AZURE ADF INTERVIEW QUESTIONS AND ANSWERS

What is Azure Data Factory?

Cloud-based integration service that allows creating data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.Using Azure data factory, you can create and schedule the data-driven workflows(called pipelines) that can ingest data from disparate data stores.It can process and transform the data by using compute services such as HDInsight Hadoop, Spark, Azure Data Lake Analytics, and Azure Machine Learning.

Windows Azure Storage?

It gives four types of storage services:

Queues for informing between web parts and worker roles

Tables for storing structural data

BLOBs (Binary Large Objects) to store contents, records, or vast information

Windows Azure Drives (VHD) to mount a page BLOB. These can be transferred and downloaded by means of BLOBs

What is the integration runtime?

The integration runtime is the compute infrastructure that Azure Data Factory uses to provide the following data integration capabilities across various network environments.

3 Types of integration runtimes:

Azure Integration Run Time: Azure Integration Run Time can copy data between cloud data stores and it can dispatch the activity to a variety of compute services such as Azure HDinsight or SQL server where the transformation takes place

Self Hosted Integration Run Time: Self Hosted Integration Run Time is software with essentially the same code as Azure Integration Run Time. But you install it on an on-premise machine or a virtual machine in a virtual network. A Self Hosted IR can run copy activities between a public cloud data store and a data store in a private network. It can also dispatch transformation activities against compute resources in a private network. We use Self Hosted IR because Data factory will not be able to directly access on-primitive data sources as they sit behind a firewall.It is sometimes possible to establish a direct connection between Azure and on-premises data sources by configuring the firewall in a specific way if we do that we don’t need to use a self-hosted IR.

AZURE ADF INTERVIEW QUESTIONS AND ANSWERS

Azure SSIS Integration Run Time: With SSIS Integration Run Time, you can natively execute SSIS packages in a managed environment. So when we lift and shift the SSIS packages to data factory, we use Azure SSIS Integration Run TIme.

What is the limit on the number of integration runtimes?

There is no hard limit on the number of integration runtime instances you can have in a data factory. There is, however, a limit on the number of VM cores that the integration runtime can use per subscription for SSIS package execution.

What is blob storage in Azure?

Azure Blob Storage is a service for storing large amounts of unstructured object data, such as text or binary data. You can use Blob Storage to expose data publicly to the world or to store application data privately. Common uses of Blob Storage include:

Serving images or documents directly to a browser

Storing files for distributed access

Streaming video and audio

Storing data for backup and restore disaster recovery, and archiving

Storing data for analysis by an on-premises or Azure-hosted service

What are the steps for creating ETL process in Azure Data Factory?

While we are trying to extract some data from Azure SQL server database, if something has to be processed, then it will be processed and is stored in the Data Lake Store.

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

AZURE ADF INTERVIEW QUESTIONS AND ANSWERS

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, it is a data structure that 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 can I schedule a pipeline?

You can use the scheduler trigger or time window trigger to schedule a pipeline.

The trigger uses a wall-clock calendar schedule, which can schedule pipelines periodically or in calendar-based recurrent patterns (for example, on Mondays at 6:00 PM and Thursdays at 9:00 PM).

Can I pass parameters to a pipeline run?

Yes, parameters are a first-class, top-level concept in Data Factory.

You can define parameters at the pipeline level and pass arguments as you execute the pipeline run on demand or by using a trigger.

Can I define default values for the pipeline parameters?

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

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.

Can an activity output property be consumed in another activity?

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

AZURE ADF INTERVIEW QUESTIONS AND ANSWERS

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.

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

Use the Data Factory V2 version to create data flows

Explain the two levels of security in ADLS Gen2?

The two levels of security applicable to ADLS Gen2 were also in effect for ADLS Gen1. Even though this is not new, it is worth calling out the two levels of security because it's a very fundamental piece to getting started with the data lake and it is confusing for many people just getting started.

Role-Based Access Control (RBAC). RBAC includes built-in Azure roles such as reader, contributor, owner or custom roles. Typically, RBAC is assigned for two reasons. One is to specify who can manage the service itself (i.e., update settings and properties for the storage account). Another reason is to permit the use of built-in data explorer tools, which require reader permissions.

Access Control Lists (ACLs). Access control lists specify exactly which data objects a user may read, write, or execute (execute is required to browse the directory structure). ACLs are POSIX-compliant, thus familiar to those with a Unix or Linux background.

What is table storage in Windows Azure?

Windows Azure Table storage service stores a lot of organized information. Windows Azure tables are perfect for putting away organized, non-relational data.

Table: A table is a collection of entities. Tables don't uphold a blueprint on elements, which implies that a solitary table can contain substances that have distinctive arrangements of properties. A record can contain numerous tables.

Entity: An entity is an arrangement of properties, like a database row. An entity can be up to 1 MB in size.

Properties: A property is a name–value pair. Every entity can incorporate up to 252 properties to store data. Every entity likewise has three system properties that determine a segment key, a row key, and a timestamp.

What is Azure Functions?

Azure Functions is a solution for executing small lines of code or functions in the cloud. We can also select the programming languages we want to use. We pay only for the time our code executes; that is,

we pay per usage. It supports a variety of programming languages, like C#, F#, Node.js, Python, PHP or Java. It supports continuous deployment and integration. Azure Functions applications let us develop serverless applications.

What is Azure HdInsight Cluster?

A: Azure HDInsight is a cloud service that makes it easy, fast and cost-effective to process massive amounts of data using open-source frameworks like Hadoop, Spark, Hive, LLAP, Kafka, Storm and R. HDInsight can enable a broad range of scenarios, including ETL, data warehousing, and Machine Learning, to name a few.

What is Azure Data Lake?

Microsoft Azure knowledge Lake may be a extremely ascendible public cloud service that enables developers, scientists, business professionals and other Microsoft customers to gain insight from large, complex data sets. As with most knowledge lake offerings, the service is composed of two parts: data storage and data analytics.

What is SQL Azure database?

SQL Azure database is just an approach to get associated with cloud services where you can store your database into the cloud. Microsoft Azure is the most ideal approach to utilize PaaS where you can have different databases on a similar account.

Microsoft SQL Azure has a similar component of SQL Server, i.e., high accessibility, versatility, and security in the core.

How to stop a running slice?

If you need to stop the pipeline from executing, you can use Suspend-**AzDataFactoryPipeline** cmdlet. Currently, suspending the pipeline does not stop the slice executions that are in progress. Once the in-progress executions finish, no extra slice is picked up.