Top 30 Azure Data Factory Interview Questions and Answers

This post will cover the **Top 30 Azure Data Factory Interview Questions**. These are well-researched, up to date and the most feasible questions that can be asked in your very next interview.

Azure Data Factory is a **cloud-based ETL service** for scaling out data Integration and transformation. It offers you to lift and shift existing SSIS packages on Azure.

Topics of discussions:

Azure Data Factory Interview Questions and Answers

I have divided Azure Data Factory Interview questions as per their difficulty level. Let's dive right into these questions.

Azure Data Factory Interview Questions Basic Level

Q1) What is Azure Data Factory?

Azure Data Factory is an integration and ETL service offered by Microsoft. You can create **data-driven workflows** to orchestrate and automate data movement. You can also transform the data

over the cloud. It lets you create and run data pipelines that can help move and transform data and run scheduled pipelines.

Q2) Why do we need Azure Data Factory?

As the world moves to the cloud and big data, data integration and migration remain an integral part of enterprises in all industries. ADF helps solve both of these problems efficiently by focusing on the data and planning, monitoring, and managing the ETL / ELT pipeline in a single view.

Check out: ***Azure Data Factory***

The reasons for the growing adoption of Azure Data Factory are:

- Increased value
- Improved results of business processes
- Reduced overhead costs
- Improved decision making
- Increased business process agility

Q3) What do we understand by Integration Runtime?

Integration runtime is referred to as a **compute infrastructure** used by Azure Data Factory. It provides integration capabilities across various network environments.

A quick look at the Types of Integration Runtimes:

- **Azure Integration Runtime** – Can copy data between cloud data stores and send activity to various computing services such as SQL Server, Azure HDInsight, etc.
- **Self Hosted Integration Runtime** – It's basically software with the same code as the Azure Integration Runtime, but it's installed on your local system or virtual machine over a virtual network.
- **Azure SSIS Integration Runtime** – It allows you to 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.

Q4) What is the difference between Azure Data Lake and Azure Data Warehouse?

Azure Data Lake	Data Warehouse
Data Lakes are capable of storing data of any form, size, or shape.	A Data Warehouse is a store for data that has previously been filtered from a specific resource.
Data Scientists are the ones who use it the most.	Business professionals are the ones who use it the most.

It is easily accessible and receives frequent changes.	Changing the Data Warehouse becomes a very strict and costly task.
When the data is correctly stored, it determines the schema.	Before storing the data, the data warehouse defines the schema.
It employs the ELT (Extract, Load, and Transform) method.	It employs the ETL (Extract, Transform, and Load) method.
It's an excellent tool for conducting in-depth research.	It is the finest platform for operational users.

Check out: ***Azure Free Trial Account***

Q5) What is the limit on the number of Integration Runtimes?

There is no restriction on the number of integration runtime instances that can be used. However, the number of VM cores used by Integration runtime for SSIS package execution is limited to one per subscription.

Q6) What is Blob Storage in Azure?

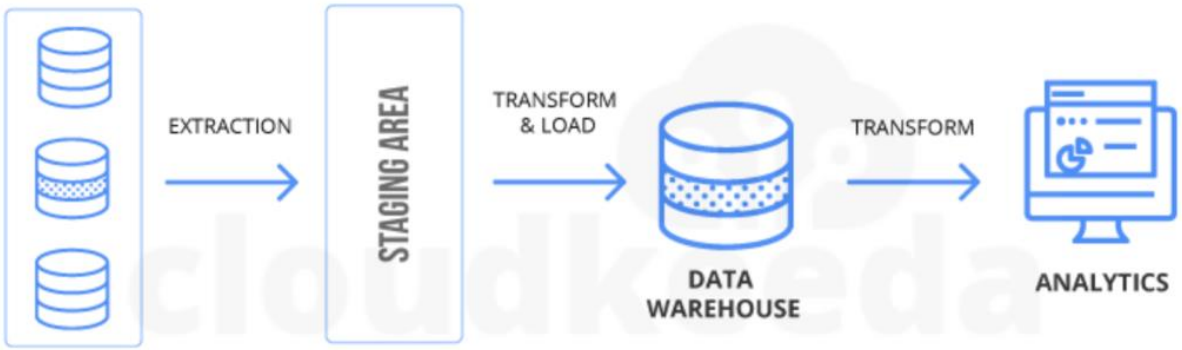
Blob storage is specially designed for **storing a huge amount of unstructured data** such as text, images, binary data. It helps make your data available public globally. The most common use of blob storage is to stream audios and videos, store data for backup, analysis etc. You can also work with data lakes to perform analytics using blob storage.

Q7) Difference between Data Lake Storage and Blob Storage.

Data Lake Storage	Blob Storage
It’s a big data analytics workload-optimized storage solution.	Blob Storage is a type of general-purpose storage that can be used in a variety of situations. It’s also capable of Big Data Analytics.
A hierarchical file system is used.	It’s based on a flat namespace object store.
Data is saved in Data Lake Storage as files within folders.	You can create a storage account with Blob Storage. The data is stored in containers in the storage account.
Batch, interactive, stream analytics, and machine learning data can all be stored in it.	Text files, binary data, media storage for streaming, and general-purpose data can all be stored on it.

Q8) Describe the process to create an ETL process in Azure Data Factory?

You can create an ETL process with a few steps.



- Create a service for linked data store i.e. SQL Server Database.
- Let's consider you have a dataset for vehicles.
- Now for this dataset, you can create a linked service for the destination store i.e. Azure Data Lake.
- Then create a Data Set for Data Saving.
- The next step is to create a pipeline and copy activity. When you are done with creating a pipeline, schedule a pipeline with the use of an added trigger.

Q9) What is the difference between Azure HDInsight and Azure Data Lake Analytics?

Azure HDInsight	Azure Data Lake Analytics
It's a Platform as a Service (PaaS) model.	It's a SaaS (Software as a Service) model.
It needs configuring the cluster with predetermined nodes in order to process data. We can also process the data using languages like pig or hive.	It's all about passing the data processing queries that have been written. To process the data set, Data Lake Analytics creates compute nodes.
HDInsight Clusters can be readily configured by users at their leisure. Users have unrestricted access to Spark and Kafka.	In terms of setting and customization, it does not offer a lot of options. However, Azure handles it for its users automatically.

Q 10) What are the top-level concepts of Azure Data Factory?

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

- **Pipeline** – It acts as a transport service where many processes take place.
- **Activities** – It represents the stages of processes in the pipeline.
- **Datasets** – This is the data structure that holds our data.
- **Linked Services** – These services store information needed when connecting other resources or services. Let's say we have a SQL server, so we need a connection string that is connected to an external device and we will mention its source and destination.

Azure Data Factory Interview Questions Intermediate Level

Q 11) How can we schedule a pipeline?

We can schedule pipelines using a trigger. It follows a world clock calendar schedule. We can schedule pipelines periodically or calendar-based recurrent patterns. Here are the two ways:

- Schedule Trigger
- Window Trigger

Q 12) Is there any way to pass parameters to a pipeline run?

Yes absolutely, passing parameters to a pipeline run is a very easy task. Pipelines are known as the first-class, top-level concepts in Azure Data Factory. We can **set** parameters at the pipeline **level** and then we can pass the arguments to run a pipeline.

Check out: ***What is Azure?***

Q 13) What is the difference between the mapping data flow and wrangling data flow transformation?

- **Mapping Data Flow:** This is a visually designed data conversion activity that allows users to design graphical data conversion logic without the need for experienced developers.
- **Wrangling Data Flow:** This is a code-free data preparation activity built into Power Query Online.

Q 14) How do I access the data using the other 80 Dataset types in Data Factory?

Dataflow mapping now enables Azure SQL databases, data warehouses, Azure Blob storage, or delimited text files in Azure Data Lake storage to build native build tools for source and receiver. You can use a copy operation to declare data from one of the other connectors, then you can run a data stream operation to transform the data.

Q 15) 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 indicated for two reasons. The first is who can manage the service themselves, and the second is to provide users with built-in data mining tools.
- **Access Control Lists** – Azure Data Lake Storage specifies exactly which data objects users can read, write, or execute.

Q 16) What is the difference between the Dataset and Linked Service in Data Factory?

- **Dataset:** A reference to a datastore described by a linked service.
- **Linked Service:** Just a description of the connection string used to connect to the data store.

Q 17) What has changed from private preview to limited public preview regarding data flows?

Some of the things that have changed are mentioned below:

- There is no need to bring your own Azure Databricks clusters now.
- Data Factory will handle cluster creation and deletion.
- We can still use Data Lake Storage Gen 2 and Blob Storage to store these files. You may use the appropriate linked services. You may also use associated services that are appropriate for the services of the storage engines.
- Blob dataset and Azure Data Lake gen 2 storage split into delimited text and Apache Parquet dataset.

Q 18) Data Factory supports two types of compute environments to execute the transform activities. What are those?

Let's take a look at the types.

- **On-Demand Computing Environment** – This is a fully managed environment provided by ADF. This type of

calculation creates a cluster to perform the transformation activity and automatically deletes it when the activity is complete.

- **Bring your own environment** – In this environment, use ADF to manage your computing environment yourself.

Q 19) What is Azure SSIS Integration Runtime?  Azure SSIS Integration is a **fully managed cluster of virtual machines** hosted in Azure and designed to run SSIS packages in your data factory. You can scale up SSIS nodes simply by configuring the node size, or you can scale out by configuring the number of nodes in the virtual machine cluster.

Q 20) What is required to execute an SSIS package in Data Factory?

You need to create an SSIS integration runtime and SSIS database catalog hosted on an Azure SQL database or an Azure SQL managed instance.

Azure Data Factory Interview Questions Advanced Level

Q 21) What is Azure Table Storage?

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

Q 22) Can we monitor and manage Azure Data Factory Pipelines?

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

- Go to the Data factory tab and click on the **monitor and manage.**
- Now click on the **resource manager**.
- You will be able to see pipelines, datasets, and linked services in a tree format.

Q 23) 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

Q 24) 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.

Q 25) What are the steps involved in the ETL process?

The ETL (Extract, Transform, Load) process follows four main steps:

- **Connect and Collect** – Helps move data to local and crowdsource data storage. Transform data using computing services such as HDInsight, Hadoop, Spark etc.
- **Publish** -Useful for loading data into Azure data warehouses, Azure SQL databases, Azure Cosmos DB, and more.
- **Monitor** – Supports Azure Monitor, API and PowerShell, log analysis, and pipeline monitoring through the Azure portal health scope.

Q 26) Can an activity output property be consumed in another activity?

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

Q 27) What is the way to access data by using the other 90 dataset types in Data Factory?

For source and sink, the mapping data flow feature supports Azure SQL Database, Azure Synapse Analytics, delimited text files from Azure Blob storage or Azure Data Lake Storage Gen2, and Parquet files from Blob storage or Data Lake Storage Gen2.

Use the Copy action to stage data from any of the other connectors, then use the Data Flow activity to transform the data once it's staged. For example, your pipeline might copy data into Blob storage first, then transform it with a Data Flow activity that uses a dataset from the source.

Q 28) Is it possible to calculate a value for a new column from the existing column from mapping in ADF?

In the mapping data flow, you can use derive transformation to generate a new column based on the logic you want. You can either create a new derived column or update an existing one when generating a derived column. Enter the name of the column you're creating in the Column textbox.

The column dropdown can be used to override an existing column in your schema. Click the Enter expression textbox to start creating the derived column's expression. You have the option of either inputting your expression or using the expression builder to create your logic.

Q 29) What is the way to parameterize column name in dataflow?

We can pass parameters to columns similar to other properties. Like in derived column customer can use **$ColumnNameParam = toString(byName($myColumnNameParamInData)).** These parameters can be passed from pipeline execution down to Data flows.

Check out: ADF Interview Questions from Microsoft (FAQs)

Q 30) In what way we can write attributes in cosmos DB in the same order as specified in the sink in ADF data flow?

Because each document in Cosmos DB is stored as a JSON object, which is an unordered set of name/value pairs, the order cannot be guaranteed.

Conclusion

Guys, no doubt there are a number of job offerings for Azure Data Engineers. And the jobs will increase drastically in the upcoming years as every other company is opting for cloud computing. But, how well you prepare for these opportunities is all what matters.

I have divided the latest Azure Data Factory interview questions as per their difficulty level. These ADF interview questions will surely help you to get that extra benefit in an interview over other candidates.