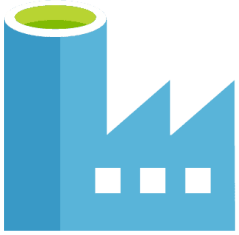
## 1) Why do we need Azure Data Factory?


Azure Data factory doesn't store any data itself; it lets you **produce workflows** that orchestrate the movement of data between supported data stores and data processing. You can **monitor** and **manage** your workflows using both **programmatic** and **UI mechanisms**. Apart from that, it is the best tool available today for ETL processes with an easy-to-use interface. This shows the need for Azure Data Factory.

## 2) What is Azure Data Factory?



Azure Data Factory is a **cloud-based integration service** offered by Microsoft that lets you **create data-driven workflows** for orchestrating and **automating data movement** and data transformation overcloud. Data Factory services also offer to create and running data pipelines that move and transform data and then run the pipeline on a specified schedule.

## 3) What is Integration Runtime?


Integration runtime is nothing but a **compute structure** used by Azure Data Factory to give integration capabilities across different network environments.

**Types of Integration Runtimes:**

- **Azure Integration Runtime** – It can copy data between cloud data stores and dispatch the activity to a variety of computing services such as **SQL Server**, Azure HDInsight
- **Self Hosted Integration Runtime** – It's software with basically the same code as Azure Integration runtime, but it's installed on on- premises systems or virtual machines over virtual networks.

- **Azure SSIS Integration Runtime** – It helps to execute SSIS packages in a managed environment. So when we lift and shift the SSIS packages to the data factory, we use Azure SSIS Integration Runtime.
- **4) How much is the limit on the number of integration runtimes?**

There's **no specific limit** on the number of integration runtime instances. But there's a limit on the number of VM cores used by Integration runtime grounded on per subscription for SSIS package execution.

## 5) What are the different components used in Azure Data Factory?

Azure Data Factory consists of several numbers of components. Some components are as follows:

- **Pipeline:** The pipeline is the logical container of the activities.
- **Activity:** It specifies the execution step in the **Data Factory pipeline**, which is substantially used for data ingestion and metamorphosis.
- **Dataset:** A dataset specifies the pointer to the data used in the pipeline conditioning.
- **Mapping Data Flow:** It specifies the data transformation UI logic.
- **Linked Service:** It specifies the descriptive connection string for the data sources used in the channel conditioning.
- **Trigger:** It specifies the time when the pipeline will be executed.
- **Control flow:** It's used to control the execution flow of the pipeline activities.

## 6) What is the key difference between the Dataset and Linked Service in Azure Data Factory?

Dataset specifies a source to the data store described by the linked service. When we put data to the dataset from a SQL Server instance, the dataset indicates the table's name that contains the target data or the query that returns data from dissimilar tables.
Linked service specifies a definition of the connection string used to connect to the data stores. For illustration, when we put data in a linked service from a SQL Server instance, the linked service contains the name for the SQL Server instance and the credentials used to connect to that case.

## 7) How many types of triggers are supported by Azure Data Factory?

Following are the three types of triggers that Azure Data Factory supports:

1. **Tumbling Window Trigger:** The Tumbling Window Detector executes the Azure Data Factory pipelines over cyclic intervals.
It's also used to maintain the state of the pipeline.

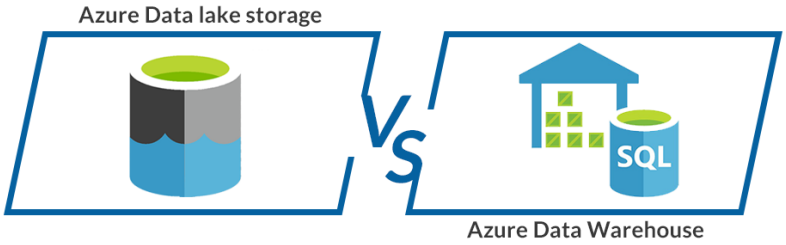
2. **Event-based Trigger:** The Event– based Trigger creates a response to any event related to blob storage. These can be created when you add or cancel blob storage.
3. **Schedule Trigger:** The Schedule Trigger executes the Azure Data Factory pipelines that follow the wall clock timetable.

## 8) What are the different rich cross-platform SDKs for advanced users in Azure Data Factory?

The Azure Data Factory V2 provides a rich set of SDKs that we can use to write, manage, and watch pipelines by applying our favourite IDE. Some popular cross-platform SDKs for advanced users in Azure Data Factory are as follows:

- Python SDK
- C# SDK
- PowerShell CLI
- Users can also use the documented REST APIs to interface with Azure Data Factory V2

## 9) What is the difference between Azure Data Lake and Azure Data Warehouse?



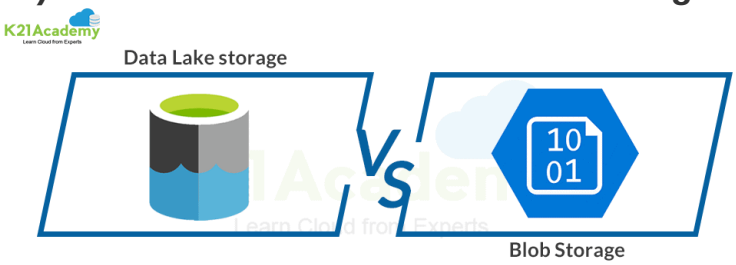

| Azure Data Lake | Data Warehouse |
|---|---|
| Data Lake is a capable way of storing any type, size, and shape of data. | Data Warehouse acts as a repository for already fi from a specific resource. |
| It is mainly used by **Data Scientists**. | It is more frequently used by **Business Profess** |
| It is **highly accessible** with quicker updates. | It becomes a **pretty rigid and costly task** to mak in Data Warehouse. |
| It defines the schema after when the data is stored successfully. | Datawarehouse defines the schema before storing |
| It uses **ELT** (Extract, Load and Transform) process. | It uses **ETL** (Extract, Transform and Load) pro |
| It is an ideal platform for doing **in-depth analysis.** | It is the best platform for **operational user** |

# Intermediate ADF Interview Questions

## 10) What is Blob Storage in Azure?

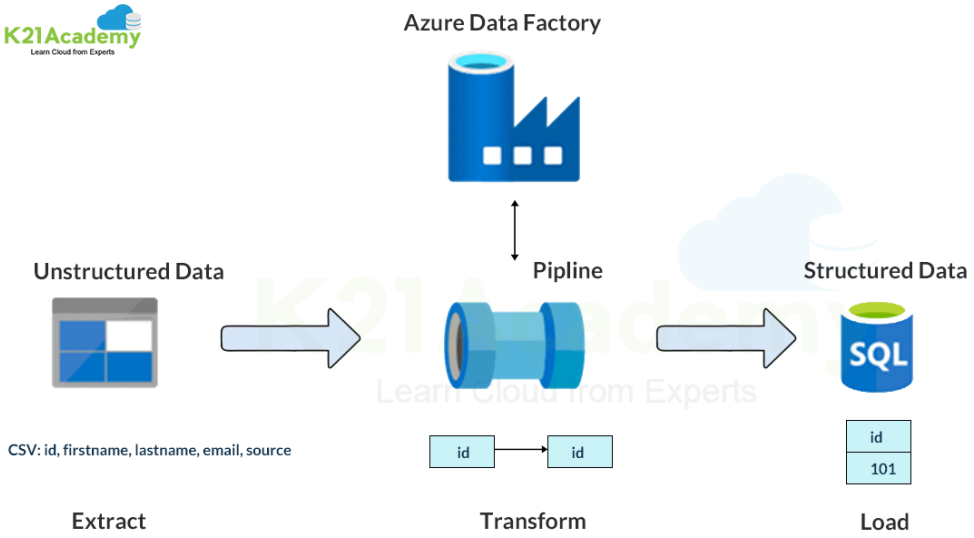

It helps to store a large amount of unstructured data similar as text, images or double data. It can be used to expose data intimately to the world. Blob storage is most commonly used for streaming audios or videos, storing data for backup, and disaster recovery, storing data for analysis etc. You can also create Data Lakes using blob storage to perform analytics.

## 11) Difference between Data Lake Storage and Blob Storage.



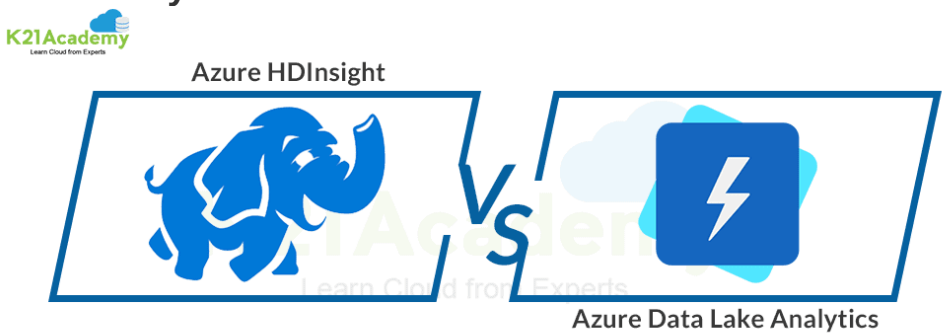

| Data Lake Storage | Blob Storage |
|---|---|
| It is an optimized storage solution for **big data analytics workloads**. | Blob Storage is **general-purpose storage** for a w of scenarios. It can also do Big Data Analyt |
| It follows a **hierarchical file system**. | It follows an object store with a **flat namesp** |
| In Data Lake Storage, data is stored as **files** inside **folders**. | Blob storage lets you create a **storage account** account has **containers** that store the dat |
| It can be used to store Batch, interactive, stream analytics, and machine learning data. | We can use it to store text files, binary data, medi for streaming and general purpose data |

## 12) What are the steps to create an ETL process in Azure Data Factory?




- There are **straightforward steps** to create an ETL process.
- We need to create a service for a **linked data store** which is an **SQL Server Database**.
- Let's assume that we have a car dataset.
- For this car's dataset, we can create a **linked service** for the **destination data store** that is Azure Data Lake.
- Now create a **data set** for Data Saving.
- Create a **Pipeline** and **Copy Activity**.

## 13) What is the difference between Azure HDInsight and Azure Data Lake Analytics?




| Azure HDInsight | Azure Data Lake Analytics |
|---|---|
| It is a **Platform as a Service**. | It is a **Software as a Service**. |
| Processing data in it **requires configuring the cluster** with predefined nodes. Further, by using languages like pig or hive, we can process the data. | It is all about **passing the queries** written for processing. Data Lake Analytics further creates nodes to process the data set. |

| | |
|---|---|
| Users can easily configure HDInsight Clusters at their convenience. Users can also use Spark, Kafka, without restrictions. | It does not give that much flexibility in terms of co and customization. But, Azure manages it automat users. |

## 14) What are the top-level concepts of Azure Data Factory?

There are four basic top-level concepts of Azure Data Factory:

- **Pipeline** – It acts as a carrier where lots of processes take place.
- **Activities** – It represents the steps of processes in the pipeline.
- **Data Sets** – It is a data structure that holds our data.
- **Linked Services**–
  These services store information that's essential while connecting the resources or other services. Let's say we've an
  SQL server, so we need a connecting string connected to an external device, and we will mention the source and the destination for it.

## 15) What are the key differences between the Mapping data flow and Wrangling data flow transformation activities in Azure Data Factory?

In Azure Data Factory, the main dissimilarity between the Mapping data flow and the Wrangling data flow transformation activities is as follows
The Mapping data flow activity is a visually allowed data transformation activity that facilitates users to plan graphical data transformation logic. It does not need the users to be expert developers. It's executed as an activity within the ADF pipeline on an ADF completely managed scaled-out Spark cluster.
On the other hand, the Wrangling data flow activity is a code–free data preparation activity. It's integrated with Power Query Online to make the Power Query M functions available for data wrangling using spark execution.

## 16) Is the knowledge of coding required for Azure Data Factory?

No, it isn't necessary to have good wisdom in coding for **Azure Data Factory**. Azure Data Factory provides 90 built-in connectors to transform the data using mapping data flow activities without the wisdom of programming skills or spark cluster knowledge. It likewise facilitates us to produce workflows veritably and quickly.

## 17) What changes can we see regarding data flows from private preview to limited public preview?

Following is a list of some important changes we can see regarding data flows from private preview to limited public preview:

- We do not need to bring our own Azure Databricks Clusters.
- We can still use the Data Lake Storage Gen 2 and Blob Storage to store those files.
- Azure Data Factory will address the cluster creation and tear–down process.
- Blob data sets and Azure Data Lake Storage Gen 2 are separated into delimited text and Apache Parquet datasets.

- We can use the appropriate linked services for those storage engines

# Advanced ADF Interview Questions

## 18) How can we schedule a pipeline?

The trigger follows a **world clock calendar schedule** that can schedule pipelines periodically or in calendar-based recurrent patterns. We can schedule a pipeline in two ways:

- **Schedule Trigger**
- **Window Trigger**

## 19) Can we pass parameters to a pipeline run?

**Yes definitely**, we can very easily pass parameters to a pipeline run. Pipeline runs are the **first-class**, **top-level concepts** in Azure Data Factory. We can define parameters at the pipeline level, and then we can pass the arguments to run a pipeline.

## 20) Can I define default values for the pipeline parameters?

You can define default values for the parameters in the pipelines.

## 21) Can an activity in a pipeline consume arguments that are passed to a pipeline run?

Each activity within the pipeline can consume the parameter value that's passed to the pipeline and run with the **@parameter** construct.

## 22) Can an activity output property be consumed in another activity?

An activity output can be consumed in a subsequent activity with the @activity construct.

## 23) How do I gracefully handle null values in an activity output?

You can use the @coalesce construct in the expressions to handle the null values gracefully.

## 24) Which Data Factory version do I use to create data flows?

Use the Data Factory V2 version to create data flows.

## 25) What has changed from private preview to limited public preview in regard to data flows?

- You'll no longer have to bring your own Azure Databricks clusters.
- Data Factory will manage cluster creation and tear– down.

- Blob datasets and Azure Data Lake Storage Gen2 datasets are separated into delimited text and Apache Parquet datasets.
- You can still use Data Lake Storage Gen2 and Blob storage to store those files. Use the appropriate linked service for those storage engines.

## 26) How do I access the data using the other 80 Dataset types in Data Factory?

The mapping data flow feature currently allows Azure SQL database, **Data Warehouse**, Delimited text-files from Azure Blob Storage or Azure Data Lake storage to generation tools natively for source and sink. You can use copy activity to states data from any of the other connectors, and then you can execute the data flow activity to transform data.

## 27) Explain the two levels of security in ADLS Gen2?

- **Role-Based Access Control** – It includes **built-in azure rules** such as reader, contributor, owner or customer roles. It is specified for two reasons. The first is, who can manage the service itself, and the second is, to permit the reasons is to permit the users built-in data explorer tools.
- **Access Control List** – Azure Data Lake Storage specifies precisely which data object users may read or write or execute.

## 28) What has changed from private preview to limited public preview regarding data flows?

There are a couple of things which have been changed mentioned below:

- You are no longer required to bring your own **Azure Databricks Clusters**.
- Data Factory will **manage cluster creation** and tear down process.
- We can still use Data Lake Storage Gen 2 and Blob Storage to store those files. You can use the appropriate linked services. You can also use the appropriate linked services for those of the storage engines.
- Blob data sets and Azure Data Lake storage gen 2 are separated into **delimited text** and **Apache Parquet datasets**.

## 29) What is the difference between the Dataset and Linked Service in Data Factory?

- **Dataset:** is a reference to the datastore that is described by Linked Service.
- **Linked Service:** is nothing but a description of the connection string that is used to connect to the data stores.

## 30) What is the difference between the mapping data flow and wrangling data flow transformation?

- **Mapping Data Flow:** It is a visually designed data transformation activity that lets users design a graphical data transformation logic without needing an expert developer.
- **Wrangling Data Flow:** This is a code-free data preparation activity that integrates with Power Query Online.

## 31) Data Factory supports two types of compute environments to execute the transform activities. Mention them briefly.

Let's go through the types:

- **On-demand compute environment** – It is a fully managed environment offered by ADF. In this compute type, a cluster is created to execute the transform activity and removed automatically when the activity is completed.
- **Bring your own environment** – In this environment, you yourself manage the compute environment with the help of **ADF**.

## 32) What is Azure SSIS Integration Runtime?


Azure SSIS Integration is a **fully managed cluster** of virtual machines that are hosted in Azure and dedicated to run SSIS packages in the data factory. We can easily scale up the SSIS nodes

by configuring the node size or scaled out by configuring the number of nodes on the Virtual Machine's cluster.

## 33) What is required to execute an SSIS package in Data Factory?

We need to create an SSIS Integration Runtime, and an SSIS Database catalogue hosted in the **Azure SQL database** or Azure SQL managed instance.

## 34) An Azure Data Factory Pipeline can be executed using three methods. Mention these methods.

Methods to execute Azure Data Factory Pipeline:

- Debug Mode
- Manual execution using trigger now
- Adding schedule, tumbling window/event trigger

## 35) If we need to copy data from an on-premises SQL Server instance using a data factory, which integration runtime should be used?

Self-hosted integration runtime should be installed on the **on-premises machine** where the SQL Server Instance is hosted.

## 36) What is Azure Table Storage?

**Azure Table Storage** is a service that helps users to **store structure data** in the cloud and also provides a Keystore with schemas designed. It is swift and effective for modern-day applications.

## 38) Can we monitor and manage Azure Data Factory Pipelines?

Yes, we can monitor and manage **ADF Pipelines** using the following steps:

- Click on the **monitor and manage** on the data factory tab.
- Click on the **resource manager**.
- Here, you will find- pipelines, datasets, and linked services in a tree format.

### 39) What are the steps involved in the ETL process?
ETL (Extract, Transform, Load) process follows four main steps:

- **Connect and Collect –** helps in moving the data on-premises and cloud source data stores
- **Transform –** lets users collect the data by using compute services such as HDInsight Hadoop, Spark etc.
- **Publish –** Helps in loading the data into Azure data warehouse, Azure SQL database, and Azure Cosmos DB etc
- **Monitor –** It helps support the pipeline monitoring via Azure Monitor, API and PowerShell, Log Analytics, and health panels on the Azure Portal.

# FAQs
### Q. Is coding required for Azure Data Factory?
**Ans:** No, coding is not required. Azure Data Factory lets you create workflows very quickly. It offers 90+ built-in connectors available in Azure Data Factory to transform the data using mapping data flow activities without programming skills or spark cluster knowledge.

### Q. Is Azure Data Factory an ETL tool?
**Ans:** Yes, ADF is the best tool available in the market for ETL processes. Without writing any complex algorithms, it simplifies the entire data migration process.

### Q. Is Azure Data Factory Certification worth doing?
**Ans:** Absolutely, there is a massive demand for Azure Data Engineers proficient in Data Factory. Since lots of companies are adopting Microsoft Azure as a cloud computing platform, so companies need skilful professionals to handle their operations.

### Q. Can we replace Synapse pipelines with other ETL like talend or SSIS?
**Ans:** We can use both azure data factory or synapse with Synapse Pipelines, Data Integration & Orchestration to integrate our data and operationalize all our code development.

### Q. ETL should always happen with Azure Data factory or Synapse Pipelines, or can we use any other ETL tool in the market?
**Ans:** Along with Azure Data Factory and Synapse Pipelines, you can also use data bricks. Data Integration & Orchestration to integrate your data and operationalize all of your code development with Synapse Pipelines.

### Q. If Azure data factory and synapse pipelines have the same functionality then which one to choose and why to choose?
**Ans:** If your requirement is only data movement and transformation then use Azure data factory and For Analytics capabilities go with synapse because Azure synapse analytics is an umbrella service which provides analytical workspace along with other services.

### Q. What is Data Flow Debug?
**Ans:** When Debug mode is on, you'll interactively build your data flow with an active Spark cluster. The session will close once you turn to debug off in **Azure Data Factory**. You should be aware of the hourly charges Azure Databricks incurred when you have the debug session turned on.

### Q. Can we use adf for running 24-by-7 jobs?

**Ans:** Yes we can run the azure data factory 24×7 for loading data If you have that much data.

## Q. Azure Data Bricks we can write transformation logic right then why we require ADF?

**Ans:** Mapping data flows are visually designed data transformations in Azure Data Factory. Data flows allow data engineers to develop data transformation logic without writing code. The resulting data flows are applied as activities within Azure Data Factory pipelines that apply scaled– out Apache Spark clusters. Data flow activities can be operationalized using existing Azure Data Factory scheduling, control, flow, and monitoring capabilities.

## Q. Linked Service – special connectors for ADF to source data?

**Ans:** Yes, you must create a linked service to link your data store to the Data Factory or Synapse Workspace.