Sayan Roy

Project NIMBUS

A **MOORC** Based on Cloud Computing

2023

Content

Azure Fundamentals

First thing first, let's start with explaining Cloud Computing. Cloud computing is the delivery of computing services over the internet. Computing services include common IT infrastructure such as virtual machines, storage, databases, and networking. Cloud services also expand the traditional IT offerings to include things like the Internet of Things (IoT), machine learning (ML), and artificial intelligence (AI). Computer Resources Hosted on the internet can be defined as a cloud.

SHARED RESPONSIBILITY MODEL

Shared responsibility model is a term I came across quite a few times, across all the vendors.
With the shared responsibility model, the responsibilities get shared between the cloud provider and the consumer. Physical security, power, cooling, and network connectivity are the responsibility of the cloud provider. whereas things like deployment, and security(to some extent), etc are the responsibilities of the Client

Major Benefits of the cloud are
1.**No Hassle of buying equipment**
vendors are responsible for providing the resources, which makes this model scalable and cheaper.
2. **On-Demand Self-Service**
It is one of the important and valuable features of Cloud Computing as the user can continuously monitor the server uptime, capabilities, and allotted network storage. With this feature, the user can also monitor the computing capabilities.
3. **Easy Maintenance**
The servers are easily maintained and the downtime is very low and even in some cases, there is no downtime. Cloud Computing comes up with an update every time by gradually making it better.
4. **Security**
Cloud Security, is one of the best features of cloud computing. It creates a snapshot of the data stored so that the data may not get lost even if one of the servers gets damaged.

Cloud MODELS

Based on **Deployment**

1. On-Premise

On-premises software is installed and runs on computers on the premises of the person or organization using the software, rather than at a remote facility such as a server farm or cloud

2. IaaS

Infrastructure-as-a-service (IaaS), also known as cloud infrastructure services, is a form of cloud computing in which IT infrastructure is provided to end users through the internet. IaaS is commonly associated with serverless computing.
Some common scenarios where IaaS might make sense include
Lift-and-shift migration: You're standing up cloud resources similar to your on-prem data centre, and then simply moving the things running on-prem to running on the IaaS infrastructure.Testing and development: You have established configurations for development and test environments that you need to rapidly replicate. You can stand up or shut down the different environments rapidly with an IaaS structure while maintaining complete control.

3. PaaS

PaaS, or platform as a service, is on-demand x access to a complete, ready-to-use, cloud-hosted platform for developing, running, maintaining and managing applications. SaaS, or software as a service, is on-demand access to ready-to-use, cloud-hosted application software..
Some common scenarios where PaaS might make sense include:
Development framework: PaaS provides a framework that developers can build upon to develop or customize cloud-based applications. Similar to the way you create an Excel macro, PaaS lets developers create applications using built-in software components. Cloud features such as scalability, high availability, and multi-tenant capability are included, reducing the amount of coding that developers must do.

4. SaaS

Software as a service (or SaaS) is a way of delivering applications over the Internet—as a service. Instead of installing and maintaining software, you simply access it via the Internet, freeing yourself from complex software and hardware management.
Some common scenarios for SaaS are:
Email and messaging.
Business productivity applications.
Finance and expense tracking

.

Based on **Ownership**

1. Private Cloud

A Private Cloud is a model of cloud computing where the infrastructure is dedicated to a single-user organization.

2. Public Cloud

A public cloud is an IT model where public cloud service providers make computing services—including compute and storage, develop-and-deploy environments, and applications—available on-demand to organizations and individuals over the public internet.

3. Hybrid Cloud

A hybrid cloud is a mixed computing environment where applications are run using a combination of computing, storage, and services in different environments—public clouds and private clouds, including on-premises data centres or “edge” locations.

4. Multi-cloud

A fourth, and increasingly likely scenario is a multi-cloud scenario. In a multi-cloud scenario, you use multiple public cloud providers. Maybe you use different features from different cloud providers. Or maybe you started your cloud journey with one provider and are in the process of migrating to a different provider. Regardless, in a multi-cloud environment, you deal with two (or more) public cloud providers and manage resources and security in both environments.

Azure Cloud Features

1. Azure Arc

Azure Arc is a set of technologies that helps manage your cloud environment. Azure Arc can help manage your cloud environment, whether it's a public cloud solely on Azure, a private cloud in your data centre, a hybrid configuration, or even a multi-cloud environment running on multiple cloud providers at once.

Azure Arc provides a centralized, unified way to

- Manage your entire environment together by projecting your existing non-Azure and/or on-premises resources into Azure Resource Manager.
- Manage virtual machines, Kubernetes clusters, and databases as if they are running in Azure.
- Use familiar Azure services and management capabilities, regardless of where they live.
- Continue using traditional ITOps while introducing DevOps practices to support new cloud-native patterns in.your environment.
- Configure custom locations as an abstraction layer on top of Azure Arc-enabled Kubernetes clusters and cluster extensions

.

2. Azure VMware

What if you're already established with VMware in a private cloud environment but want to migrate to a public or hybrid cloud? Azure VMware Solution lets you run your VMware workloads in Azure with seamless integration and scalability.

availability, responsibilities - are agreed upon between the service provider and the service user.

.

CapEx & OpEx

A **capital expenditure, or Capex**, is money invested by a company to acquire or upgrade fixed, physical or nonconsumable assets. Capex is primarily a one-time investment in nonconsumable assets used to maintain existing levels of operation within a company and to foster its future growth.

In contrast, **OpEx** is spending money on services or products over time. Renting a convention centre, leasing a company vehicle, or signing up for cloud services are all examples of OpEx.

Cloud uses the **OpEx model,** you pay when you need
This consumption-based model has many benefits, including

- No upfront costs.
- No need to purchase and manage the costly infrastructure that users might not use to its fullest potential.
- The ability to pay for more resources when they're needed.
- The ability to stop paying for resources that are no longer needed.

Service Level Agreement (SLAS)

A service-level agreement is a commitment between a service provider and a customer. Particular aspects of the service - quality, availability, responsibilities - are agreed upon between the service provider and the service user

SLAs. 99% will have 432 min unavailable per month
but 99.9% will be unavailable for 43.2min per month. so choose wisely!

Scaling

can be divided into two types

- **Vertical scaling**
  With vertical scaling, if you were developing an app and you needed more processing power, you could vertically scale up to add more CPUs or RAM to the virtual machine. Conversely, if you realized you had over-specified the needs, you could vertically scale down by lowering the CPU or RAM specifications.

- **Horizontal scaling**
  With horizontal scaling, if you suddenly experienced a steep jump in demand, your deployed resources could be scaled out (either automatically or manually). For example, you could add additional virtual machines or containers, scaling out. In the same manner, if there was a significant drop in demand, deployed resources could be scaled in (either automatically or manually), scaling in.

An availability set is a logical grouping of VMs that allows Azure to understand how your application is built to provide redundancy and availability. We recommended that two or more VMs are created within an availability set to provide for a highly available application and to meet the 99.95% Azure SLA.

.Availability sets do this by grouping VMs in two ways: **Update domain** and **fault domain**

Update domain: The update domain groups VMs that can be rebooted at the same time. This allows you to apply updates while knowing that only one update domain grouping will be offline at a time. All of the machines in one updated domain will be updated. An update group going through the update process is given a 30-minute time to recover before maintenance on the next update domain starts.

Fault domain: The fault domain groups your VMs by a common power source and network switch. By default, an availability set will split your VMs across up to three fault domains. This helps protect against a physical power or networking failure by having VMs in different fault domains (thus being connected to different power and networking resources).

.

3. Azure Virtual Desktop

Another type of virtual machine is the Azure Virtual Desktop. **Azure Virtual Desktop** is a desktop and application virtualization service that runs on the cloud. It enables you to use a cloud-hosted version of Windows from any location. Azure Virtual Desktop works across devices and operating systems and works with apps that you can use to access remote desktops or most modern browsers.

4.Containers

While virtual machines are an excellent way to reduce costs versus the investments that are necessary for physical hardware, they're still limited to a single operating system per virtual machine. If you want to run multiple instances of an application on a single host machine, containers are an excellent choice

Containers are a virtualization environment. Much like running multiple virtual machines on a single physical host, you can run multiple containers on a single physical or virtual host. Unlike virtual machines, you don't manage the operating system for a container. Virtual machines appear to be an instance of an operating system that you can connect to and manage. Containers are lightweight and designed to be created, scaled out, and stopped dynamically. It's possible to create and deploy virtual machines as application demand increases, but containers are a lighter-weight, more agile method. Containers are designed to allow you to respond to changes on demand. With containers, you can quickly restart if there's a crash or hardware interruption. One of the most popular container engines is Docker, which is supported by Azure.

Popular Container Engine **DOCKER**

Container Virtualizes VM, much faster but less controlled

Containers are often used to create solutions by using a microservice architecture. This architecture is where you break solutions into smaller, independent pieces. For example, you might split a website into a container hosting your front end, another hosting your back end, and a third for storage. This split allows you to separate portions of your app into logical sections that can be maintained, scaled, or updated independently.

Imagine your website's back end has reached capacity but the front end and storage aren't being stressed. With containers, you could scale the back end separately to improve performance. If something necessitated such a change, you could also choose to change the storage service or modify the front end without impacting any of the other components.

.

Serverless computing

Serverless technologies feature automatic scaling, built-in high availability, and a pay-for-use billing model to increase agility and optimize costs. These technologies also eliminate infrastructure management tasks like capacity provisioning and patching, so you can focus on writing code that serves your customers.

5. Azure Functions

Azure Functions is a serverless solution that allows you to write less code, maintain less infrastructure, and save on costs. Instead of worrying about deploying and maintaining servers, the cloud infrastructure provides all the up-to-date resources needed to keep your applications running. Amazon's alternative is **lambda** and Google's alternative is Google **Cloud Function**

6. Azure App Service

App Service enables you to build and host web apps, background jobs, mobile back-ends, and RESTful APIs in the programming language of your choice without managing infrastructure. It offers automatic scaling and high availability. App Service supports Windows and Linux. It enables automated deployments from GitHub, Azure DevOps, or any Git repo to support a continuous deployment model.

7. Azure virtual Networks

Azure virtual networks and virtual subnets enable Azure resources, such as VMs, web apps, and databases, to communicate with each other, with users on the internet, and with your on-premises client computers. You can think of an Azure network as an extension of your on-premises network with resources that link other Azure resources.

Azure virtual networks provide the following key networking capabilities:

- Isolation and segmentation
- Internet communications
- Communicate between Azure resources
- Communicate with on-premises resources
- Route network traffic
- Filter network traffic
- Connect virtual networks

.

Isolation and segmentation
Azure virtual network allows you to create multiple isolated virtual networks. When you set up a virtual network, you define a private IP address space by using either public or private IP address ranges. The IP range only exists within the virtual network and isn't internet routable. You can divide that IP address space into subnets and allocate part of the defined address space to each named subnet

.

VPN Gateway

A VPN gateway is a type of virtual network gateway. Azure VPN Gateway instances are deployed in a dedicated subnet of the virtual network and enable the following connectivity

Connect on-premises data centres to virtual networks through a site-to-site connection.
Connect individual devices to virtual networks through a point-to-site connection.
Connect virtual networks to other virtual networks through a network-to-network connection.

8. stock keeping units (SKUs)

In regions that support availability zones, VPN gateways and ExpressRoute gateways can be deployed in a zone-redundant configuration. This configuration brings resiliency, scalability, and higher availability to virtual network gateways. Deploying gateways in Azure availability zones physically and logically separates gateways within a region while protecting your on-premises network connectivity to Azure from zone-level failures. These gateways require different gateway stock-keeping units (SKUs) and use Standard public IP addresses instead of Basic public IP addresses.

9. Azure ExpressRoute

Azure ExpressRoute lets you extend your on-premises networks into the Microsoft cloud over a private connection, with the help of a connectivity provider. This connection is called an ExpressRoute Circuit. With ExpressRoute, you can establish connections to Microsoft cloud services, such as Microsoft Azure and Microsoft 365. This allows you to connect offices, data centres, or other facilities to the Microsoft cloud. Each location would have its own ExpressRoute circuit.
Express route data is not encrypted but provides a direct connection to MSC's services. whereas VPN provides secure encrypted access to resources all over the internet

10. Azure DNS

Azure DNS is a hosting service for DNS domains that provides name resolution by using Microsoft Azure infrastructure. By hosting your domains in Azure, you can manage your DNS records using the same credentials, APIs, tools, and billing as your other Azure services.

DNS for websitesAzure DNS is based on Azure Resource Manager, which provides features such as

- Azure role-based access control (Azure RBAC) to control who has access to specific actions for your organization.
- Activity logs to monitor how a user in your organization modified a resource or to find an error when troubleshooting.
- Resource locking to lock a subscription, resource group, or resource. Locking prevents other users in your organization from accidentally deleting or modifying critical resources.

You can't use Azure DNS to buy a domain name. For an annual fee, you can buy a domain name by using App Service domains or a third-party domain name registrar. Once purchased, your domains can be hosted in Azure DNS for record management.

11. Azure Storage Tiers

Blob Storage (including Data Lake Storage), Queue Storage, Table Storage, and Azure Files Standard storage account type for blobs, file shares, queues, and tables.

Recommended for most scenarios using Azure Storage. If you want support for the network file system (NFS) in Azure Files, use the premium file shares account type.

Blob Storage (including Data Lake Storage)
Premium storage account type for block blobs and append blobs. Recommended for scenarios with high transaction rates or that use smaller objects or require consistently low storage latency.

Azure Files
Premium storage account type for file shares only. Recommended for enterprise or high-performance scale applications. Use this account type if you want a storage account that supports both Server Message Block (SMB) and NFS file shares

Page blobs only
Premium storage account type for page blobs only.
Storage service
Endpoint

- Blob Storage
  https://<storage-account-name>.blob.core.windows.net
- Data Lake Storage Gen2

  https://<storage-account-name>.dfs.core.windows.net
- Azure Files
  https://<storage-account-name>.file.core.windows.net
- Queue Storage

  https://<storage-account-name>.queue.core.windows.net
- Table Storage
  https://<storage-account-name>.table.core.windows.net

Data in an Azure Storage account is always replicated three times in the primary region. Azure Storage offers two options

zone-redundant storage (ZRS).

Locally redundant storage
(LRS) replicates your data three times within a single data centre in the primary region. LRS provides at least 11 nines of durability (99.999999999%) of objects over a given year.

For Availability Zone-enabled Regions, **zone-redundant storage** (ZRS) replicates your Azure Storage data synchronously across three Azure availability zones in the primary region. ZRS offers durability for Azure Storage data objects of at least 12 nines (99.9999999999%) over a given year.

MR robot uses GEo redundant Storage
Azure Storage offers two options for copying your data to a secondary region: geo-redundant storage (GRS) and geo-zone-redundant storage (GZRS). GRS is similar to running LRS in two regions, and GZRS is similar to running ZRS in the primary region and LRS in the secondary region.

Because data is replicated to the secondary region asynchronously, a failure that affects the primary region may result in data loss if the primary region can't be recovered. The interval between the most recent writes to the primary region and the last write to the secondary region is known as the recovery point objective (RPO). The RPO indicates the point in time to which data can be recovered. Azure Storage typically has an RPO of less than 15 minutes, although there's currently no SLA on how long it takes to replicate data to the secondary region.

- Locally redundant Storage
- Zone Redundant Storage
- Geo-redundant Storage

The Azure Storage platform includes the following data services

for how your data is replicated in the primary region, locally redundant storage (LRS) and
Azure Blobs: A massively scalable object store for text and binary data. Also includes support for big data analytics through Data Lake Storage Gen2.
Azure Files: Managed file shares for cloud or on-premises deployments.
Azure Queues: A messaging store for reliable messaging between application components.
Azure Disks: Block-level storage volumes for Azure VMs

Hot access tier: Optimized for storing data that is accessed frequently (for example, images for your website).
Cool access tier: Optimized for data that is infrequently accessed and stored for at least 30 days (for example, invoices for your customers).
Archive access tier: Appropriate for data that is rarely accessed and stored for at least 180 days, with flexible latency requirements (for example, long-term backups).

12. Azure Migrate

Azure Migrate is a service that helps you migrate from an on-premises environment to the cloud. Azure Migrate functions as a hub to help you manage the assessment and migration of your on-premises datacenter to Azure. It provides the following:
Unified migration platform: A single portal to start, run, and track your migration to Azure.
Range of tools: A range of tools for assessment and migration. Azure Migrate tools include Azure Migrate: Discovery and assessment and Azure Migrate: Server Migration. Azure Migrate also integrates

with independent software vendor (ISV) offerings.
Assessment and migration: In the Azure Migrate hub, you can assess and migrate your on-premises infrastructure to Azure.

12. Az Copy

AzCopy is a command-line utility that you can use to copy blobs or files to or from your storage account. With AzCopy, you can upload files, download files, copy files between storage accounts, and even synchronize files. AzCopy can even be configured to work with other cloud providers to help move files back and forth between clouds.

13. Azure Active Directory

(Azure AD) is a directory service that enables you to sign in and access both Microsoft cloud applications and the cloud applications that you develop. Azure AD can also help you maintain your on-premises Active Directory deployment.

14. Security

Single sign-on (SSO) enables a user to sign in one time and use that credential to access multiple resources and applications from different providers. For SSO to work, the different applications and providers must trust the initial authenticator.

The objective of defence-in-depth is to protect information and prevent it from being stolen with other Azure services and tools, and Role-based access control is applied to a scope, which is a resource or set of resources that this access applies to.

The following diagram shows the relationship between roles and scopes. A management group, subscription, or resource admin might be given the role of owner, so they have increased control and authority. An observer, who isn't expected to make any updates, might be given a role of Reader for the same scope, enabling them to review or observe the management group, subscription, or resource group.

Zero Trust is a security model that assumes the worst case scenario and protects resources with that expectation. Zero Trust assumes breach at the outset and then verifies each request as though it originated from an uncontrolled network.

Today, organizations need a new security model that effectively adapts to the complexity of the modern environment; embraces the mobile workforce: and protects people, devices, applications, and data wherever they're located.

To address this new world of computing, Microsoft highly recommends the Zero Trust security model, which is based on these guiding principles:

Verify explicitly - Always authenticate and authorize based on all available data points.
Use least privilege access - Limit user access with Just-In-Time and Just-Enough-Access (JIT/JEA), risk-based adaptive policies, and data protection.
Assume breach - Minimize blast radius and segment access. Verify end-to-end encryption. Use analytics to get visibility, drive threat detection, and improve

by those who aren't authorized to access it.

A defence-in-depth strategy uses a series of mechanisms to slow the advance of an attack that aims at acquiring unauthorized access to data

14. Pricing Calculator

The pricing calculator and the total cost of ownership (TCO) calculator are two calculators that help you understand potential Azure expenses. Both calculators are accessible from the internet, and both calculators allow you to build out a configuration. However, the two calculators have very different purposes.

15. Pricing Blue Prints

Azure Blueprints lets you standardize cloud subscriptions or environment deployments. Instead of having to configure features like Azure Policy for each new subscription, with Azure Blueprints, you can define repeatable settings and policies that are applied as new subscriptions are created. Need a new test/dev environment? Azure Blueprints lets you deploy a new Test/Dev environment with security and compliance settings already configured. In this way, development teams can rapidly build and deploy new environments with the knowledge that they're building within organizational requirements

defences.

15. Artefacts

Each component in the blueprint definition is known as an artifact.

It is possible for artifacts to have no additional parameters (configurations). An example is the Deploy threat detection on SQL servers policy, which requires no additional configuration.

Artifacts can also contain one or more parameters that you can configure. The following screenshot shows the Allowed locations policy. This policy includes a parameter that specifies the allowed locations.

A resource lock prevents resources from being accidentally deleted or changed.

Even with Azure role-based access control (Azure RBAC) policies in place, there's still a risk that people with the right level of access could delete critical cloud resources. Resource locks prevent resources from being deleted or updated, depending on the type of lock. Resource locks can be applied to individual resources, resource groups, or even an entire subscription. Resource locks are inherited, meaning that if you place a resource lock on a resource group, all of the resources within the resource group will also have the resource lock applied.

Some Other Terms:

Azure Advisor evaluates your Azure resources and makes recommendations to help improve reliability, security, and performance, achieve operational excellence, and reduce costs. Azure Advisor is designed to help you save time on cloud optimization. The recommendation service includes suggested actions you can take right away, postpone, or dismiss.

Azure Resource Manager (ARM) is the deployment and management service for Azure. It provides a management layer that enables you to create, update, and delete resources in your Azure account. Anytime you do anything with your Azure resources, ARM is involved.

To get the most out of Azure, you need a way to interact with the Azure environment, the management groups, subscriptions, resource groups, resources, and so on. Azure provides multiple tools for managing your environment, including the:

Azure portal
Azure PowerShell
Azure Command Line Interface (CLI)

The Microsoft Service Trust Portal is a portal that provides access to various content, tools, and other resources about Microsoft security, privacy, and compliance practices.

The Service Trust Portal contains details about Microsoft's implementation of controls and processes that protect our cloud services and the customer data therein. To access some of the resources on the Service Trust Portal, you must sign in as an authenticated user with your Microsoft cloud services account (Azure Active Directory organization account). You'll need to review and accept the Microsoft non-disclosure agreement for compliance materials.

AWS Skill Builder Essentials

Amazon Elastic Cloud EC2

Amazon Elastic Compute Cloud is a part of Amazon.com's cloud-computing platform, Amazon Web Services, that allows users to rent virtual computers on which to run their own computer applications

Multitenancy

Multitenancy is a reference to the mode of operation of software where multiple independent instances of one or multiple applications operate in a shared environment. The instances (tenants) are logically isolated, but physically integrated.
Sharing Hardware between multiple VMs with the help of a hypervisor

Compute as a Service Module(CaaS)

CaaS (Containers as a Service) is a pay-as-you-go cloud-based service offering organizations a way to manage their virtualized applications, clusters, and containers to make deployments faster and easier.

Instance Types

- **General Purpose** (Balanced)
- **Compute Optimized** Resource(Computing tasks like gaming)
- **Memory Optimized** for memory intensive tasks
- **Accelerated Computing** (Uses hardware accelerators) - good for graphics processing, data pattern matching etc etc

Compute-optimized instances are ideal for compute-bound applications that benefit from high-performance processors. Like general-purpose instances, you can use compute-optimized instances for workloads such as web, application, and gaming servers.

However, the difference is compute optimized applications are ideal for high-performance web servers, compute-intensive applications servers, and dedicated gaming servers. You can also use compute-optimized instances for batch-processing workloads that require processing many transactions in a single group.

Memory-optimized instances are designed to deliver fast performance for workloads that process large datasets in memory. In computing, memory is a temporary storage area. It holds all the data and instructions that a central processing unit (CPU) needs to be able to complete actions. Before a computer program or application is able to run, it is loaded from storage into memory. This preloading process gives the CPU direct access to the computer program.

Suppose that you have a workload that requires large amounts of data to be preloaded before running an application. This scenario might be a high-performance database or a workload that involves performing real-time processing of a large amount of unstructured data. In these types of use cases, consider using a memory-optimized instance. Memory-optimized instances enable you to run workloads with high memory needs and receive great performance.

Accelerated computing instances use hardware accelerators, or coprocessors, to perform some functions more efficiently than is possible in software running on CPUs.

Examples of these functions include floating-point number calculations, graphics processing, and data pattern matching.

In computing, a **hardware accelerator** is a component that can expedite data processing. Accelerated computing instances are ideal for workloads such as graphics applications, game streaming, and application streaming.

Storage-optimized instances are designed for workloads that require high, sequential read and write access to large datasets on local storage. Examples of workloads suitable for storage-optimized instances include distributed file systems, data warehousing applications, and high-frequency online transaction processing (OLTP) systems.

In computing, the term input/output operations per second (IOPS) is a metric that measures the performance of a storage device. It indicates how many different input or output operations a device can perform in one second. Storage-optimized instances are designed to deliver tens of thousands of low-latency, random IOPS to applications.

You can think of input operations as data put into a system, such as records entered into a database. An output operation is data generated by a server. An example of output might be the analytics performed on the records in a database. If you have an application that has a high IOPS requirement, a storage-optimized instance can provide better performance over other instance types not optimized for this kind of use case.

Pricing

Based on Pricing models are of the type:

On-Demand Instances are ideal for short-term, irregular workloads that cannot be interrupted. No upfront costs or minimum contracts apply. The instances run continuously until you stop them, and you pay for only the computing time you use. (Best when you start)

AWS offers **Savings Plans** for several computing services, including Amazon EC2. Amazon EC2 Savings Plans enable you to reduce you compute costs by committing to a consistent amount of compute usage for a 1-year or 3-year term. This term commitment results in savings of up to 72% over On-Demand costs.

Reserved Instances are a billing discount applied to the use of On-Demand Instances in your account. You can purchase Standard Reserved and Convertible Reserved Instances for a 1-year or 3-year term, and Scheduled Reserved Instances for a 1-year term. You realize greater cost savings with the 3-year option.

Dedicated Hosts are physical servers with Amazon EC2 instance capacity that is fully dedicated to your use.

You can use your existing per-socket, per-core, or per-VM software licenses to help maintain license compliance. You can purchase On-Demand Dedicated Hosts and Dedicated Hosts Reservations. Of all the Amazon EC2 options that were covered, Dedicated Hosts are the most expensive.

Scalability

Amazon EC2 Auto Scaling
Amazon EC2 Auto Scaling enables you to automatically add or remove Amazon EC2 instances in response to changing application demand.

Amazon EC2 Auto Scaling helps you maintain application availability and lets you automatically add or remove EC2 instances using scaling policies that you define. Dynamic or predictive scaling policies let you add or remove EC2 instance capacity to service established or real-time demand patterns. The fleet management features of Amazon EC2 Auto Scaling helps maintain the health and availability of your fleet.

Elastic Load Balancing

Load balancing
Divide requests between the front end and back end

Elastic Load Balancing is the AWS service that automatically distributes incoming application traffic across multiple resources, such as Amazon EC2 instances.

A load balancer acts as a single point of contact for all incoming web traffic to your Auto Scaling group. This means that as you add or remove Amazon EC2 instances in response to the amount of incoming traffic, these requests route to the load balancer first. Then, the requests spread across multiple resources that will handle them.

For example, if you have multiple Amazon

EC2 instances, Elastic Load Balancing distributes the workload across the multiple instances so that no single instance has to carry the bulk of it

Another way to achieve decoupled system is by making use of Message and Query boards

.

Messaging and queuing

Suppose that you have an application with **tightly coupled components**. These components might include databases, servers, the user interface, business logic, and so on. This type of architecture can be considered a monolithic application.

In a **microservices approach**, application components are loosely coupled. In this case, if a single component fails, the other components continue to work because they are communicating with each other. The loose coupling prevents the entire application from failing.

Two services facilitate application integration: **Amazon Simple Notification Service** (Amazon SNS) and **Amazon Simple Queue Service** (Amazon SQS).

Amazon Simple Notification Service (Amazon SNS) is a **publish/subscribe** service. Using Amazon SNS topics, a publisher publishes messages to subscribers. This is similar to the coffee shop; the cashier provides coffee orders to the barista who makes the drinks.

Amazon Simple Queue Service (SQS) lets you send, store, and receive messages between software components at any volume, without losing messages or requiring other services to be available.

Compute Services

An AWS service for serverless computing is AWS Lambda.

AWS Lambda

AWS Lambda is an event-driven, serverless computing platform provided by Amazon as a part of Amazon Web Services. It is a computing service that runs code in response to events and automatically manages the computing resources required by that code. It was introduced on November 13, 2014.AWS Lambda is a service that lets you run code without needing to provision or manage servers. Suitable for task which requires a processing time of less than **15min**.

Azure has **Azure functions** and **Google** has **Google Cloud Functions**

Containers

An AWS service for serverless computing is AWS Lambda.

Containers provide you with a standard way to package your application's code and dependencies into a single object. You can also use containers for processes and workflows in which there are essential requirements for security, reliability, and scalability.

Amazon offers two kinds of container services: **Amazon Elastic Container Service** and **Amazon Elastic Kubernetes Service**.

Amazon Elastic Container Service (Amazon ECS)

Amazon Elastic Container Service (Amazon ECS) is a highly scalable, high-performance container management system that enables you to run and scale containerized applications on AWS.

Amazon Elastic Kubernetes Service (Amazon EKS)

Amazon Elastic Kubernetes Service (Amazon EKS) is a fully managed service that you can use to run Kubernetes on AWS.

You can run them either on EC2 or on a serverless application like **AWS Fargate**

AWS Fargate is a serverless computing engine for containers. It works with both Amazon ECS and Amazon EKS.

Interact with AWS

The AWS Management Console is a web-based interface for accessing and managing AWS services. You can quickly access recently used services and search for other services by name, keyword, or acronym. The console includes wizards and automated workflows that can simplify the process of completing tasks.

To save time when making API requests, you can use the **AWS Command Line Interface** (AWS CLI). AWS CLI enables you to control multiple AWS services directly from the command line within one tool. AWS CLI is available for users on Windows, macOS, and Linux.

Another option for accessing and managing AWS services is the software development kits (SDKs). SDKs make it easier for you to use AWS services through an API designed for your programming language or platform. SDKs enable you to use AWS services with your existing applications or create entirely new applications that will run on AWS.

AWS Elastic Beanstalk

AWS Elastic Beanstalk is an orchestration service offered by Amazon Web Services for deploying applications which orchestrate various AWS services

With AWS Elastic Beanstalk, you provide code and configuration settings, and Elastic Beanstalk deploys the resources necessary to perform the following tasks:

- Adjust capacity
- Load balancing
- Automatic scaling
- Application health monitoring

AWS CloudFormation

With AWS CloudFormation, you can treat your **infrastructure as code.** This means that you can build an environment by writing lines of code instead of using the AWS Management Console to individually provision resources.

Networking

Amazon Virtual Private Cloud (Amazon VPC)

Imagine the millions of customers who use AWS services. Also, imagine the millions of resources that these customers have created, such as Amazon EC2 instances. Without boundaries around all of these resources, network traffic would be able to flow between them unrestricted.

A **subnet** is a section of a VPC that can contain resources such as Amazon EC2 instances.

Virtual private gateway
To access private resources in a VPC, you can use a virtual private gateway.

AWS Direct Connect
AWS Direct Connect is a service that enables you to establish a dedicated private connection between your data centre and a VPC.

A subnet is a section of a VPC in which you can group resources based on security or operational needs. Subnets can be public or private.

Public subnets contain resources that need to be accessible by the public, such as an online store's website.

Private subnets contain resources that should be accessible only through your private network, such as a database that contains customers' personal information and order histories.

Network access control lists (ACLs)

A network access control list (ACL) is a virtual firewall that controls inbound and outbound traffic at the subnet level..

They have **Stateless packet filtering**

Network ACLs perform stateless packet filtering. They remember nothing and check packets that cross the subnet border each way: inbound and outbound.

Network access control lists (ACLs)

Security groups

A security group is a virtual firewall that controls inbound and outbound traffic for an Amazon EC2 instance.

They have **Stateful packet filtering**

Security groups perform stateful packet filtering. They remember previous decisions made for incoming packets.

Consider the same example of sending a request out from an Amazon EC2 instance to the internet.

When a packet response for that request returns to the instance, the security group remembers your previous request. The security group allows the response to proceed, regardless of inbound security group rules.

AWS Route 53

Amazon Route 53 is a highly available and scalable **Domain Name System** (DNS) web service. Route 53 connects user requests to internet applications running on AWS or on-premises.

Storage

Instance storage
Block-level storage volumes behave like physical hard drives.Best not to write here as data is lost as soon as the instance is terminated

Amazon Elastic Block Store (Amazon EBS) is a service that provides block-level storage volumes that you can use with Amazon EC2 instances. If you stop or terminate an Amazon EC2 instance, all the data on the attached EBS volume remains available.

An **EBS snapshot** is an incremental backup. This means that the first backup taken of a volume copies all the data. For subsequent backups, only the blocks of data that have changed since the most recent snapshot are saved.

Storage Spaces are called **Buckets**

Storage Tiers

Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service (Amazon S3) is a service that provides object-level storage. Amazon S3 stores data as objects in buckets.

You can upload any type of file to Amazon S3, such as images, videos, text files, and so on. For example, you might use Amazon S3 to store backup files, media files for a website, or archived documents.

Amazon S3 offers unlimited storage space.
The maximum file size for an object in Amazon S3 is 5 TB.

WORM(Write Once Read Many)
Amazon S3 Object Lock is an Amazon S3 feature that allows you to store objects using a write once, read many (WORM) model. You can use WORM protection for scenarios where it is imperative that data is not changed or deleted after it has been written.

Standard
Designed for frequently accessed data
Stores data in a minimum of three Availability Zones **Amazon S3 Standard** provides high availability for objects. This makes it a good choice for a wide range of use cases, such as websites, content distribution, and data analytics. Amazon S3 Standard has a higher cost than other storage classes intended for infrequently accessed data and archival storage.

IA(infrequent access)
is ideal for infrequently accessed data
Similar to Amazon S3 Standard but has a lower storage price and higher retrieval price
Amazon S3 Standard-IA is ideal for data infrequently accessed but requires high availability when needed. Both Amazon S3 Standard and Amazon S3 Standard-IA store data in a minimum of three Availability Zones. Amazon S3 Standard-IA provides the same level of availability as Amazon S3 Standard but with a lower storage price and a higher retrieval price.

One zone IA
Stores data in a single Availability Zone and has a lower storage price than Amazon S3 Standard-IA
Compared to Amazon S3 Standard and Amazon S3 Standard-IA, which store data in a minimum of three Availability Zones, Amazon S3 One Zone-IA stores data in a single Availability Zone. This makes it a good storage class to consider if the following conditions apply:

- You want to save costs on storage.
- You can easily reproduce your data in the event of an Availability Zone failure.

Intelligent tiering
Ideal for data with unknown or changing access patterns requires a small monthly monitoring and automation fee per object
In the Amazon S3 Intelligent-Tiering storage class, Amazon S3 monitors objects' access patterns. If you haven't accessed an object for 30 consecutive days, Amazon S3 automatically moves it to the infrequent access tier, Amazon S3 Standard-IA. **If you access an object in the infrequent access tier, Amazon S3 automatically moves it to the frequent access tier, Amazon S3 Standard.**

Glacier retrieval
Works well for archived data that requires immediate access to retrieve objects within a few milliseconds
When you decide between the options for archival storage, consider how quickly you must retrieve the archived objects. You can retrieve objects stored in the Amazon S3 Glacier Instant Retrieval storage class within milliseconds, with the same performance as Amazon S3 Standard.

Glacier Flexible Retrieval
Low-cost storage designed for data archiving
Able to retrieve objects within a few minutes to hours
Amazon S3 Glacier Flexible Retrieval is a low-cost storage class that is ideal for data archiving. For example, you might use this storage class to store archived customer records or older photos and video files.

Glacier deep archive
is the Lowest-cost object storage class ideal for archiving. Able to retrieve objects within 12 hours. Amazon S3 Deep Archive supports long-term retention and digital preservation of data that might be accessed once or twice in a year. This storage class is the lowest-cost storage in the AWS Cloud, with data retrieval from 12 to 48 hours. All objects from this storage class are replicated and stored across at least three geographically dispersed Availability Zones.

Block Storage VS Object Storage
Block storage can be changed and micro updates can be applied however in the case of object storage it can't be changed and need to be reuploaded as a whole if any change is required

AWS File Storage

In file storage, multiple clients (such as users, applications, servers, and so on) can access data that is stored in shared file folders. In this approach, a storage server uses block storage with a local file system to organize files. Clients access data through file paths.

EBS is a single zone but EFS is multi-zone..EFS is a file system as a whole

Amazon Elastic File System is a cloud storage service provided by Amazon Web Services designed to provide scalable, elastic, concurrent with some restrictions, and encrypted file storage for use with both AWS cloud services and on-premises resources.

RDBMS

Amazon RDS database engines
Amazon Relational Database Service (RDS) is a managed SQL database service provided by Amazon Web Services (AWS). Amazon RDS supports an array of database engines to store and organize data. It also helps with relational database management tasks, such as data migration, backup, recovery and patching.

Amazon RDS is available on six database engines, which optimize for memory, performance, or input/output (I/O). Supported database engines include:

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

Amazon Aurora
is an enterprise-class relational database. It is compatible with MySQL and PostgreSQL relational databases. It is up to five times faster than standard MySQL databases and up to three times faster than standard PostgreSQL databases.

Nonrelational databases

In a nonrelational database, you create tables. A table is a place where you can store and query data.

Nonrelational databases are sometimes referred to as “NoSQL databases” because they use structures other than rows and columns to organize data. One type of structural approach for nonrelational databases is key-value pairs. With key-value pairs, data is organized into items (keys), and items have attributes (values). You can think of attributes as being different features of your data.

Amazon DynamoDB
Amazon DynamoDB is a key-value database service. It delivers single-digit millisecond performance at any scale.

Amazon Redshift
Amazon Redshift is a data warehousing service that you can use for big data analytics. It offers the ability to collect data from many sources and helps you to understand relationships and trends across your data.

Amazon DocumentDB
Amazon DocumentDB is a document database service that supports MongoDB workloads. (MongoDB is a document database program.)

Amazon Neptune
Amazon Neptune is a graph database service. You can use Amazon Neptune to build and run applications that work with highly connected datasets, such as recommendation engines, fraud detection, and knowledge graphs.

Amazon Neptune
Amazon Neptune is a graph database service

Amazon Managed Blockchain

is a service that you can use to create and manage blockchain networks with open-source frameworks.

A **blockchain** is a distributed ledger with growing lists of records (blocks) that are securely linked together via cryptographic hashes. Each block contains a cryptographic hash of the previous block, a timestamp, and transaction data (generally represented as a Merkle tree, where data nodes are represented by leaves)..

Security

Security is a **shared responsibility**.
In the shared responsibility model, AWS treat your AWS environment as a single object. Rather, you treat the environment as a collection of parts that build upon each other. AWS is responsible for some parts of your environment and you (the customer) are responsible for other parts. This concept is known as the shared responsibility model.**AWS** secures cloud infrastructure whereas Customers are responsible for services that run in the cloud.

AWS Identity and Access Management, or IAM

IAM users
An IAM user is an identity that you create in AWS. It represents the person or application that interacts with AWS services and resources. It consists of a name and credentials.

By default, when you create a new IAM user in AWS, it has no permissions associated with it. To allow the IAM user to perform specific actions in AWS, such as launching an Amazon EC2 instance or creating an Amazon S3 bucket, you must grant the IAM user the necessary permissions.

AWS Organizations

Suppose that your company has multiple AWS accounts. You can use AWS Organizations to consolidate and manage multiple AWS accounts within a central location.

When you create an organization, AWS Organizations automatically creates a root, which is the parent container for all the accounts in your organization.

AWS Artifact

Suppose that your company has multiple AWS accounts. You can use AWS Organizations to consolidate and manage multiple AWS accounts within a central location.

When you create an organization, AWS Organizations automatically creates a root, which is the parent container for all the accounts in your organization.

Basically purpose is to get documentation

AWS Shield

AWS Shield is a service that protects applications against DDoS attacks. AWS Shield provides two levels of protection: Standard and Advanced.
AWS Shield

AWS Shield is a service that protects applications against DDoS attacks. AWS Shield provides two levels of protection: Standard and Advanced.

AWS Shield Advanced
AWS Shield Advanced is a paid service that provides detailed attack diagnostics and the ability to detect and mitigate sophisticated DDoS attacks.

It also integrates with other services such as Amazon CloudFront, Amazon Route 53, and Elastic Load Balancing. Additionally, you can integrate AWS Shield with AWS WAF by writing custom rules to mitigate complex DDoS attacks.

AWS Key Management Service (AWS KMS)

AWS Key Management Service (AWS KMS) lets you create, manage, and control cryptographic keys across your applications and more than 100 AWS services.

Other Tools

AWS WAF

AWS WAF is a web application firewall that lets you monitor network requests that come into your web applications.
AWS WAF works together with Amazon CloudFront and an Application Load Balancer. Recall the network access control lists that you learned about in an earlier module. AWS WAF works in a similar way to block or allow traffic. However, it does this by using a web access control list (ACL) to protect your AWS resources.

Amazon Inspector

Suppose that the developers at the coffee shop are developing and testing a new ordering application. They want to make sure that they are designing the application in accordance with security best practices. However, they have several other applications to develop, so they cannot spend much time conducting manual assessments. To perform automated security assessments, they decide to use Amazon Inspector.

Amazon CloudWatch

Amazon CloudWatch is a web service that enables you to monitor and manage various metrics and configure alarm actions based on data from those metrics.
metrics to CloudWatch. CloudWatch then uses these metrics to create graphs automatically

CloudWatch uses metrics to represent the data points for your resources. AWS services send
that show how performance has changed over time.

AWS CloudTrail

AWS CloudTrail records API calls for your account. The recorded information includes the identity of the API caller, the time of the API call, the source IP address of the API caller, and more. You can think of CloudTrail as a “trail” of breadcrumbs (or a log of actions) that someone has left behind them.

AWS Trusted Advisor

AWS Trusted Advisor is a web service that inspects your AWS environment and provides real-time recommendations in accordance with AWS best practices.

Trusted Advisor compares its findings to AWS best practices in five categories: cost optimization, performance, security, fault tolerance, and service limits. For the checks in each category, Trusted Advisor offers a list of recommended actions and additional resources to learn more about AWS best practices.

.

Basic CLoud Terminologies

IP Address

An IP address is a unique address that identifies a device on the internet or a local network. IP stands for "Internet Protocol," which is the set of rules governing the format of data sent via the internet or local network.

RFC1918

An RFC1918 address is an IP address that is assigned by an enterprise organization to an internal host. These IP addresses are used in private networks, which are not available, or reachable, from the Internet. In fact, one of the basic requirements of the Internet is that each host has a unique **IP address**.
Each RFC1918 address is of the form IPV4-x.x.x.x x=(0-255)

Network Address Translation (NAT)

Network address translation is a method of mapping an IP address space into another by modifying network address information in the IP header of packets while they are in transit across a traffic routing device.

Zero Trust Security

Zero Trust is a strategic approach to cybersecurity that secures an organization by eliminating implicit trust and continuously validating every stage of digital interaction. In short

- Trust None
- Authenticate evert request
- No inherent, automatic security within a perimeter

Escalation of privileges

Privilege escalation is the act of exploiting a bug, a design flaw, or a configuration oversight in an operating system or software application to gain elevated access to resources that are normally protected from an application or user.

.

Single Sign-on (SSO)

Single sign-on (SSO) is an authentication method that enables users to securely authenticate with multiple applications and websites by using just one set of credentials.

Principle of Least Privilege

The principle of least privilege (PoLP) is an information security concept which maintains that a user or entity should only have access to the specific data, resources and applications needed to complete a required task.

East-West VS North-South Network Traffic

In computer networking, east-west traffic is network traffic among devices within a specific data centre. The other direction of traffic flow is north-south traffic, data flowing from or to a system physically residing outside the data centre.

the modern term for them would be:
Ingress- entering traffic
Egress- Leaving Traffic

Hybrid Network

A hybrid cloud network is a network that enables data transfers between on-premises IT resources, private clouds and public clouds, in other words, a hybrid cloud.

VPN Tunnelling

A VPN is a secure, encrypted connection over a publicly shared network. Tunnelling is the process by which VPN packets reach their intended destination, which is typically a private network. Many VPNs use the IPsec protocol suite.

Imperative vs Declarative code

imperative code focuses on writing an explicit sequence of commands to describe how you want the computer to do things, and **declarative code** focuses on specifying the result of what you want.

Big Data and Machine Learning Fundamentals

Data analysis products are known as **Big Query** for **Google**, **Redshift** for **Amazon** and **Machine learning** for **Azure**

Google Machine Learning Products

Compute Engine
Google Compute Engine (GCE) is an infrastructure as a service (IaaS) offering that allows clients to run workloads on Google's physical hardware. Google Compute Engine provides a scalable number of virtual machines (VMs) to serve as large compute clusters for that purpose. It's an **infrastructure as a service** (IAAS)

Google Kubernetes Engine(GKE)
Google Kubernetes Engine (GKE) provides a managed environment for deploying, managing, and scaling your containerized applications using Google infrastructure. The GKE environment consists of multiple machines (specifically, Compute Engine instances) grouped together to form a cluster

App Engine
App Engine is a fully managed, serverless platform for developing and hosting web applications at scale. You can choose from several popular languages, libraries, and frameworks to develop your apps, and then let App Engine take care of provisioning servers and scaling your app instances based on demand. It's a **Software as a Service**(SaaS)

Cloud Function
helps you to run your code in the cloud with no servers or containers to manage with our scalable, pay-as-you-go functions as a service (FaaS) product.

App Engine supports many different services within a single application, **Cloud Functions** support individualized services. It's an important detail when comparing Google App Engine vs Cloud Functions. If your requirements don't include multiple services then Cloud Functions is a great choice.

Cloud run
is a managed computing platform that lets you run containers directly on top of Google's scalable infrastructure. You can deploy code written in any programming language on Cloud Run if you can build a container image from it. In fact, building container images is optional.

Tensor Processing Unit (TPU)

Tensor Processing Units (TPUs) are Google's custom-developed application-specific integrated circuits (ASICs) used to accelerate machine learning workloads. TPUs are designed from the ground up with the benefit of Google's deep experience and leadership in machine learning. It's google's special processor to process Machine learning tasks.

Google Storage Units

Standard/Hot Data
Frequently Accessed data.ome examples of the uses for this type of storage would be interactive video editing, web content, online transactions and the like.

Nearline/Once per month
examples of nearline storage include tape and disk libraries. Nearline storage is slower than online storage, which uses media that are permanently mounted on the system.

Coldine Storage/Once every 90 days

Archive Storage
this is the cheapest storage type, but the retrieval cost of data is high. These are mostly used for tasks like backups

GFS(google file storage)
Google File System is a proprietary distributed file system developed by Google to provide efficient, reliable access to data using large clusters of commodity hardware. Generally used to share petabytes of data

Map Reduce
MapReduce is a popular computing framework that is widely used for big data processing in cloud platforms. Cloud computing as a distributed computing paradigm, provides an environment to perform large-scale data processing. It enables massive data analytics on available computer nodes by using the MapReduce platform.

Colossus
Colossus is our cluster-level file system, a successor to the Google File System (GFS). Spanner is our globally-consistent, scalable relational database. Borg is a scalable job scheduler that launches everything from compute to storage services.

Machine Learning Cycle

The Process has 4 stages

- ingestion and Process
- Storage
- Analytics
- Machine Learning

Ingestion and Process

Data ingestion is the process of importing large, assorted data files from multiple sources into a single, cloud-based storage medium—a data warehouse, data mart or database—where it can be accessed and analyzed.

Pub/Sub/Data Flow
Pub-Sub (or Pub/Sub messaging), as it is often called, works like this: a publisher (any source of data) sends messages out to interested subscribers (the receivers of data) using live-feed or real-time data streams known as channels (or sometimes called topics).

Pub/Sub is a scalable, durable event ingestion and delivery system. Dataflow compliments Pub/Sub's scalable, at-least-once delivery model with message deduplication and exactly-once, in-order processing if you use windows and buffering.

Data Processing Pipeline
is a method in which raw data is ingested from various data sources and then ported to data store, like a data lake or data warehouse, for analysis. Before data flows into a data repository, it usually undergoes some data processing.

For achieve this **Apache Beam** can be used Apache Beam is an open-source unified programming model to define and execute data processing pipelines, including ETL, batch and stream processing

We can also use **Data flow pipelines** so that we don't have to create pipelines forms scratch

Storage

For storage we can make use of Google's Cloud Storage or a DBMS solution like Cloud SQL- RDBMS or Cloud Spanner- RDBMS

Analytics

we can make use of **Big Query** for data analytics

BigQuery is Google's fully managed, serverless data warehouse that enables scalable analysis over petabytes of data. It is a Platform as a Service that supports querying using ANSI SQL. It also has built-in machine-learning capabilities. BigQuery was announced in May 2010 and made generally available in November 2011.

Interactive Vs Batch Queries
In interactive queries you can execute your queries all at once, but however in Batch queries the queries are executed one by one.

The advantage of using Big Query is that you can create ML models with simple SQL models.

Hyperparameters
are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. The prefix 'hyper_' suggests that they are 'top-level' parameters that control the learning process and the model parameters that result from it. it short, the settings we use for creating the prediction model

Supervised Model
Supervised learning is a machine learning paradigm for problems where the available data consists of labelled examples, meaning that each data point contains features and an associated label. They are of two types: **Regression** and **Classification**

Unsupervised learning
is a type of algorithm that learns patterns from untagged data. The goal is that through mimicry, which is an important mode of learning in people, the machine is forced to build a concise representation of its world and then generate imaginative content from it. Model. The two types of unsupervised learning are **clustering** and **association rules.**

The different stages in the lifecycle of Big Query are:

- Extract transform and load data
- Select and Preprocess Features
- Create a model inside Big Query
- Evaluate the Performance of the model
- Use the model to make predictions

Repeat

Machine Learning

Google is an AI-first company.

The various tools that can be used to create the models are:

Big Query ML with SQL
BigQuery ML lets you create and execute machine learning models in BigQuery using GoogleSQL queries. BigQuery ML democratizes machine learning by letting SQL practitioners build models using existing SQL tools and skills. BigQuery ML increases development speed by eliminating the need to move data.

Pre-built API
Pre-trained APIs for vision, video, natural language, and more. Easily infuse vision, video, translation, and natural language ML into existing applications or build entirely new intelligent applications across a broad range of use cases (including Translation and Speech to Text). Advantage of this is no data is required. Best for used cases like NLP, Text to voice, etc.

AutoML
AutoML enables developers with limited machine learning expertise to train high-quality models specific to their business needs. Build your own custom machine learning model in minutes. This enables normal users to build models without much expertise.

Vertex AI
The unified platform for all Machine learning solutions.Vertex AI is a machine learning (ML) platform that lets you train and deploy ML models and AI applications. Vertex AI combines data engineering, data science, and ML engineering workflows, enabling your teams to collaborate using a common toolset.

Ai Solutions

Horizontal Ai solutions
aim towards solving larger, generalized problems. Horizontal AI focuses on solving a larger problem statement, For instance, Apple's Siri and Amazon's Alexa are examples of Horizontal AI applications.

Vertical Ai Solutions
Vertical AI companies are focused on mastering a single, large use case. Waymo and Vara are two often given examples of such vertical AI companies. Other examples are Ai used in Retail Product Discovery, Healthcare data engine, lending DocAI.

Ai Platform

An enterprise AI platform is an integrated set of technologies that enables organizations to design, develop, deploy, and operate enterprise AI applications at scale.Top Artificial Intelligence Platforms: Google AI Platform, TensorFlow, Rainbird, Dialogflow

Feature Store

A feature store is an emerging, ML-specific data system used to centralize storage, processing, and access to frequently used features, making them available for reuse in the development of future machine learning models.

Data Preparation

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labelling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data.Upload data and perform feature engineering. Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

Supervised learning

Task-driven machine learning which identifies a goal., needs a label.

Classification Model
A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. Feature: A feature is an individual measurable property of a phenomenon being observed.

Regression Model
A regression model provides a function that describes the relationship between one or more independent variables and a response, dependent, or target variable. For example, the relationship between height and weight may be described by a linear regression model. For example use past sales of an item to predict future trend.

Unsupervised learning

Clustering
Grouping unlabeled examples is called clustering. As the examples are unlabeled, clustering relies on unsupervised machine learning. If the examples are labelled, then clustering becomes classification.Use customer demographics to determine customer segmentation

Association
allow you to predict which items are most likely to appear together, and predict the strength of the relationship between them. Identifying underlying relationships. Example Correlation between two products placed in a grocery store

Dimensional Reduction
Dimensionality reduction, or dimension reduction, is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension. Combing characteristics to create a quote.

Model Evaluation

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

Confusion Matrix

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm. it shows how much the data is off as compared to the original results.

Recall vs Precision

Precision means the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by your algorithm.

Model Serving

The basic meaning of model serving is to host machine-learning models (on the cloud or on-premises) and to make their functions available via API so that applications can incorporate AI into their systems.

Model Deployment

Model deployment is the process of putting machine learning models into production. This makes the model's predictions available to users, developers or systems, so they can make business decisions based on data, interact with their application (like recognising a face in an image) and so on.

Deploying is the process of putting the model into the server. **Serving** is the process of making a model accessible from the server (for example with REST API or web sockets).

RESTful API is an interface that two computer systems use to exchange information securely over the internet. Most business applications have to communicate with other internal and third-party applications to perform various tasks.

Linux & SQL on Azure

Linux infrastructure as a service (IaaS)
With IaaS, you deploy your applications and workloads to virtual machines that run in Azure. These VMs are connected to each other and to the internet by a virtual network infrastructure that you define and configure.

Linux platforms as a service (PaaS)
PaaS enables you to deploy applications to the cloud without managing infrastructure. These managed service platforms trade the flexibility of virtual infrastructure for reduced maintenance concerns and easier scalability. Azure PaaS services let you control, configure, and deploy your applications with the same centralized, globally available Azure Resource Manager management tools and libraries that you use for provisioning. - A Serverless Solution.

Azure Resource Manager (ARM)
Azure Resource Manager provides a standard interface and set of concepts for managing every kind of Azure service and platform. Manage all of your resources and workloads in a single, browser-based graphical experience with the Azure portal. Automate resource provisioning and management from the command line with Azure PowerShell and the Azure CLI, available locally and in the browser via the Azure Cloud Shell.

.

VM types:

General purpose: Balanced CPU-to-memory ratio. Ideal for testing and development, small to medium databases, and low to medium traffic web servers.

Compute optimized: High CPU-to-memory ratio. Suitable for medium-traffic web servers, network appliances, batch processes, and application servers.
Memory optimized: High memory-to-CPU ratio. Great for relational database servers, medium to large caches, and in-memory analytics.

Burstable VMs: For workloads that don't require the full CPU performance all the time. These VMs can be purchased with VM-size baseline performance. This means that if a VM is using less than the baseline, it builds up credits. When higher CPU performance is required, it can burst up to 100 per cent of the CPU.

Storage optimized: Optimized for storage-intensive workloads. High disk throughput and input/output (I/O), ideal for big data, SQL databases, NoSQL databases, data warehousing, and large transactional databases.

GPU-enabled VM(Hardware Accelerated)s: Specialized VMs targeted for heavy graphic rendering and video editing, as well as model training and inferencing (ND series) with deep learning. Available with single or multiple GPUs.

High-performance compute Fastest and most powerful CPU VMs available with optional high-throughput network interfaces (remote direct memory access).

Azure Automage

Azure Automanage machine best practices eliminate the need to discover and know how to onboard and configure certain Azure services to benefit your virtual machines. These Azure services help enhance reliability, security, and management for virtual machines. basically scheduled servicing

Polybase

PolyBase enables your SQL Server instance to query data with T-SQL directly from SQL Server, Oracle, Teradata, MongoDB, Hadoop clusters, Cosmos DB, and S3-compatible object storage without separately installing client connection software. You can also use the generic ODBC connector to connect to additional providers using third-party ODBC drivers. PolyBase allows T-SQL queries to join the data from external sources to relational tables in an instance of SQL Server.

MEAN

MEAN is a development stack for building and hosting web applications. MEAN is an acronym for its component parts: **MongoDB**, **Express**, **AngularJS**, and **Node.js.**

Why would I pick MEAN?
All of the components of the MEAN stack are reliable, well-understood, and open source, but so are many other development stacks. Here are some reasons you might choose MEAN over other development stacks.

Your data isn't highly structured
MongoDB is what's called a NoSQL database. A NoSQL database doesn't require data to be structured in a pre-defined way as it would with a relational database like Microsoft SQL Server or MySQL. Instead, MongoDB stores its data in JSON-like documents that don't require the rigid data structures that MySQL or other relational databases require.

MEAN is well documented
The components of the MEAN stack are all popular right now. Resources for working with MongoDB, Express, AngularJS, and Node.js are easy to find.

MEAN runs almost anywhere
You can also develop MEAN stack applications from your favourite development environment - whether that's Windows, macOS, or Linux.

MEAN is scalable
In addition to being cross-platform, MEAN stack applications can be scaled out and easily tested for accelerated growth in enterprise environments and offer high performance.

Common SQL Server performance counters

Many performance counters are included with SQL Server 2019. Each counter can give detailed and precise information, and help to diagnose performance bottlenecks. But you must understand clearly what each counter means. The following list explains some of the most commonly used counters

Batch Requests/Sec. This counter measures the rate at which SQL Server is receiving requests from clients. Use this counter to measure demand on the server.

User Connections. This counter measures the number of users currently connected to the database. Again, use this counter to measure demand on the server.

Buffer Cache Hit Ratio. This counter measures the proportion of requests that SQL Server can satisfy by returning data pages from its buffer cache in memory. When the proportion is high, most requests are returned without obtaining data from the hard drives, which respond more slowly than memory. A high value indicates optimal performance.

SQL Compilations/Sec. This counter measures the rate at which SQL Server compiles execution plans. This process is resource-intensive. If this counter is more than 10% of the value of Batch Requests/Sec, then some complication may be lowering performance by rendering execution plans invalid.

Page Life Expectancy. This counter measures the average time a page remains in the buffer cache. In general, a page life expectancy of less than 300 seconds might indicate that your server would do better with extra physical memory.

When diagnosing performance issues, it's often necessary to identify changes in demand or behavior over time. If your company is growing, for example, user demand on the database server might increase over months or years. You should record values of common counters when you know that your server is doing well, as a baseline. Compare later measurements of these counters against the baseline to spot changes that might create bottlenecks.

Future Plans

1. Microsoft Certified: Azure Fundamentals

Next goal would be to aim for the the Microsoft Certification on Cloud Computing, which maybe followed by the more difficult Google's Cloud leader certification exam.

2. Use the knowledge of clouds to host my projects

Till now I had been making static projects, but it would be interesting to creating more complex projects with more complex Backend and Frontend infrastructure.

3. Try out a ML model on cloud

Futre plans are to create an ML model which can predict the Air Quality Index of an area, I still need more research, but most probably I am going to try it out for my chemistry project