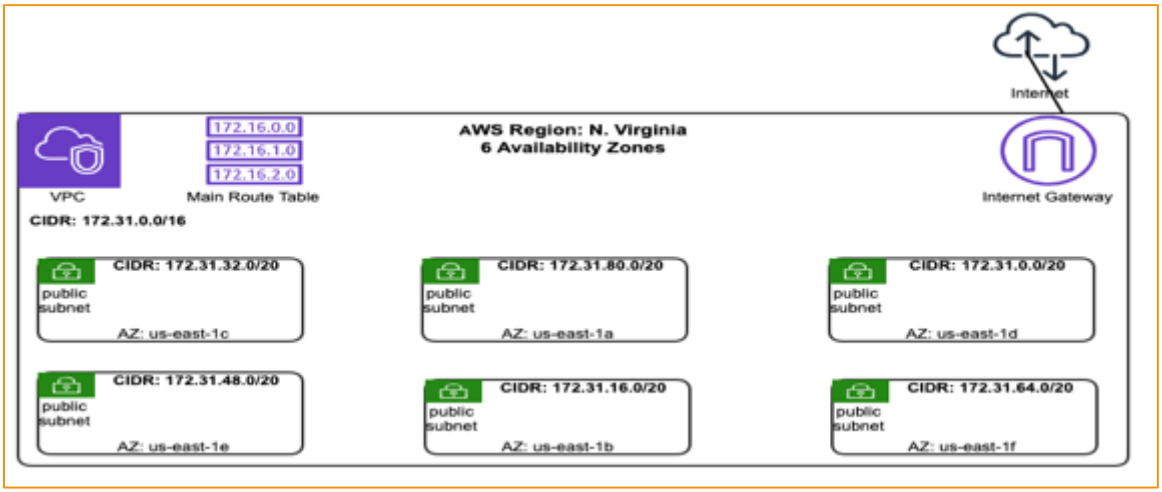
People tend to use AWS as if it were a virtual data center. As a result, they deploy the same kind of stuff in their on-premises data center. For example, they might launch EC2 instances, launch databases, launch EMR instances, etc. These instances need to live in some network and talk to the Internet.

AWS has VPC (Virtual Private Cloud), a private, isolated network to group AWS resources. You also think of VPC as a private virtual network on AWS Cloud.



The above screenshot shows different defaults you get related to AWS VPC in your account.

When you create a new AWS account, AWS gives many things to help start with a network in your AWS account. It provides you with a default VPC with a CIDR range. It provides you subnets for AZs for resiliency. It gives Route tables and Internet Gateway to connect AWS resources to the Internet. It also gives you security groups and Network ACLs to provide security.



Default VPC: AWS Region - N. Virginia

The above diagram is another way to get an idea about a default VPC that AWS gives you when creating a new account. In the diagram, the AWS Region is N. Virginia with 6 AZs. Each is associated with a subnet. The main Route table is related to each subnet in the VPC. There is also an Internet Gateway connected to the VPC to manage Internet traffic. You can notice that the CIDR block is assigned to each subnet and the VPC. There are no hosts shown in any subnet as this is the default setup as no EC2 instance has been launched yet.

Default VPC is a good starting point. Next, we will understand different components of AWS networking to help build your VPC or modify the default VPC for your requirements.

VPC Concepts and Fundamentals

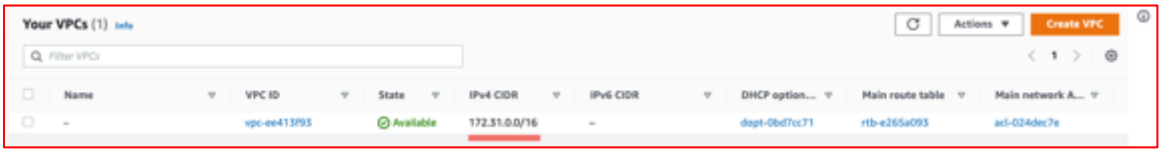
We will learn about IP addresses, how to create subnets, how to set up routing in a VPC, and how to add security.

What is VPC?

Virtual Private Cloud (VPC) is conceptually like a traditional network in a data center, with the additional benefits of leveraging the massively scalable infrastructure of AWS. Virtual Private Cloud (VPC) enables you to launch AWS resources into the network you have set up. A Virtual Private Cloud (VPC) allows selecting your IP address range, creating subnets, and configuring route tables and network gateways. A Virtual Private Cloud (VPC) controls how the AWS resources inside your network are exposed to the Internet.

IP Address CIDR Block

Let's start with IP addresses. At the time of AWS account creation, you get a default VPC. So, for example, in my AWS account, with N. Virginia as the default AWS Region, the CIDR range for the VPC is 172.31.0.0/16.



In the CIDR block (172.31.0.0/16.), the first half portion of the first two octets ("172.31") describes the network portion of the IP address, and the next two octets ("0.0") represent the host portion of the IP address. The "172.31.*" is a private IP range. If you use this IP range in your private VPC, you will not have overlapping or conflict with any other IP address on the Internet. We should not have overlapping IP addresses because overlapping IP addresses can't talk to each other.

Let's talk about the host portion of the CIDR block (172.31.0.0/16). In the host portion of the CIDR block, we have "0.0/16," -- which means we can get around 65536 IP addresses. There are many IP addresses to launch resources in a private network.

When defining your IP CIDR block, think about how many resources you will be launching and choose the IP address CIDR block accordingly.

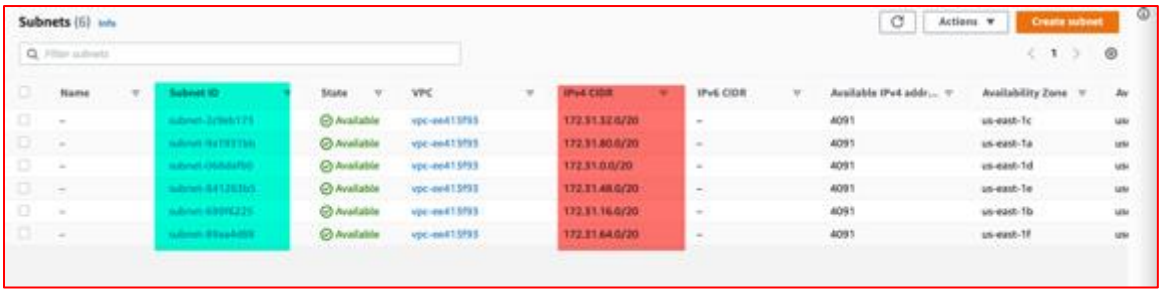
VPC Subnets

What is a subnet? -- A subnet is a range of IP addresses in your VPC. You can launch AWS resources into a specific subnet, such as EC2 instances. When you create a subnet, you specify a subset of the VPC CIDR block for the subnet. Each subnet must reside entirely within one Availability Zone (AZ) and cannot span zones. By launching instances in separate Availability Zones, you can protect your applications from the failure of a single zone.

Ref: https://docs.aws.amazon.com/vpc/latest/userguide/configure-subnets.html

When a default VPC is created, AWS creates subnets based on the number of Availability Zones (AZs) and assigns a CIDR range to each subnet. It establishes one subnet for each AZ.

When creating yourself, assign IP address CIDR to each subnet of the VPC and make sure there are no overlapping IP addresses.



As you can see in the screenshot, in the default VPC with six subnets – there are six AZs in the us-east-1 (N. Virginia) Region. Each of them has been assigned a CIDR range. Each AZ has one or more data centers with a separate power grid, away from the city to avoid floods or other natural calamities.

Each AZ is associated with a subnet in a Region. This subnet association with an AZ helps resources in a subnet talk to resources in the other subnet using their private address yet having logical separation.

When defining your subnet, we need to put subnet in AZs so that they can talk to a subnet of other AZs in the same Region. To do that, we need to divide the CIDR range of VPC into different subnets. For example, from the screenshot of the subnets given above, in the case of subnet-2c9eb173, the IPv4 CIDR is 172.31.32.0/20. That means the network address for this subnet is 172.31.32.0 and starting IP address for the first host is 172.31.33.0, and the ending IP address is 172.31.46.255, with the max possible hosts in this subnet being 4094.

You can use the CIDR calculator https://codebeautify.org/cidr-calculator to get the idea network address and max hosts about other subnets based on the CIDR block address.

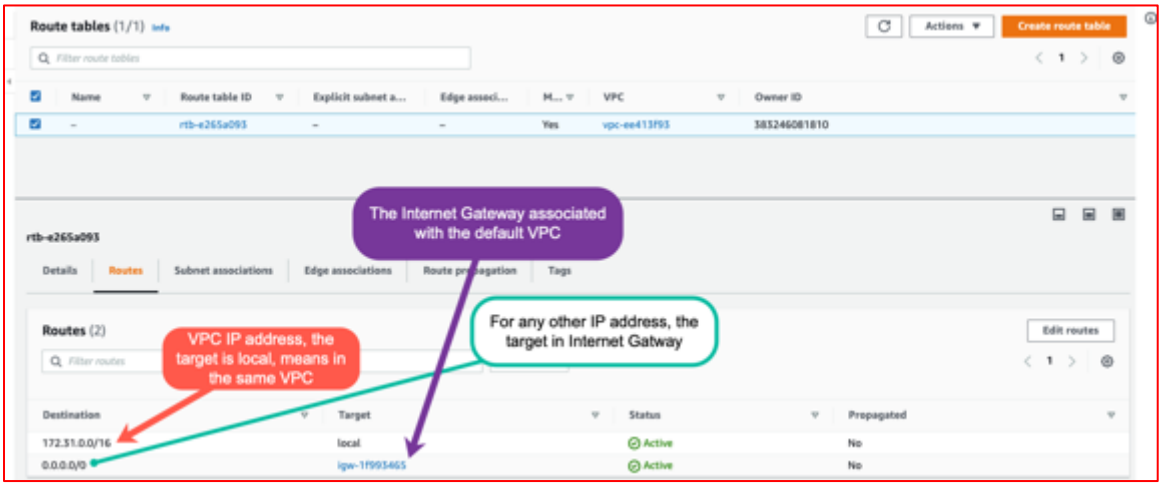
As you can see in the default setting with VPC IP CIDR (172.31.0.0/16), the first two octets are used for the network portion for subnets.

Routing in VPC

Routing is critical to talk to two addresses on a VPC. How two IP addresses talk to each other – the rules – are contained in a Route table. In other words, a Route table includes rules about how two IP addresses talk to each other on a VPC or where to send the next packet.

A routing table specifies how packets are forwarded between the subnets within your VPC, the internet, and your VPN connection.

Every VPC has a default Route table, in which there is a rule that says a target for every request in CIDR range is inside the VPC.



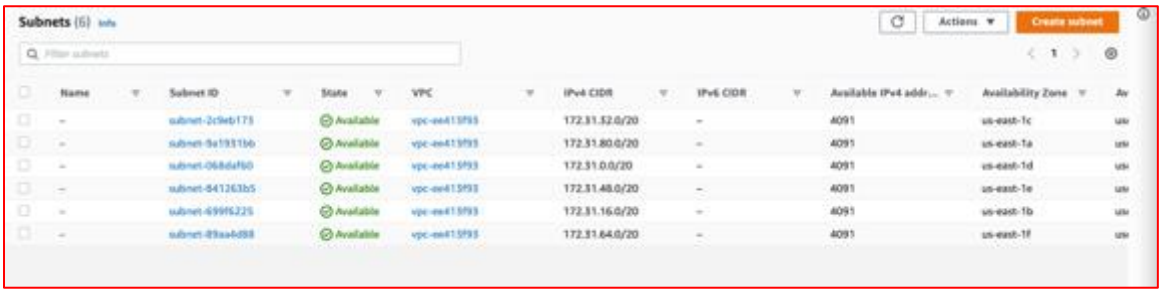
As you can see, this is the screenshot of the default Route table for the default VPC. When a destination is the IP address of the VPC, the target is local – means in the VPC. If the destination IP address is not from the VPC IP address CIDR, the target is Internet Gateway, associated with the default VPC.

You can also create a Route table and assign it to any subnet in your VPC. Then that Route table will replace the default route table of the VPC.

What About VPC Resources Talking to the Internet

For resources in VPC to connect to the Internet:

- Your VPC needs to have a connection to the Internet. You will get the default Internet Gateway when you get your AWS account.
- You need a route to the Internet Gateway from the Route table associated with the subnet. Next, you need to have a public IP address for the resource in the subnet trying to connect to the Internet.



The above screenshot shows all the subnets in the N. Virginia Region in my AWS account. Since there are six AZs in the N. Virginia, that's why if you notice in the screenshot, there are six subnets.

In the case of default subnets that AWS creates for the default VPC, each of them, by default, is associated with the default route table. As you saw earlier, that default Route table has an entry to the Internet Gateway. In other words, if we launch an EC2 instance in a subnet created by AWS for the default VPC, the EC2 instance will get a public IP address.

Please keep in mind that any subnet in which the Route table has an entry for the Internet Gateway means the subnet is public. Therefore, all subnets associated with default VPC is, by default, a public subnet unless we add a different Route table that doesn't have an entry for Internet Gateway.

Private Subnet

In many use cases, you need a private subnet where you want AWS resources not reachable from the Internet. In the case of a private subnet, the Route table associated with the subnet doesn't have association with the Internet Gateway. In other words, there will be no entry for the Internet Gateway in the Route table.

A typical use case for EC2 instances running in a private subnet is a scenario in which instances should not be accessible from the Internet. Still, they are, however, allowed to connect to the Internet to download any software or get an update about their installed software.

The question is how to allow a host inside a private subnet to access the Internet. The answer is NAT Gateway. Unfortunately, there is no default NAT Gateway created with the default VPC.

NAT Gateway

What is NAT Gateway? A NAT gateway is a Network Address Translation (NAT) service. You can use a NAT gateway so that instances in a private subnet can connect to services outside of your VPC. However, the external services cannot initiate a connection with instances inside the private subnet.

When you create a NAT gateway, you specify one of the following connectivity types:

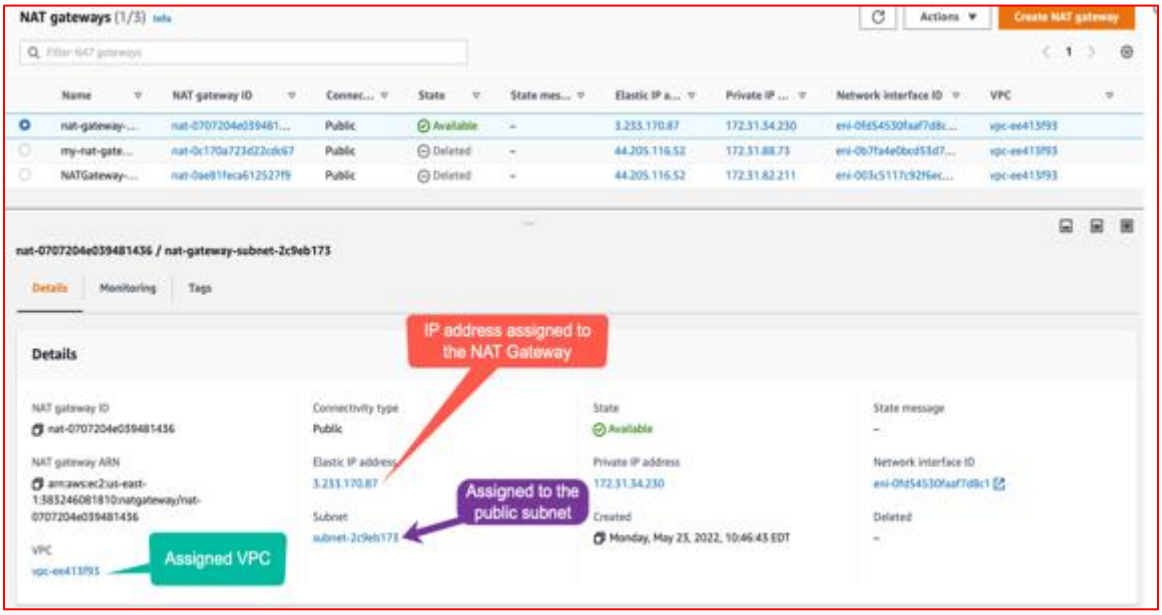
Public – (Default) Instances in private subnets can connect to the internet through a public NAT gateway but cannot receive unsolicited inbound connections from the internet. You create a public NAT gateway in a public subnet and must associate an elastic IP address with the NAT gateway at creation. Then, you route traffic from the NAT gateway to the internet gateway for the VPC. Alternatively, you can use a public NAT gateway to connect to other VPCs or your on-premises network. In this case, you route traffic from the NAT gateway through a transit gateway or a virtual private gateway.

Private – Instances in private subnets can connect to other VPCs or your on-premises network through a private NAT gateway. You can route traffic from the NAT gateway through a transit gateway or a virtual private gateway. You cannot associate an elastic IP address with a private NAT gateway. You can attach an internet gateway to a VPC with a private NAT gateway, but if you route traffic from the private NAT gateway to the internet gateway, the internet gateway drops the traffic.

Reference:
https://docs.aws.amazon.com/vpc/latest/userguide/vpc-nat-gateway.html

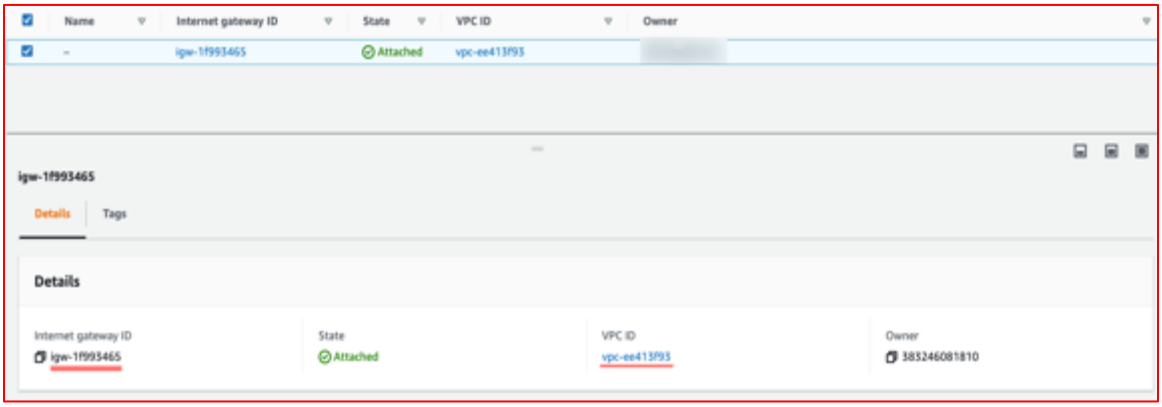
Suppose you would like hosts in the private subnet to connect to the Internet. You will need to create NAT Gateway and associate the NAT Gateway to the subnet.

Please note that the NAT Gateway is one way street to connect. What it means inbound traffic from the Internet to the NAT Gateway is not allowed.



The above screenshot is related to a NAT Gateway. The NAT Gateway is assigned an Elastic IP address, which is a must. For devices to connect to the Internet and find and talk to other devices or machines on the Internet, a public IP address is necessary.

The next point to notice is that it is assigned to a VPC, and VPC is connected to the Internet Gateway.



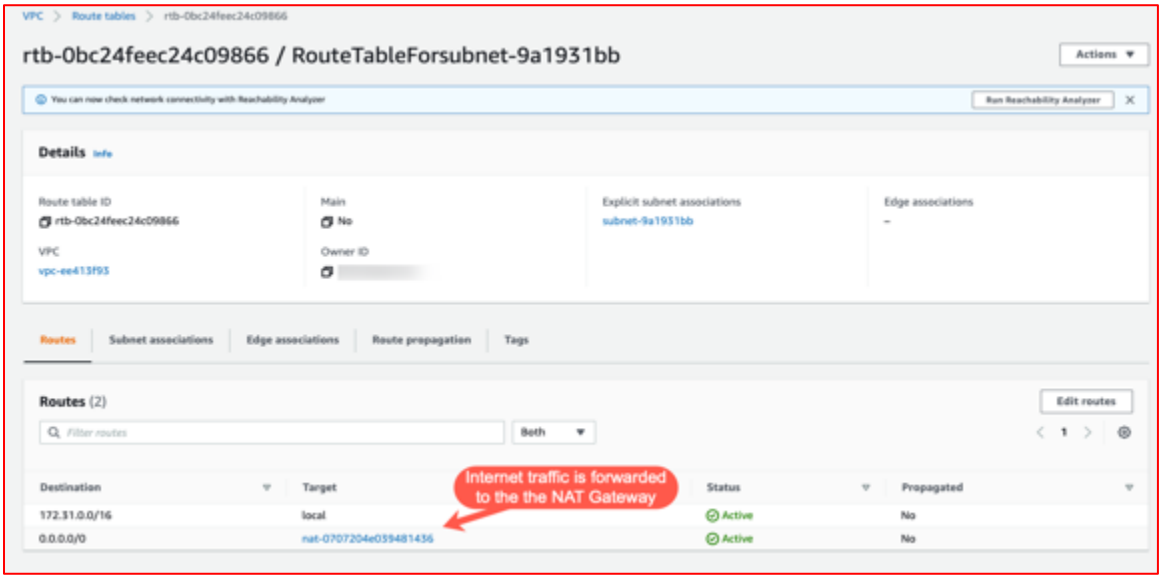
As you can notice in the above screenshot of an Internet Gateway, it is connected to a VPC.

Going back to the screenshot of the NAT Gateway, another vital point to notice is that it is assigned a public subnet. It is an important point – a NAT Gateway must be associated with a public subnet.

You have created a NAT Gateway in the public subnet. The NAT Gateway is associated with a VPC assigned Internet Gateway, which will take care of Internet traffic. Please keep in mind that NAT

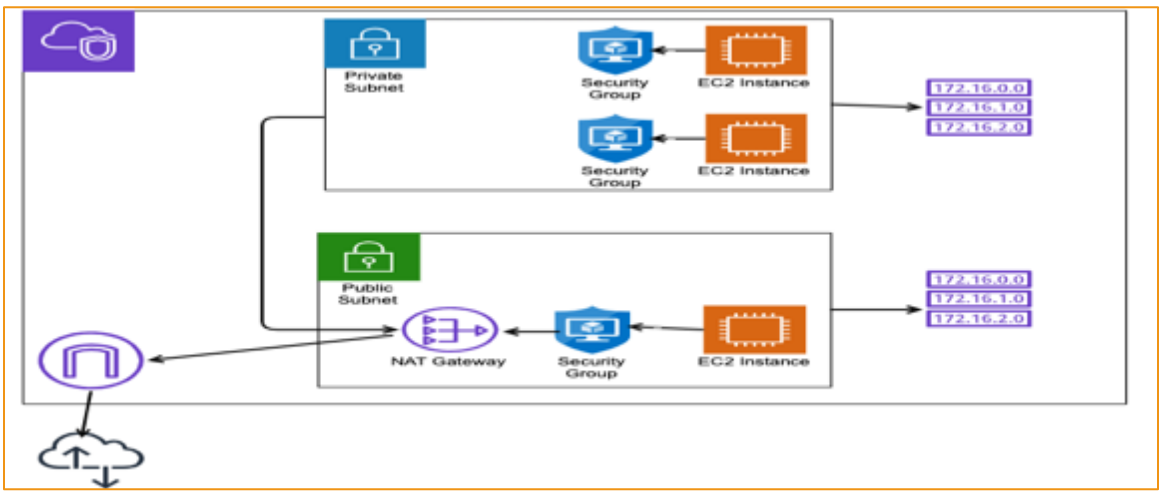
Gateway is a device for one-way traffic. In other words, a NAT Gateway allows making outbound calls to the Internet from resources in the private subnet -- not the inbound calls from the Internet to the subnet. The typical use case is to get patches or install software on the machines in a private subnet.

The next question is: how will a host in a private subnet connect to the NAT Gateway to go to the Internet, for example, download a patch or a software.



The answer in the Route Table of the private subnet is straightforward: add an entry for the Internet traffic to forward to the NAT Gateway, as you can see in the screenshot above.

For hosts in a private subnet to make outbound requests to the Internet, in the Route table associated with the private subnet, an entry will be added to direct Internet 0.0.0.0/0 traffic to the NAT Gateway, as you can notice in the screenshot above.



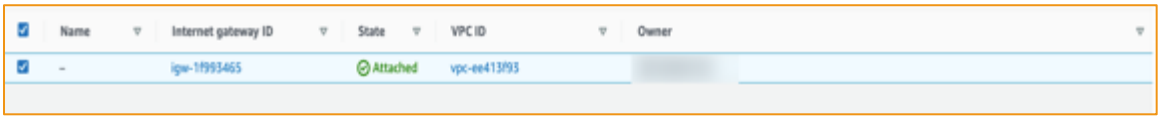
Setting up NAT Gateway

The above diagram shows setting up NAT Gateway in the public subnet to access the Internet from the hosts in the private subnet. If you notice, traffic from the private subnet goes to the NAT Gateway. And from there, then, it is sent to the Internet Gateway connected to the Internet.

NAT Gateway and Internet Gateway

- Internet Gateway allows Internet access to public AWS resources -- within a VPC.
- NAT Gateway allows Internet access to AWS resources – that are in a private subnet.

In other works, Internet Gateway is way out to the Internet for your public AWS resources. However, NAT Gateway, which is connected to the Internet Gateway, allows your private AWS resources way out to the Internet.



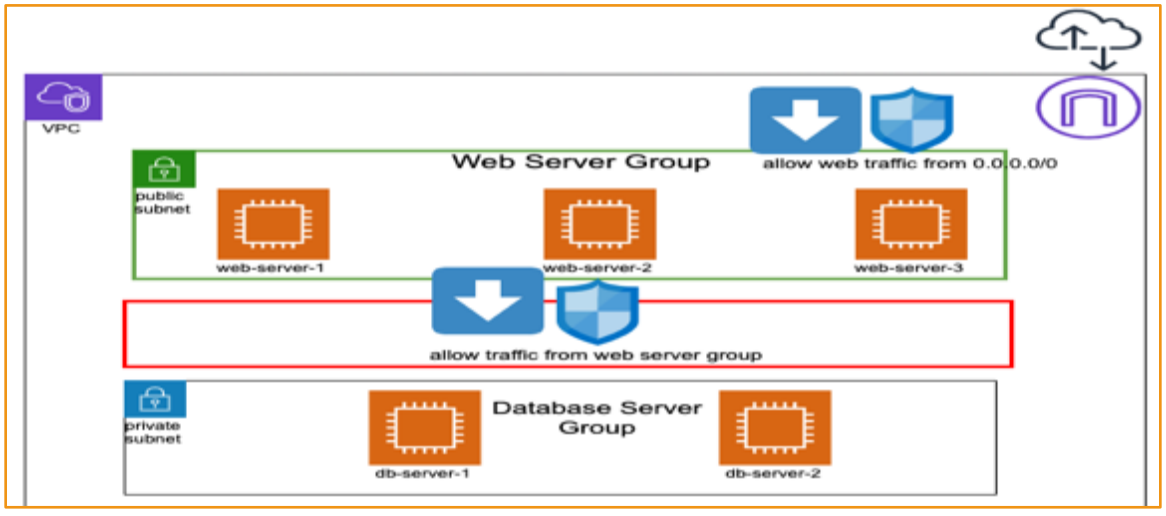
Screenshot showing Internet Gateway in the default VPC.

AWS Network Security

When we talk about AWS network security, it is essential to understand security group, Network Access Control List (NACLs), and VPC Flow Logs.

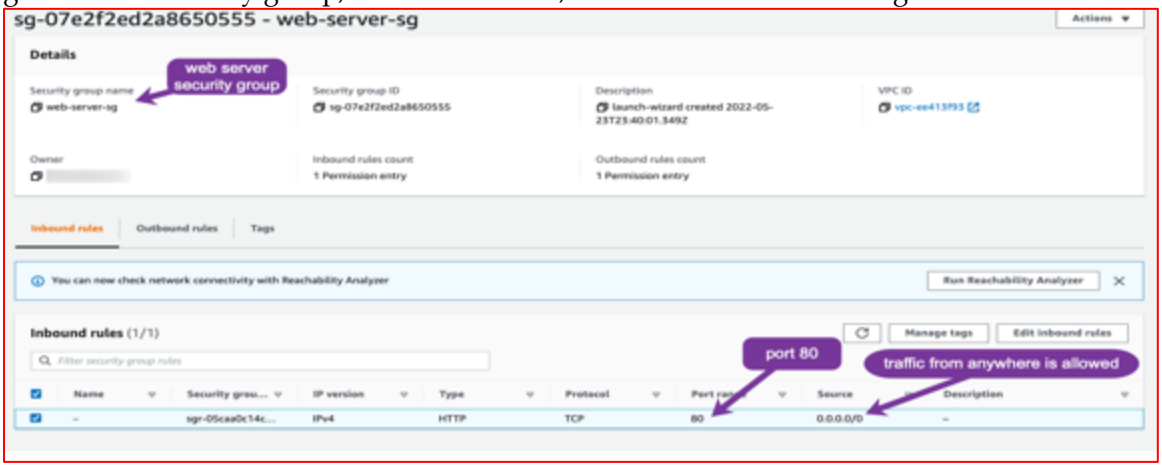
Security Group

What is a security group? Security groups are AWS distributed firewalls. The important point to know about the AWS Security Group, or in general, about any firewall, is that they are stateful. So, for example, if a request is allowed, a response is automatically allowed.



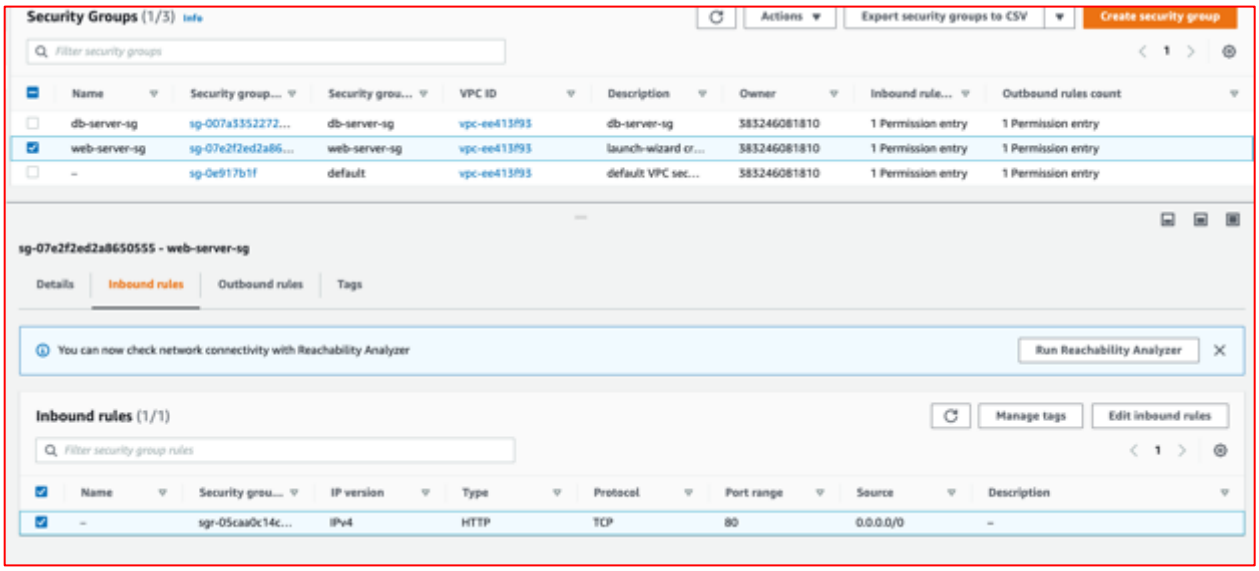
Let's try to understand AWS Security Group with the diagram shown above. Each of the three EC2 instances runs a web server, and there are two EC2 instances for databases. It's always a good practice to protect your databases. So EC2 instances running database servers are in a private subnet. And they only accept traffic from the webserver group. On the other hand, EC2 instances running web servers are in a public subnet. The reason is that web servers need to be accessed from the Internet.

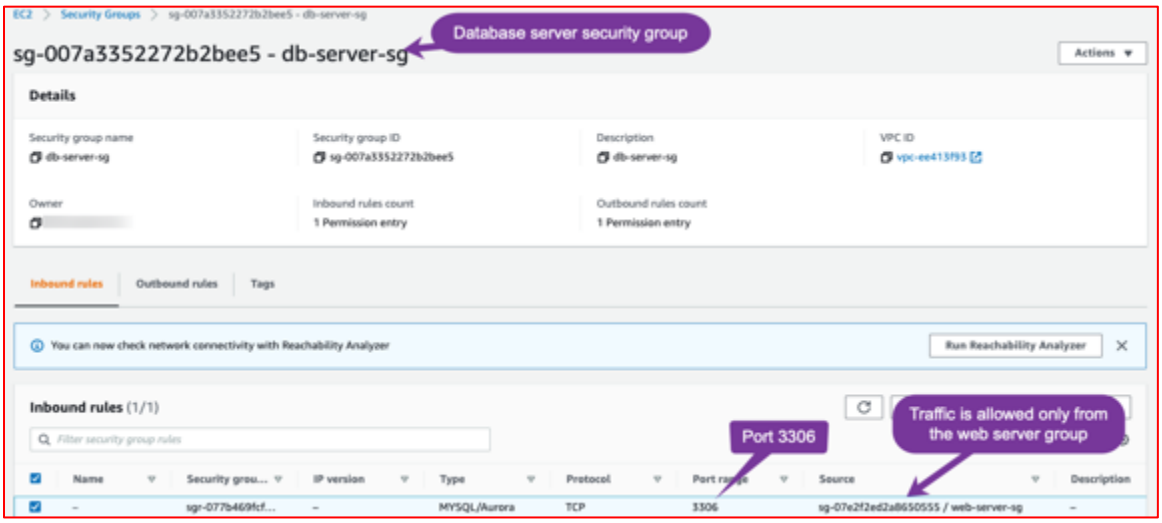
Now that was the concept of the security group. How will you create it? You can create it by clicking on the Security Group link on VPC or EC2 instance. Also, when you launch an EC2 instance, you can assign a default security group, create a new one, or choose from the existing ones.



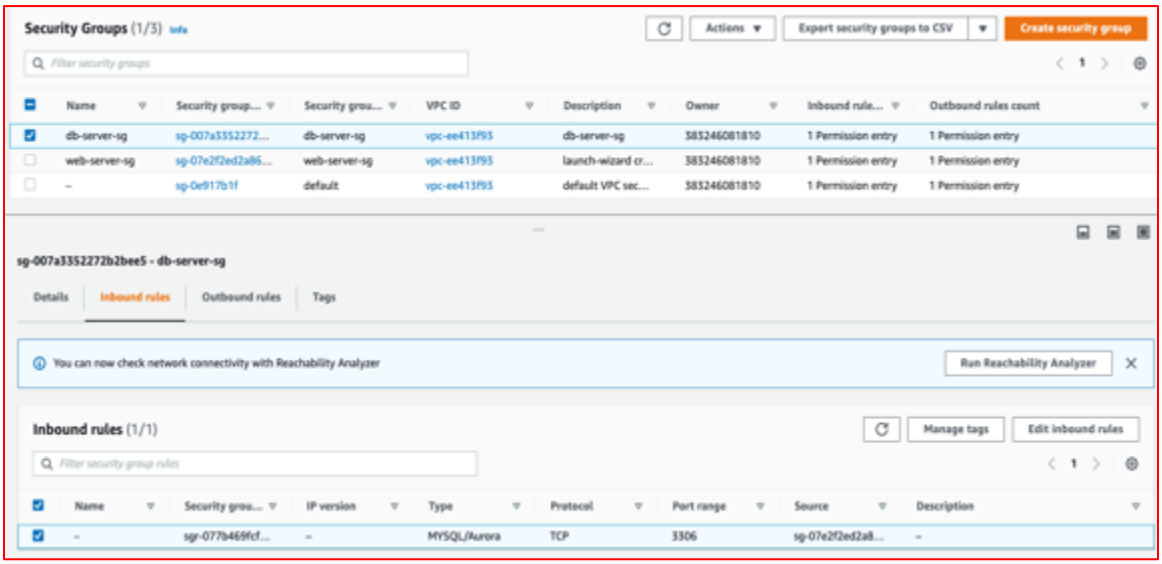
This is the screenshot of the security group of the web server group. If you notice, inbound traffic from anywhere is allowed at port 80.

The screenshot below is another view of the web server security group.





The above screenshot is for the security group of the database server group. This security group allows traffic only from the web server security group -- this is an excellent concept. First, providing a security group as a source protects the input traffic by ensuring that the input is allowed from the machines that have that security group (for example, web-server-sg in this case) traffic. And secondly, if more web servers are added to the webserver group, you will have not to change the database security group – imagine a scenario where you have to provide the IP address of each web server machine. Adding a security group as a source is a scalable solution.
The screenshot below is another view of the database server security group.

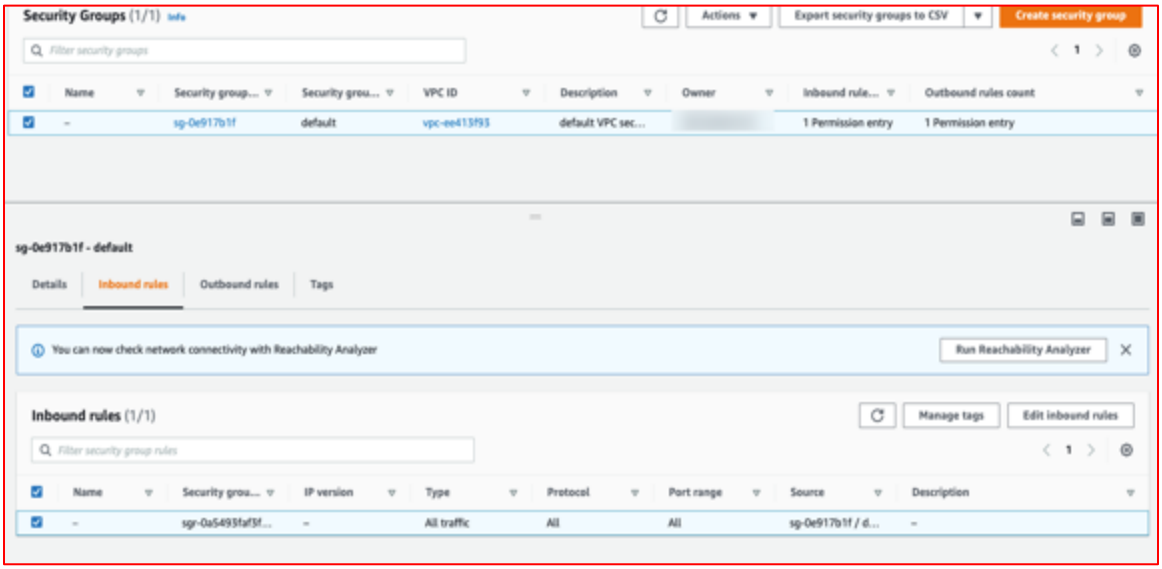


Another critical point is that though more than one security group can be assigned to one EC2 instance, at least, one security group must be assigned to an EC2 instance.

A security group can be modified, changed, and deleted.

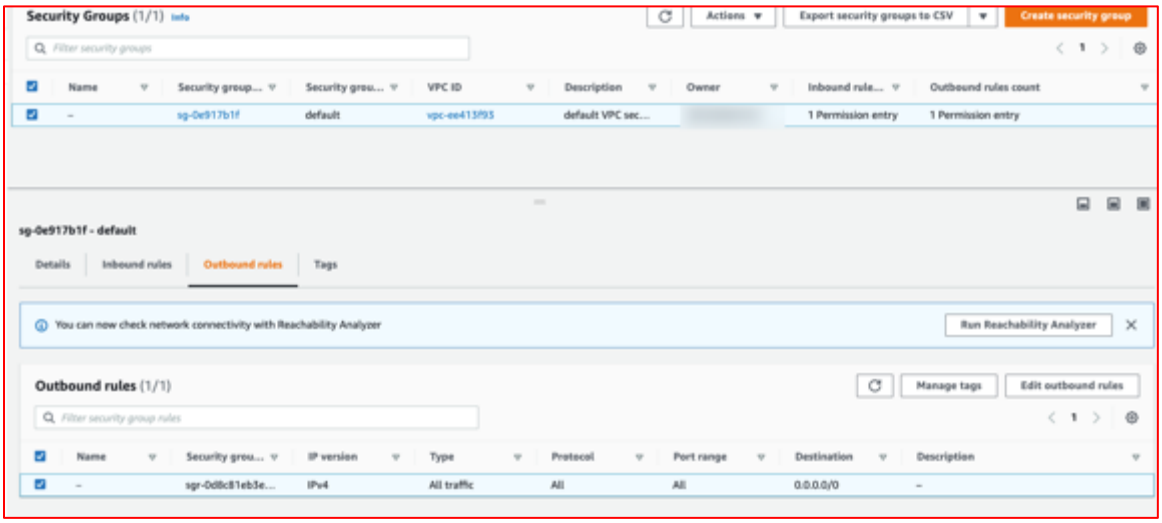
Default Inbound Rule

When you create a security group, it has no inbound rules. Therefore, no inbound traffic originating from another host to your instance is allowed until you add inbound rules to the security group.



Default Outbound Rule

A security group includes an outbound rule that allows all outbound traffic by default. It is best to remove this default rule and add outbound rules that would enable specific outbound traffic only.



Key Points About Security Group

- Security Group acts as a virtual firewall controlling inbound and outbound traffic on an EC2 instance. When you create VPC, it comes with a default security group. You can modify the default security group or create an additional security group. The default security group has

no inbound rules until you add inbound rules. You only add Allow rules – not Deny rules. For each security group, you add rules that control the traffic based on protocols and port numbers. There are a separate set of rules for inbound and outbound traffic.

- There are quotas about how many security groups can be created in a VPC, how many rules can be added to a security group, and how many security groups can be associated with a network interface. A security group can only be assigned to the resources in the security group's VPC.

- Security groups are stateful. What it means for each allowed inbound request, the response is also allowed.

- You can assign multiple security groups. Based on the aggregation of rules, it is decided whether particular traffic is allowed or not on a resource.

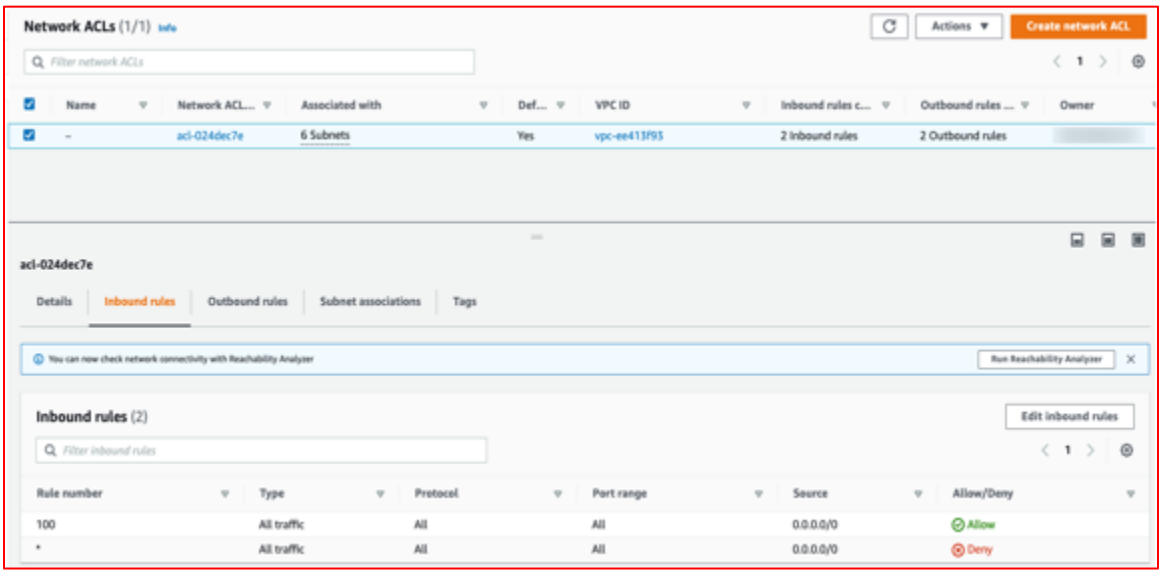
Network Access Control List (NACLs)

To understand the Network Access Control List, it would be better if we go through the differences between security group and Network ACL.

Security Group	Network ACL
Security groups operate at *instance level*	Network ACLs operate at *subnet level*
Security groups support *allow rules* only	Network ACLs support *allow and deny rules*
Security groups are *stateful* – the return traffic is automatically allowed regardless of any rules	Network ACLs are *stateless* – the return traffic must be allowed by rules.
All rules are evaluated before deciding if the traffic is allowed	Rules are *evaluated in order (low to high)* in deciding whether to allow traffic
Applies only to *instances explicitly associated* with the security group	Automatically applies *all instances* launched into associated subnet

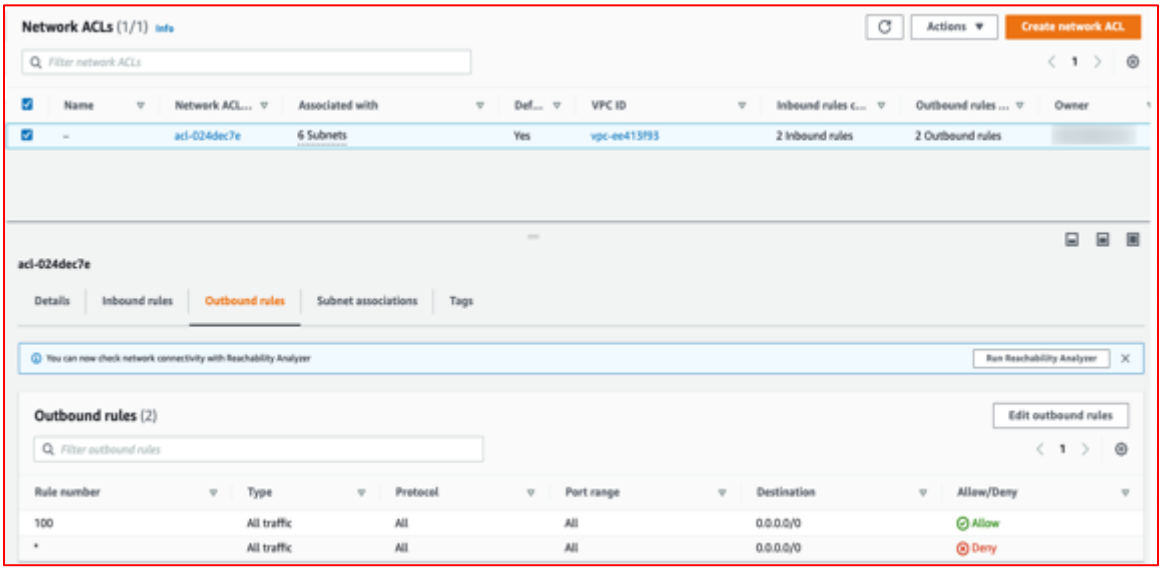
Network Access Control List are coarse-grained controls, and they should be allowed to work at the edges of the network. However, suppose you have too many complex sets of rules configured in NACLs. It could be highly likely that you are using NACLs as a security group. If this is the case, it is better to review them to ensure that you are not configuring security groups in the NACLs. To understand the Network Access Control List (NACLs), it would be better if we go through the differences between a security group and NACLs.

NACLs should be coarse grained.



The above is the Screenshot showing default Network ACLs inbound rules. You can notice that the NACLs are associated with six subnets. There are six subnets because my default Region is N. Virginia which has 6 AZs. Another important to notice is that inbound traffic is allowed from anywhere.

Below is the Screenshot showing NACLs outbound rules -- this is the default setup. In the Screenshot, the outbound traffic is allowed from anywhere.



Key Points of NACLs

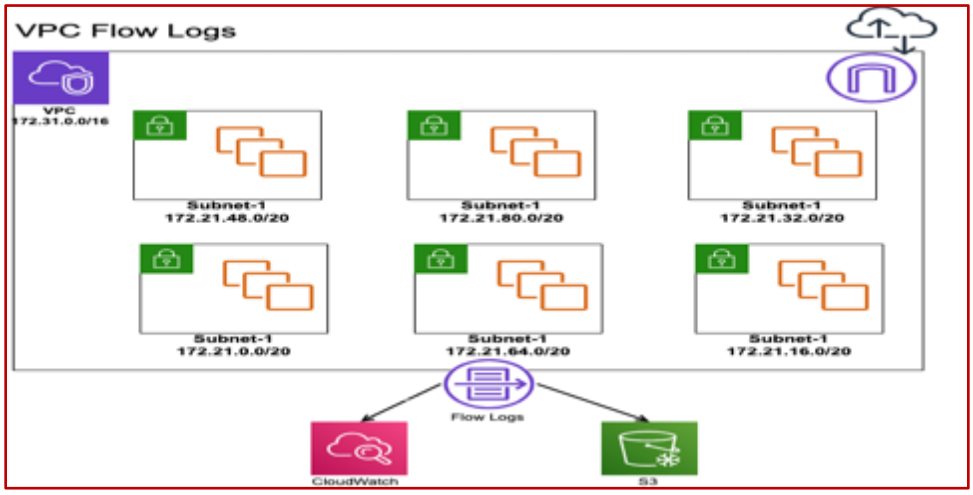
- Network ACLs provide an optional layer of security for your VPC. It acts as a firewall controlling inbound and outbound traffic for one or more subnets. By default, your VPC has a default Network ACLs which allows all inbound and outbound traffic. However, this Network ACL can be modified.

- You can create a custom Network ACL and assign it to a subnet. By default, each custom Network denies all inbound and outbound rules until you add rules.
- Each subnet in a VPC must be assigned to a Network ACL. If the subnet is not assigned to a Network ACL, the subnet is automatically assigned to a default Network ACL.
- Network ACLs have a separate inbound and outbound rules -- each rule can deny or allow traffic.
- Network ACLs are stateless, which means the response to inbound traffic is only allowed if outbound traffic is permitted and vice-versa
- Network ACL contains numbered rules highest number can be 32766. The order is evaluated for the lowest numbers; as soon as the lowest number rule matches, it is applied, and higher number rules are ignored.
- Network ACLs are different from security groups – security groups are applied at the instance level, while Network ACLs are used at the subnet level.

Flow Logs

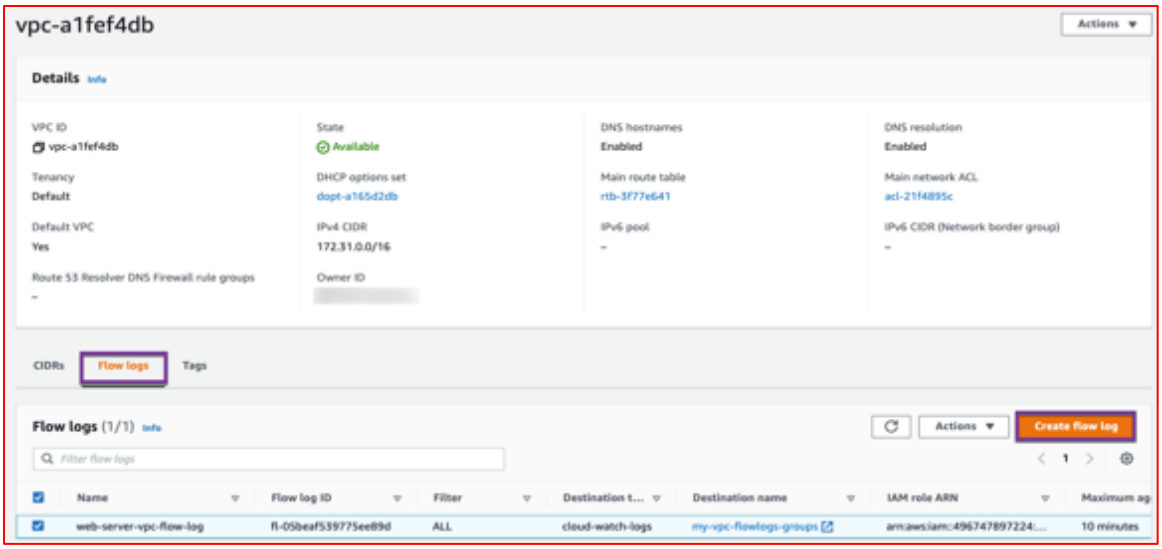
In the previous sections, we discussed ensuring that correct traffic is allowed to the subnet (via NACLs) and instances (via Security Groups).

But the question is how to look into traffic: the question is Flow Logs. We can create Flow Logs at the VPC and the Subnet levels. When we create Flow Logs at a VPC, it applies to all subnets in the VPC. However, when we create Flow Logs at the subnet, it applies only to the associated subnet.



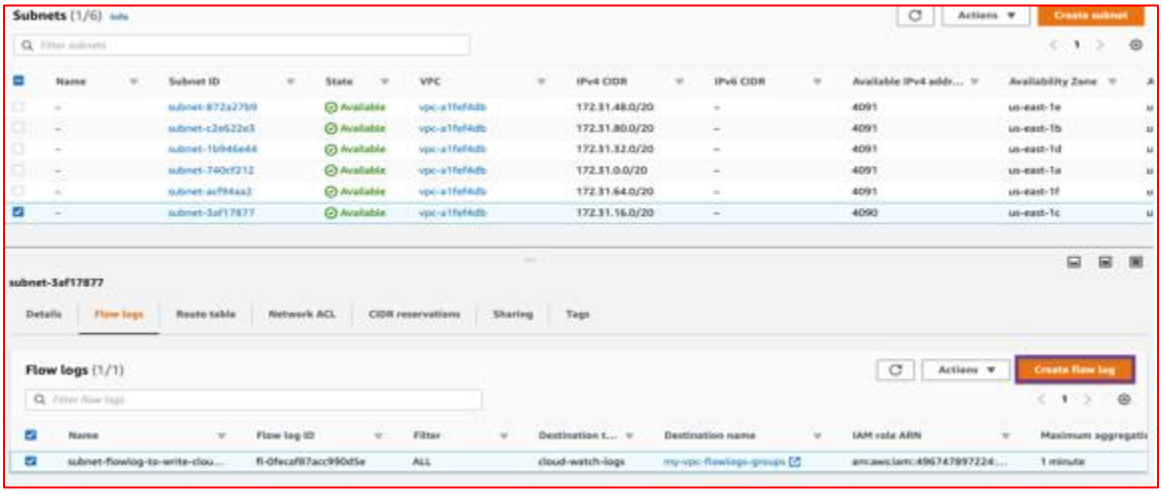
The VPC Flow Logs can be written either to a CloudWatch group or an S3 bucket. It provides visibility about what's going on in your VPC, such as troubleshooting if wrong rules are set up or analyzing traffic flows. One important point is that Flow Logs do not contain the payload of a request and response. Instead, the Flow Logs only include a packet description, such as a source and destination address, port, payload size, and whether the request is denied or accepted.

How to Create VPC Flow Logs



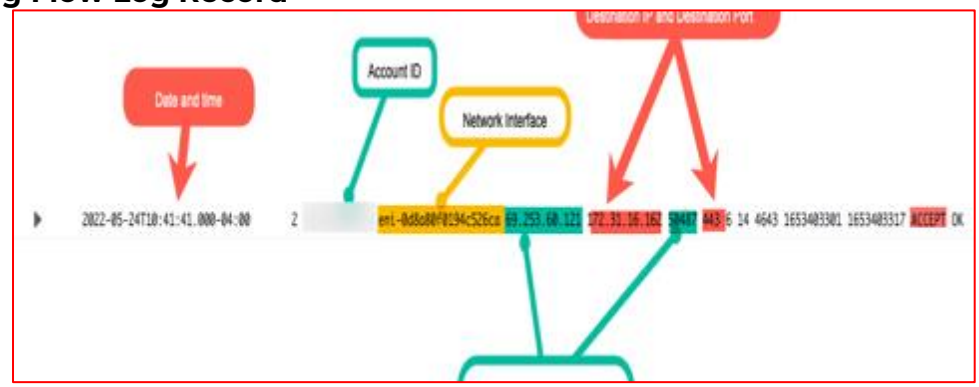
Go to VPC, click on the Flow Logs tab and then click on Create flow Log button. You can notice in the screenshot above that there is the Flow Log associated with the VPC, and the destination is CloudWatch logs.

How to Create Subnet Flow Logs



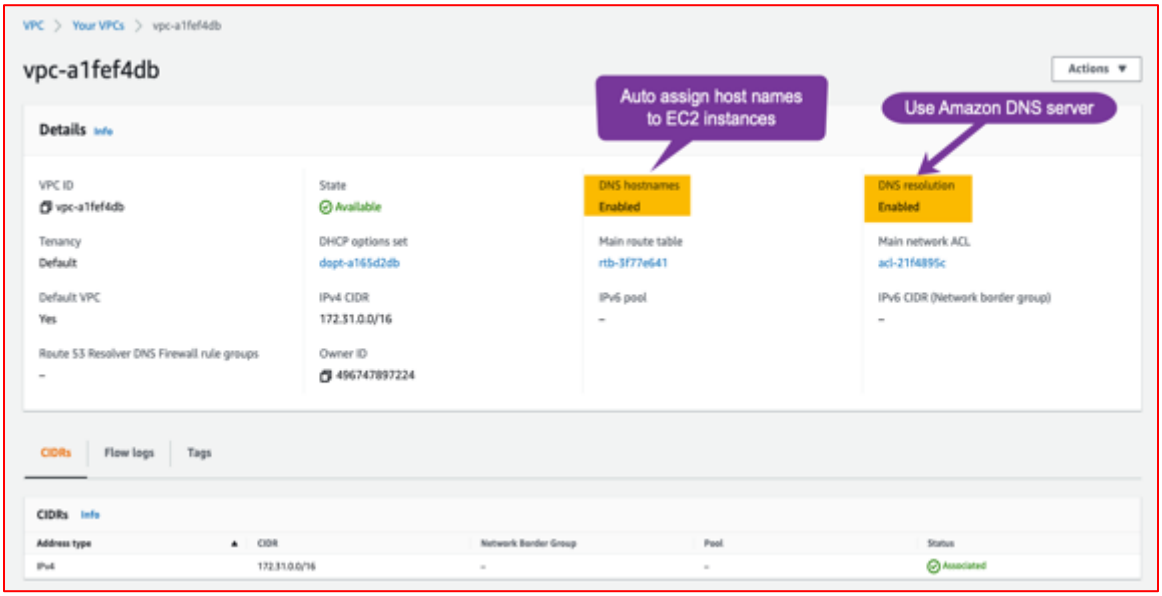
You can also create flow logs for subnet as well. Go to VPC, select subnets. Select the subnet to which you would like to add a flow log. Then click on Flow Logs tab and click on Create flow log button. You can notice in the screenshot above, there is flow log associated with the subnet subnet-3af17877 and the destination is CloudWatch logs.

Analyzing Flow Log Record



The above screenshot is for one sample Flow Log record. You can notice different parts of a request in the Flow Log.

DNS in VPC



As you can see in the screenshot, AWS provides options for DNS hostnames and DNS hostname resolutions – by default, these options are enabled. However, you can disable them if you would like for your use case.

DNS hostname resolutions help resolve the public hostname of your EC2 instance. DNS hostnames option adds the ability to add a hostname to an EC2 instance. This helps to avoid using IP addresses.

For details about managing DNS for your servers, please look into the Route53 service.