Azure Data Analytics Interview Questions

1. What is Cloud Technology?

A cloud is a combination of services, networks, hardware, storage, and interfaces that help in delivering computing as a service. It broadly has three users. These are the end-user, business management user, and cloud service provider. The end-user is the one who uses the services provided by the cloud. The responsibility of the data and the services provided by the cloud is taken by the business management user in the cloud. The one who takes care of or is responsible for the maintenance of the IT assets of the cloud is the cloud service provider. The cloud acts as a common center for its users to fulfill their computing needs.

2. What are some of the key features of Cloud Computing?

The following are some of the key features of cloud computing:

- Agility: Helps in quick and inexpensive re-provisioning of resources.
- Location Independence: This means that the resources can be accessed from everywhere.
- Multi-Tenancy: The resources are shared amongst a large group of users.
- Reliability: Resources and computation can be dependable for accessibility.
- Scalability: Dynamic provisioning of data helps in scaling.

3. What do you mean by cloud delivery models?

Cloud delivery models are models that represent the computing environments. These are as follows:

- Infrastructure as a Service (IaaS): Infrastructure as a Service (IaaS) is the delivery of services, including an operating system, storage, networking, and various utility software elements, on a request basis.
- Platform as a Service (PaaS): Platform as a Service (PaaS) is a mechanism for combining Infrastructure as a Service with an abstracted set of middleware services, software development, and deployment tools. These allow the organization to have a consistent way to create and deploy applications on a cloud or on-premises environment.

- Software as a Service (SaaS): Software as a Service (SaaS) is a business application created and hosted by a provider in a multi-tenant model.
- Function as a Service (FaaS): Function as a Service (FaaS) gives a platform for customers to build, manage and run app functionalities without the difficulty of maintaining infrastructure. One can thus achieve a "serverless" architecture.

4. What are the different versions of the cloud?

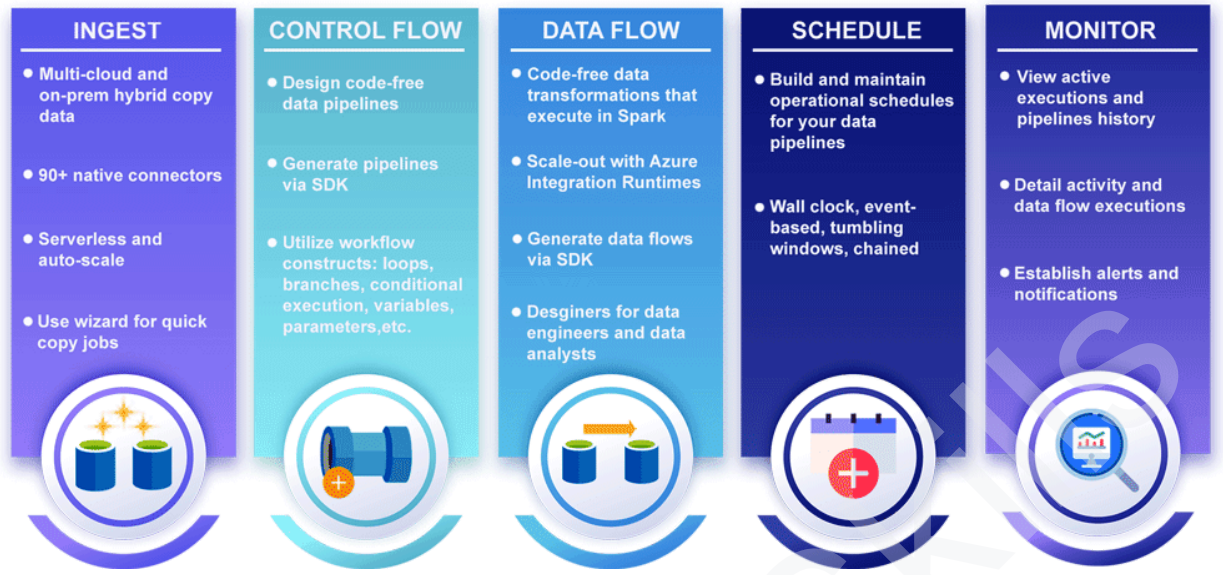
There are two primary deployment models of the cloud: Public and Private.

- Public Cloud: The set of hardware, networking, storage, services, applications, and interfaces owned and operated by a third party for use by other companies or individuals is the public cloud. These commercial providers create a highly scalable data center that hides the details of the underlying infrastructure from the consumer. Public clouds are viable because they offer many options for computing, storage, and a rich set of other services.
- Private Cloud: The set of hardware, networking, storage, services, applications, and interfaces owned and operated by an organization for the use of its employees, partners, or customers is the private cloud. This can be created and managed by a third party for the exclusive use of one enterprise. The private cloud is a highly controlled environment not open for public consumption. Thus, it sits behind a firewall.
- Hybrid Cloud: Most companies use a combination of private computing resources and public services, called the hybrid cloud environment.
- Multi-Cloud: Some companies, in addition, also use a variety of public cloud services to support the different developer and business units – called a multi-cloud environment.

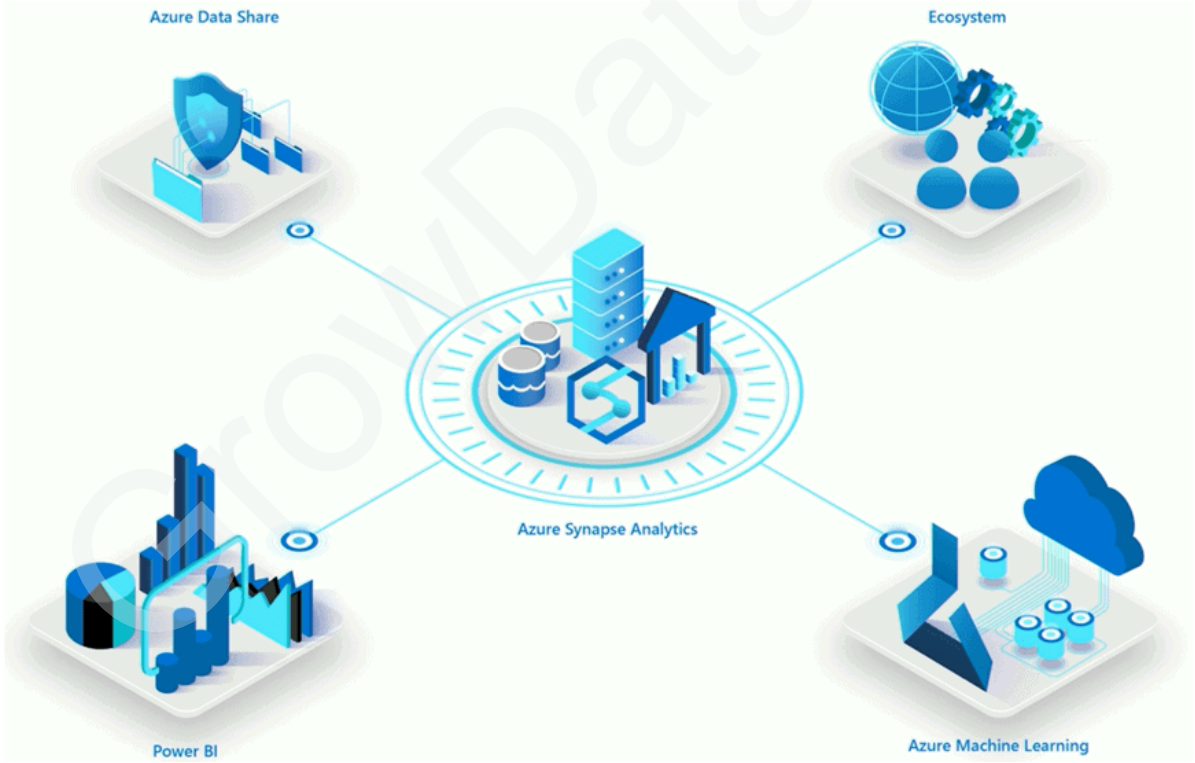
5) What is Microsoft Azure?

Microsoft Azure is a cloud computing platform that provides both hardware and software. The service provider creates a managed service here to enable users to access these services on demand.

6) What is the primary ETL service in Azure?



7) Which service would you use to create Data Warehouse in Azure?



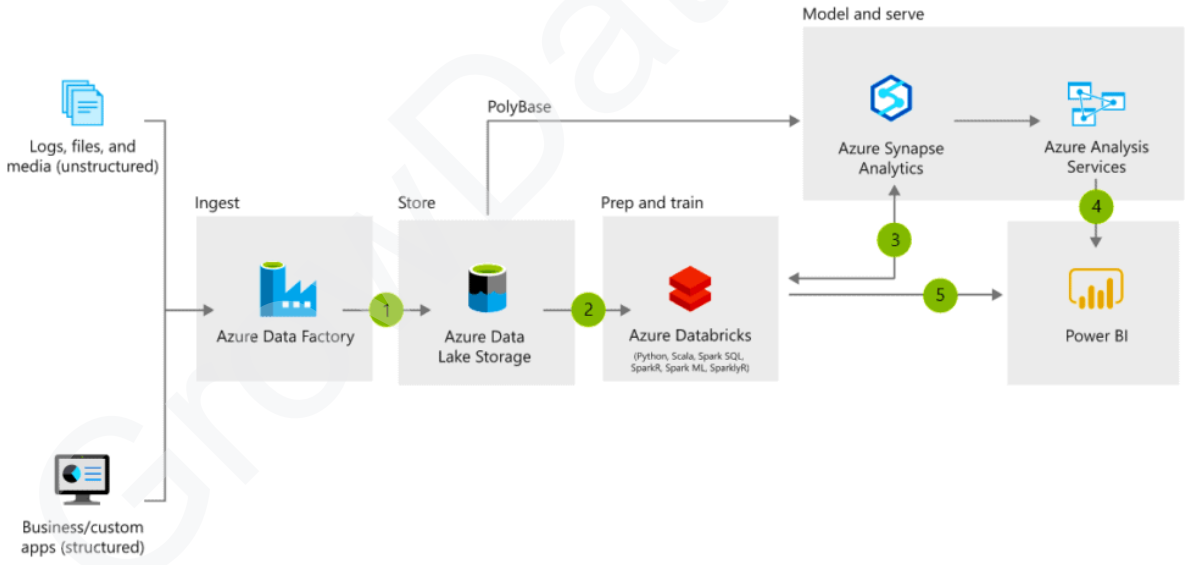
Azure Synapse is a limitless analytics service that brings together Big Data analytics and enterprise data warehousing. It gives users the freedom to query data on individual terms for using either serverless on-demand or provisioned resources at scale.

8) Explain the architecture of Azure Synapse Analytics

It is designed to process massive amounts of data with hundreds of millions of rows in a table. Azure Synapse Analytics processes complex queries and returns the query results within seconds, even with massive data, because Synapse SQL runs on a Massively Parallel Processing (MPP) architecture that distributes data processing across multiple nodes.

Applications connect to a control node that acts as a point of entry to the Synapse Analytics MPP engine. On receiving the Synapse SQL query, the control node breaks it down into MPP optimized format. Further, the individual operations are forwarded to the compute nodes that can perform the operations in parallel, resulting in much better query performance.

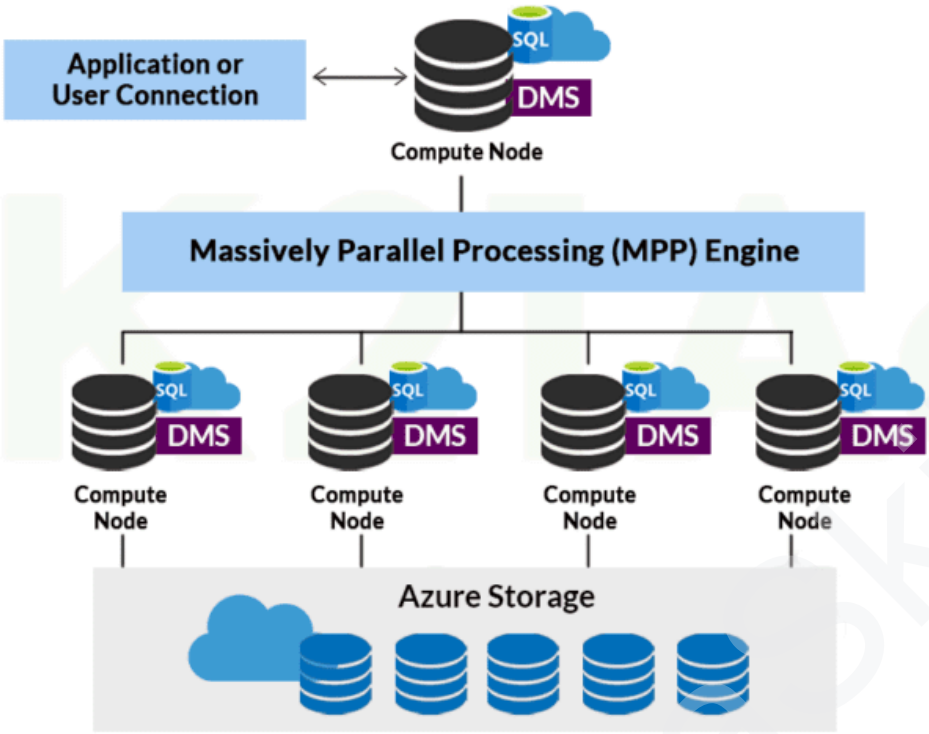
9) Difference between ADLS and Azure Synapse Analytics



Both Azure Data Lake Storage Gen2 and Azure Synapse Analytics are highly scalable and can ingest and process vast amounts of data (on a Peta Byte scale). But there are some differences:

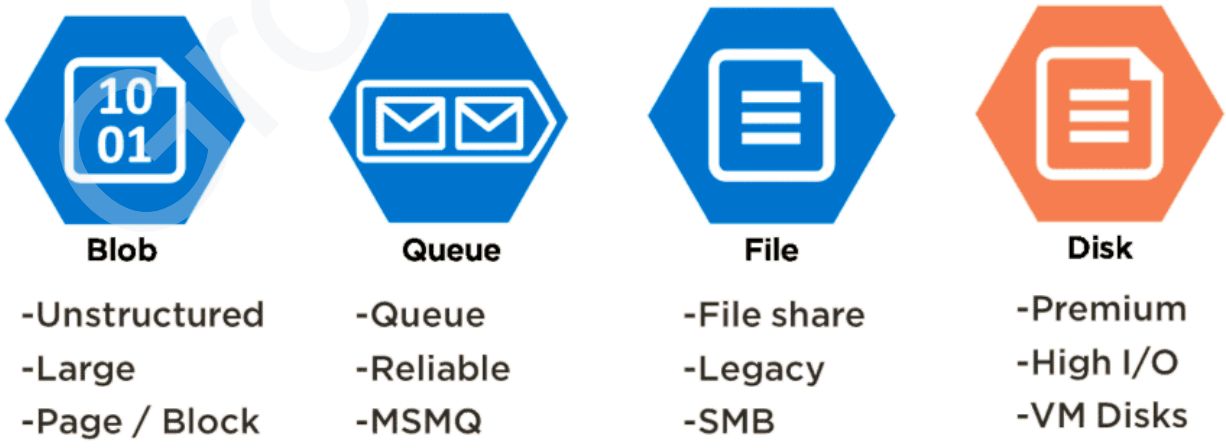
ADLS Gen2	Azure Synapse Analytics
• Optimised for storing and processing structured and non-structured data	• Optimised for processing structured data in a well-defined schema
• Used for data exploration and analytics by data scientists and engineers	• Used for Business Analytics or disseminating data to business users
• Built to work with Hadoop	• Built on SQL Server
• No regulatory compliance	• Compliant with regulatory standards such as HIPAA
• USQL (combination of C# and TSQL) and Hadoop are used for accessing data	• Synapse SQL (improved version of TSQL) is used for accessing data
• Can handle data streaming using tools such as Azure Stream Analytics	• Built-in data pipelines and data streaming capabilities

10) What are Dedicated SQL Pools?



Dedicated SQL Pool is a collection of features that enable the implementation of the more traditional Enterprise Data Warehousing platform using Azure Synapse Analytics. The resources are measured in Data Warehousing Units (DWU) that are provisioned using Synapse SQL. A dedicated SQL pool uses columnar storage and relational tables to store data, improving query performance and reducing the required amount of storage.

11) What are the different types of storage in Azure?



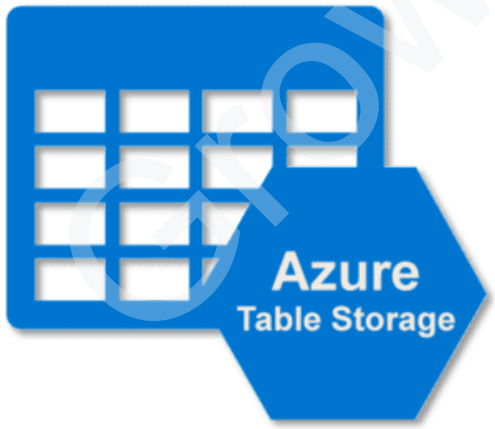
There are five types of storage in Azure:

- Azure Blobs: Blob stands for a large binary object. It can support all kinds of files including, text files, videos, images, documents, binary data etc.
- Azure Queues: Azure Queues is a cloud-based messaging store for establishing and brokering communication between various applications and components.
- Azure Files: It is an organised way of storing data in the cloud. Azure Files has one main advantage over Azure Blobs, it allows organising the data in a folder structure, and it is SMB compliant, i.e. it can be used as a file share.
- Azure Disks: It is used as a storage solution for Azure VMs (Virtual Machines).
- Azure Tables: A NoSQL storage solution for storing structured data which does not meet the standard relational database schema.

12) Explore Azure storage explorer and its uses?

It is a versatile standalone application available for Windows, Mac OS and Linux to manage Azure Storage from any platform. Azure Storage can be downloaded from Microsoft.
It provides access to multiple Azure data stores such as ADLS Gen2, Cosmos DB, Blobs, Queues, Tables, etc., with an easy to navigate GUI.
One of the key features of Azure Storage Explorer is that it allows users to work even when they are disconnected from the Azure cloud service by attaching local emulators.

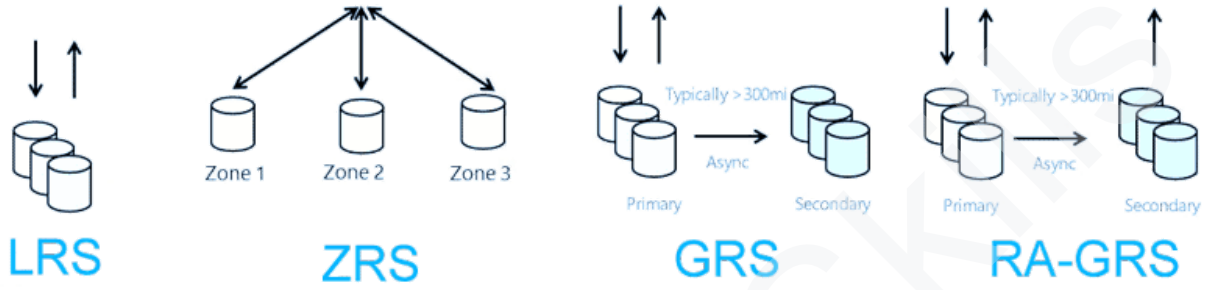
13) What is Azure table storage?



It is a storage service optimised for storing structured data. In structured data, table entities are the basic units of data equivalent to rows in a relational database table. Each entity represents a key-value pair, and the properties for table entities are as follows:

- *PartitionKey:* It stores the key of the partition to which the table entity belongs.
- *RowKey:* It identifies the entity uniquely within the partition.
- *TimeStamp:* It stores the last modified date/time value for the table entity.

14) What is data redundancy in Azure?



Azure constantly retains several copies of data to provide high levels of data availability. Some data redundancy solutions are accessible to clients in Azure, depending on the criticality and duration necessary to provide access to the replica.

- Locally Redundant Storage (LRS): In this type, data is replicated across different racks in the same data center. It is the cheapest redundancy option and ensures that there are at least three copies of the data.
- Zone Redundant Storage (ZRS): It ensures that data is replicated across three zones within the primary region. Azure takes care of DNS repointing automatically in case of zone failure. It might require a few changes to the network settings for any applications accessing data after the DNS repointing.
- Geo-Redundant Storage (GRS): In this type, data is replicated across two regions and ensures that data can be recovered if one entire region goes down. It may take some time for the Geo failover to complete and make data accessible in the secondary region.
- Read Access Geo Redundant Storage (RA-GRS): It is much similar to GRS but with the added option of reading access to the data in the secondary region in case of a failure in the primary region.

15) What are pipelines and activities in Azure?

The grouping of activities arranged to accomplish a task together is known as Pipelines. It allows users to manage the individual activities as a single group and provide a quick overview of the activities involved in a complex task with many steps.

ADF activities are grouped into three parts:

- Data Movement Activities – Used to ingest data into Azure or export data from Azure to external data stores.
- Data Transformation Activities – Related to data processing and extracting information from data.
- Control Activities – Specify a condition or affect the progress of the pipeline.

16) What is the trigger execution in Azure Data Factory?

In Azure Data Factory, pipelines can be triggered or automated.

Some ways to automate or trigger the execution of Azure Data Factory Pipelines are:

- Schedule Trigger: It invokes a pipeline execution at a fixed time or on a fixed schedule such as weekly, monthly etc.
- Tumbling Window Trigger: It executes Azure Data Factory Pipeline at fixed periodic time intervals without overlap from a specified start time.
- Event-Based Trigger: It executes an Azure Data Factory Pipeline based on the occurrence of some event, such as the arrival or deletion of a new file in Azure Blob Storage.

17) What is Azure Data Lake Analytics used for?

Azure Data Lake Analytics is used for running big data analysis and processing jobs on data stored in Azure Data Lake Store or Azure Blob Storage.

18) Which Azure service is used for real-time analytics?

Azure Stream Analytics is used for real-time analytics.

19) What is Azure Data Factory used for?

Azure Data Factory is used for orchestrating and automating data movement and data transformation.

20) How does Azure Data Lake Store differ from Azure Blob Storage?

Azure Data Lake Store is optimized for big data analytics and is designed for storing large amounts of data in a hierarchical structure, whereas Azure Blob Storage is a general-purpose object storage solution.

21) What is the purpose of Azure Data Explorer?

Azure Data Explorer is used for ingesting, storing, and analyzing large volumes of data quickly.

22) What are the benefits of using Azure Synapse Analytics?

Benefits include seamless integration with various data sources, scalability, and the ability to perform both data warehousing and big data analytics tasks.

23) Explain the difference between Azure Data Factory and Azure Logic Apps.

Azure Data Factory is primarily used for data integration and orchestration tasks, while Azure Logic Apps is a platform for automating workflows and integrating systems and services.

24)Explain the difference between batch processing and stream processing in the context of Azure Data Analytics.

Batch processing involves processing large volumes of data at scheduled intervals, while stream processing involves processing data in real-time as it is generated.

25) How does Azure Data Lake Storage Gen2 differ from its predecessor?

Azure Data Lake Storage Gen2 includes features such as hierarchical namespace, improved scalability, and lower latency compared to its predecessor.

26) Describe the role of Azure Data Catalog in Azure Data Analytics.

Azure Data Catalog is used for discovering, understanding, and managing data assets across various data sources within an organization.

27) What are some common security measures for securing data in Azure Data Analytics solutions?

Security measures include role-based access control, encryption at rest and in transit, auditing, and monitoring of data access and activities.

28) Explain the difference between cloud and on-premise?

Cloud computing and on-premise computing are two distinct models for delivering computing resources and applications, each with its own set of characteristics, advantages, and challenges. Here's an explanation of the key differences between them:

1. Location:

- Cloud: In cloud computing, resources such as servers, storage, databases, networking, software, and applications are hosted on remote servers maintained by a third-party provider, typically accessed over the internet. Users access these resources on-demand and pay based on usage.

- On-premise: On-premise computing refers to the traditional model where all hardware, software, and supporting infrastructure are owned, operated, and maintained by the organization itself, physically located within their premises (office, data center, etc.).

2. Ownership and Control:

- Cloud: In a cloud model, the cloud provider owns and manages the infrastructure, including hardware, networking, and virtualization layers. Users have control over their applications and data but have limited control over the underlying infrastructure.

- On-premise: With on-premise computing, the organization has full control and ownership of all hardware, software, and infrastructure. This allows for greater customization and control over security and compliance.

3. Scalability and Flexibility:

- Cloud: Cloud computing offers scalability and flexibility, allowing users to easily scale resources up or down based on demand. Users can quickly provision additional resources as needed without upfront investment in hardware.

- On-premise: On-premise solutions may require significant upfront investment in hardware and infrastructure. Scaling up may involve purchasing and deploying additional hardware, which can be time-consuming and expensive.

4. Cost Structure:

- Cloud: Cloud computing typically operates on a pay-as-you-go or subscription-based pricing model, where users pay only for the resources they use. This can result in cost savings for organizations, particularly for smaller businesses or startups that may not have the capital for large upfront investments.

- On-premise: On-premise solutions involve significant upfront capital expenditure for purchasing hardware and software licenses. While there may be lower ongoing costs for maintenance and support, the total cost of ownership over time can be higher compared to cloud solutions.

5. Security and Compliance:

- Cloud: Cloud providers invest heavily in security measures to protect data and infrastructure. However, some organizations may have concerns about data privacy and compliance with industry regulations when storing sensitive data in the cloud.

-On-premise: On-premise solutions offer more direct control over security measures and data compliance, which may be necessary for industries with strict regulatory requirements. However, this also means that the organization bears full responsibility for implementing and maintaining security measures.

6. Maintenance and Updates:

- Cloud: Cloud providers handle maintenance, updates, and upgrades of the underlying infrastructure and services, relieving the users from these tasks. This allows organizations to focus on their core business activities rather than IT maintenance.

- On-premises`: With on-premise solutions, organizations are responsible for managing and maintaining all hardware, software, and infrastructure components, including applying patches, updates, and upgrades. This can require dedicated IT staff and resources.

In summary, while cloud computing offers benefits such as scalability, flexibility, and cost-effectiveness, on-premise solutions provide greater control, security, and compliance for organizations with specific requirements or regulatory constraints. The choice between cloud and on-premise often depends on factors such as budget, security needs, regulatory compliance, and the organization's overall IT strategy.

29. Is Azure Data Factory ETL or ELT tool?

It is a cloud-based Microsoft tool that provides a cloud-based integration service for data analytics at scale and supports ETL and ELT paradigms.

30. Why is ADF needed?

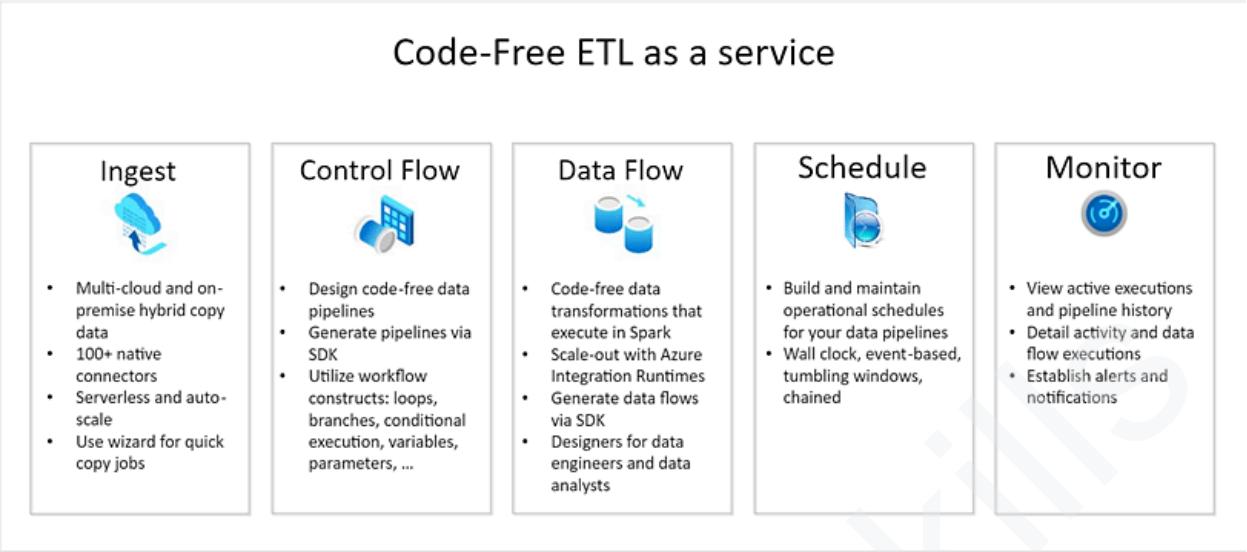
With an increasing amount of big data, there is a need for a service like ADF that can orchestrate and operationalize processes to refine the enormous stores of raw business data into actionable business insights.

31. What sets Azure Data Factory apart from conventional ETL tools?

Azure Data Factory stands out from other ETL tools as it provides: -

- Enterprise Readiness: Data integration at Cloud Scale for big data analytics!
- Enterprise Data Readiness: There are 90+ connectors supported to get your data from any disparate sources to the Azure cloud!
- Code-Free Transformation: UI-driven mapping dataflows.
- Ability to run Code on Any Azure Compute: Hands-on data transformations
- Ability to rehost on-prem services on Azure Cloud in 3 Steps: Many SSIS packages run on Azure cloud.
- Making DataOps seamless: with Source control, automated deploy & simple templates.
- Secure Data Integration: Managed virtual networks protect against data exfiltration, which, in turn, simplifies your networking.

Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers. The below snippet summarizes the same.



32. What are the major components of a Data Factory?

To work with Data Factory effectively, one must be aware of below concepts/components associated with it: -

- Pipelines: Data Factory can contain one or more pipelines, which is a logical grouping of tasks/activities to perform a task. An activity can read data from Azure blob storage and load it into Cosmos DB or Synapse DB for analytics while transforming the data according to business logic. This way, one can work with a set of activities using one entity rather than dealing with several tasks individually.
- Activities: Activities represent a processing step in a pipeline. For example, you might use a copy activity to copy data between data stores. Data Factory supports data movement, transformations, and control activities.
- Datasets: Datasets represent data structures within the data stores, which simply point to or reference the data you want to use in your activities as inputs or outputs.
- Linked Service: This is more like a connection string, which will hold the information that Data Factory can connect to various sources. In the case of reading from Azure Blob storage, the storage-linked service will specify the connection string to connect to the blob, and the Azure blob dataset will select the container and folder containing the data.
- Integration Runtime: Integration runtime instances bridged the activity and linked Service. The linked Service or activity references it and provides the computing

environment where the activity runs or gets dispatched. This way, the activity can be performed in the region closest to the target data stores or compute Services in the most performant way while meeting security (no publicly exposing data) and compliance needs.

- Data Flows: These are objects you build visually in Data Factory, which transform data at scale on backend Spark services. You do not need to understand programming or Spark internals. Design your data transformation intent using graphs (Mapping) or spreadsheets (Power query activity).

33. What are the different ways to execute pipelines in Azure Data Factory?

There are three ways in which we can execute a pipeline in Data Factory:

- Debug mode can be helpful when trying out pipeline code and acts as a tool to test and troubleshoot our code.
- Manual Execution is what we do by clicking on the 'Trigger now' option in a pipeline. This is useful if you want to run your pipelines on an ad-hoc basis.
- We can schedule our pipelines at predefined times and intervals via a Trigger. As we will see later in this article, there are three types of triggers available in Data Factory.

34. What is the purpose of Linked services in Azure Data Factory?

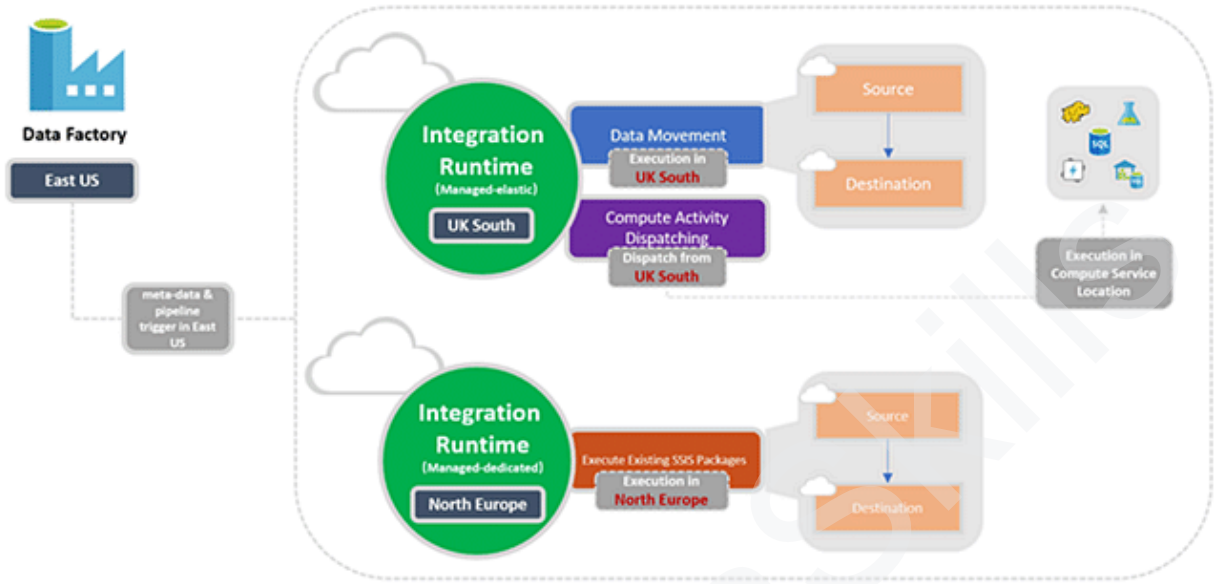
Linked services are used majorly for two purposes in Data Factory:

1. For a Data Store representation, i.e., any storage system like Azure Blob storage account, a file share, or an Oracle DB/ SQL Server instance.
2. For Compute representation, i.e., the underlying VM will execute the activity defined in the pipeline.

35. Can you Elaborate more on Data Factory Integration Runtime?

The Integration Runtime, or IR, is the compute infrastructure for Azure Data Factory pipelines. It is the bridge between activities and linked services. The linked Service or Activity references it and provides the computing environment where the activity is run directly or dispatched. This allows the activity to be performed in the closest region to the target data stores or computing Services.

The following diagram shows the location settings for Data Factory and its integration runtimes:



Source: https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

Azure Data Factory supports three types of integration runtime, and one should choose based on their data integration capabilities and network environment requirements.

1. Azure Integration Runtime: To copy data between cloud data stores and send activity to various computing services such as SQL Server, Azure HDInsight, etc.
2. Self-Hosted Integration Runtime: Used for running copy activity between cloud data stores and data stores in private networks. Self-hosted integration runtime is software with the same code as the Azure Integration Runtime but installed on your local system or machine over a virtual network.
3. Azure SSIS Integration Runtime: You can run SSIS packages in a managed environment. So, when we lift and shift SSIS packages to the data factory, we use Azure SSIS Integration Runtime.

36). Which three activities can you run in Microsoft Azure Data Factory?

Azure Data Factory supports three activities: data movement, transformation, and control activities.

- Data movement activities: As the name suggests, these activities help move data from one place to another.
  e.g., Copy Activity in Data Factory copies data from a source to a sink data store.
- Data transformation activities: These activities help transform the data while we load it into the data's target or destination.
  e.g., Stored Procedure, U-SQL, Azure Functions, etc.
- Control flow activities: Control (flow) activities help control the flow of any activity in a pipeline.

e.g., wait activity makes the pipeline wait for a specified time.

36. How does Azure Synapse differ from Azure Databricks?

Azure Synapse is a data integration service with some amazing transformation capabilities while Azure Databricks is data analytics focused platform build on top of Spark. Azure Synapse integrates big data analytics and enterprise data warehouse into a single platform. Databricks allows customers to develop complex machine learning algorithms and perform big data analytics. However, both both Synapse Analytics and Azure Databricks can be used together when building a data pipeline.