**What is the use of Azure Active Directory?**

Azure Active Directory is an identify and access management system. It is very much similar to the active directories. It allows you to grant your employee in accessing specific products and services within the network.

**Briefly describe the purpose of the ADF Service**

ADF is used mainly to orchestrate the data copying between different relational and non-relational data sources, hosted in the cloud or locally in your datacentres’. In addition, ADF can be used for transforming the ingested data to meet your business requirements. It is ETL, or ELT tool for data ingestion in most Big Data solutions.

**Data Factory consists of a number of components. Mention these components briefly**

* Pipeline: The activities logical container
* Activity: An execution step in the Data Factory pipeline that can be used for data ingestion and transformation
* Mapping Data Flow: A data transformation UI logic
* Dataset: A pointer to the data used in the pipeline activities
* Linked Service: A descriptive connection string for the data sources used in the pipeline activities
* Trigger: Specify when the pipeline will be executed
* Control flow: Controls the execution flow of the pipeline activities

**What is the difference between the Dataset and Linked Service in Data Factory?**

Linked Service is a description of the connection string that is used to connect to the data stores. For example, when ingesting data from a SQL Server instance, the linked service contains the name for the SQL Server instance and the credentials used to connect to that instance.

Dataset is a reference to the data store that is described by the linked service. When ingesting data from a SQL Server instance, the dataset points to the name of the table that contains the target data or the query that returns data from different tables.

**What is Data Factory Integration Runtime?**

Integration Runtime is a secure compute infrastructure that is used by Data Factory to provide the data integration capabilities across the different network environments and make sure that these activities will be executed in the closest possible region to the data store.

**What is the difference between the Mapping data flow and Wrangling data flow transformation activities in Data Factory?**

Mapping data flow activity is a visually designed data transformation activity that allows us to design a graphical data transformation logic without the need to be an expert developer, and executed as an activity within the ADF pipeline on an ADF fully managed scaled-out Spark cluster.

Wrangling data flow activity is a code-free data preparation activity that integrates with Power Query Online in order to make the Power Query M functions available for data wrangling using spark execution.

[What is blob storage in Azure?](https://intellipaat.com/blog/interview-question/azure-data-factory-interview-questions/#6)  
[What is the difference between Azure Data Lake store and Blob storage?](https://intellipaat.com/blog/interview-question/azure-data-factory-interview-questions/#7)  
[What are the steps for creating ETL process in Azure Data Factory?](https://intellipaat.com/blog/interview-question/azure-data-factory-interview-questions/#8)  
[What is the difference between HDInsight & Azure Data Lake Analytics?](https://intellipaat.com/blog/interview-question/azure-data-factory-interview-questions/#9)  
[What are the top-level concepts of Azure Data Factory?](https://intellipaat.com/blog/interview-question/azure-data-factory-interview-questions/#10)

Explain triggers in ADF.

**Steps for Creating ETL**

* Create a Linked Service for source data store which is SQL Server Database
* Assume that we have a cars dataset
* Create a Linked Service for destination data store which is Azure Data Lake Store
* Create a dataset for Data Saving
* Create the pipeline and add copy activity
* Schedule the pipeline by adding a trigger
* What are the top-level concepts of Azure Data Factory?
* **Pipeline**: It acts as a carrier in which we have various processes taking place.
* This individual process is an activity.
* **Activities**: Activities represent the processing steps in a pipeline. A pipeline can have one or multiple activities. It can be anything i.e process like querying a data set or moving the dataset from one source to another.
* **Datasets**: Sources of data. In simple words, a data structure holds our data.
* **Linked services**: These store information that is very important when it comes to connecting an external source.
* For example: Consider SQL server, you need a connection string that you can connect to an external device. You need to mention the source and the destination of your data.

How to create and connect to Azure SQL Database?

First, we need to log into the Azure Portal with our Azure credentials. Then we need to create an Azure SQL database in the Azure portal.

Click on “Create a resource” on the left side menu and it will open an “Azure Marketplace”. There, we can see the list of services. Click “Databases” then click on the “SQL Database”.

Create a SQL database

After clicking the “SQL Database”, it will open another section. There, we need to provide the basic information about our database like Database name, Storage Space, Server name, etc.

What is Azure Blob Storage?

Azure Storage is one of the cloud computing PaaS (Platform as a Service) services provided by the Microsoft Azure team. It provides cloud storage that is highly available, secure, durable, scalable, and redundant. It is massively scalable and elastic. It can store and process hundreds of terabytes of data or you can store the small amounts of data required for a small business website.

What is Blob?

Blob is a service for storing large amounts of unstructured data that can be accessed from anywhere in the world via HTTP or HTTPS.” Blob stands for ” Binary Large Object “. It’s designed to store large amounts of unstructured text or binary data like virtual hard disks, videos, images or even log files.

The data can be exposed to the public or stored privately. It scales up or down as your needs change. We no longer manage it, we only pay for what we use.

**Control flows and scale**

To support the diverse integration flows and patterns in the modern data warehouse, Data Factory enables flexible data pipeline modeling. This entails full control flow programming paradigms, which include conditional execution, branching in data pipelines, and the ability to explicitly pass parameters within and across these flows. Control flow also encompasses transforming data through activity dispatch to external execution engines and data flow capabilities, including data movement at scale, via the Copy activity.

Data Factory provides freedom to model any flow style that's required for data integration and that can be dispatched on demand or repeatedly on a schedule. A few common flows that this model enables are:

**Control flows**:

Activities can be chained together in a sequence within a pipeline.

Activities can be branched within a pipeline.

**Parameters:**

Parameters can be defined at the pipeline level and arguments can be passed while you invoke the pipeline on demand or from a trigger.

Activities can consume the arguments that are passed to the pipeline.

**Custom state passing:**

Activity outputs, including state, can be consumed by a subsequent activity in the pipeline.

**Looping containers:**

The foreach activity will iterate over a specified collection of activities in a loop.

**Trigger-based flows:**

Pipelines can be triggered on demand, by wall-clock time, or in response to driven by event grid topics

**Delta flows:**

Parameters can be used to define your high-water mark for delta copy while moving dimension or reference tables from a relational store, either on-premises or in the cloud, to load the data into the lake.

**What are the top-level concepts of Azure Data Factory?**

An Azure subscription can have one or more Azure Data Factory instances (or data factories). Azure Data Factory contains four key components that work together as a platform on which you can compose data-driven workflows with steps to move and transform data.

**Pipelines**

A data factory can have one or more pipelines. A pipeline is a logical grouping of activities to perform a unit of work. Together, the activities in a pipeline perform a task. For example, a pipeline can contain a group of activities that ingest data from an Azure blob and then run a Hive query on an HDInsight cluster to partition the data. The benefit is that you can use a pipeline to manage the activities as a set instead of having to manage each activity individually. You can chain together the activities in a pipeline to operate them sequentially, or you can operate them independently, in parallel.

**Data flows**

Data flows are objects that 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. Just design your data transformation intent using graphs (Mapping) or spreadsheets (Wrangling).

**Activities**

Activities represent a processing step in a pipeline. For example, you can use a Copy activity to copy data from one data store to another data store. Similarly, you can use a Hive activity, which runs a Hive query on an Azure HDInsight cluster to transform or analyze your data. Data Factory supports three types of activities: data movement activities, data transformation activities, 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 services**

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources. Think of it this way: A linked service defines the connection to the data source, and a dataset represents the structure of the data. For example, an Azure Storage linked service specifies the connection string to connect to the Azure Storage account. And an Azure blob dataset specifies the blob container and the folder that contains the data.

Linked services have two purposes in Data Factory:

To represent a *data store* that includes, but is not limited to, a SQL Server instance, an Oracle database instance, a file share, or an Azure Blob storage account. For a list of supported data stores, see [Copy Activity in Azure Data Factory](https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview).

To represent a *compute resource* that can host the execution of an activity. For example, the HDInsight Hive activity runs on an HDInsight Hadoop cluster. For a list of transformation activities and supported compute environments, see [Transform data in Azure Data Factory](https://docs.microsoft.com/en-us/azure/data-factory/transform-data).