1. What is Azure data factory?

Azure Data Factory is a cloud-based data integration service that allows you to create data-driven workflows in the cloud for integrate, automating data movement and data transformation. We can make use of Azure Data Factory to create and schedule data-driven workflows that can ingest data from various data stores. It can process and transform the data by using compute services. We can publish the output data to data stores for Business Intelligence (BI) applications to perform visualisation or analytics for better business decisions. By using ADF, we can organise the raw data into meaningful data stores.

1. Difference between ADF V1 and V2?

|  |  |  |
| --- | --- | --- |
| Difference | V1 | V2 |
| Scheduling | All scheduling was done at Data set level. Scheduling is based on pipeline start/end times and dataset availability. | All scheduling Done through Triggers |
| Optimizations | Copying 100 tables of the same type but different names required 100 data sets | This can be done with a single Data set and single for each activity |
| Activity runs | NA | An instance of an activity execution within a pipeline |
| Trigger runs | NA | An instance of a trigger execution |
| Pipeline runs | NA | A single instance of a pipeline execution |
| Expressions | Data Factory V1 allows you to use functions and system variables in data selection queries and activity/dataset properties | In the current version of Data Factory, you can use expressions anywhere in a JSON string value |
| Parameters | NA | Parameters are key-value pairs of read-only configuration settings that are defined in pipelines. You can pass arguments for the parameters when you are manually running the pipeline. If you are using a scheduler trigger, the trigger can pass values for the parameters too. Activities within the pipeline consume the parameter values |
| Linked services | Linked services are much like connection strings, which define the connection information that's necessary for Data Factory to connect to external resources | New **connectVia** property to utilize the Integration Runtime compute environment of the current version of Data Factory |
| Datasets | A named view of data that references the data that you want to use in your activities as inputs and outputs. Availability defines the processing | no need to Availability defines the processing |

1. What is Azure Data lake storage?

The Azure Data Lake Storage service provides a platform for organizations to park – and process and analyse – vast volumes of data in any format

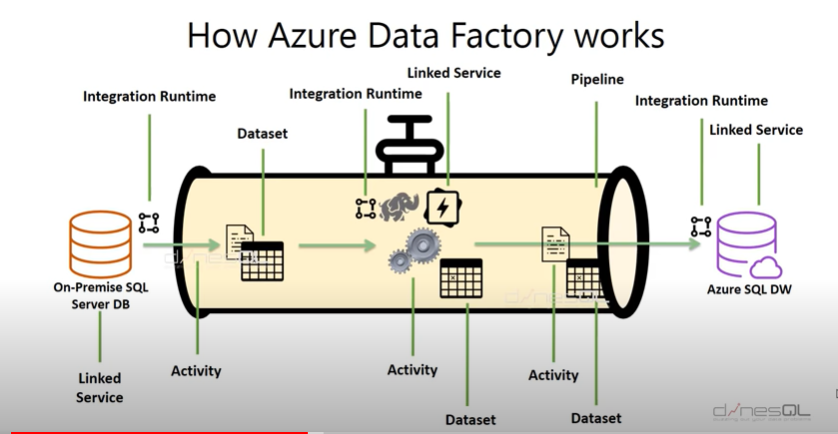
The Data Lake store provides a single repository where organizations upload data of just about infinite volume.

Azure Data Lake is a repository **for large quantities and varieties of both structured and unstructured data** in their native formats.

1. What is Azure Blob Storage?

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.

1. How does Data Azure Factory Work?





1. Key Components in Data Factory

Pipeline Activity Data sets Linked Service Trigger Parameter Variables

1. How the components work Together?

Pipelines Activities Dataflows Data sets Linked Services

Integration Runtimes Triggers

1. What is Azure HDInsight?

Azure HDInsight is a cloud distribution of Hadoop components. Azure HDInsight makes it easy, fast, and cost-effective to process massive amounts of data. You can use the most popular open-source frameworks such as Hadoop, Spark, Hive, LLAP, Kafka, Storm, R, and more.

Azure HDInsight is a cloud-based service from Microsoft for [big data analytics](https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics) that helps organizations process large amounts of streaming or historical data.

1. What is linked Service?

Linked services is like a connection string(s), defines the connection information for Data Factory to connect the external resources

1. What is the Integration Runtime?

Integration Runtime Provides integration capabilities across different network environments.

**It Facilitates**

Data movements

Activity Dispatch

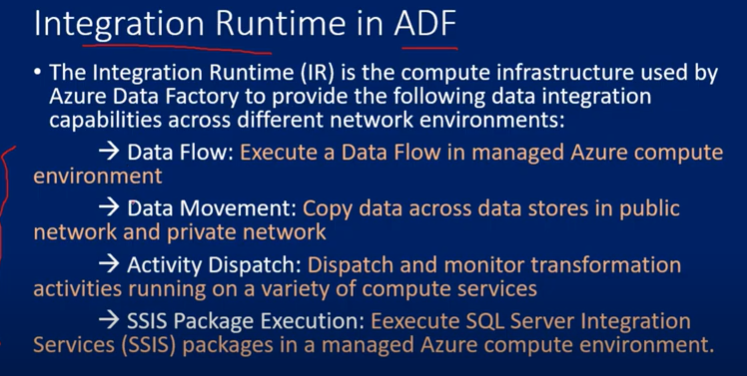
SSIS package execution

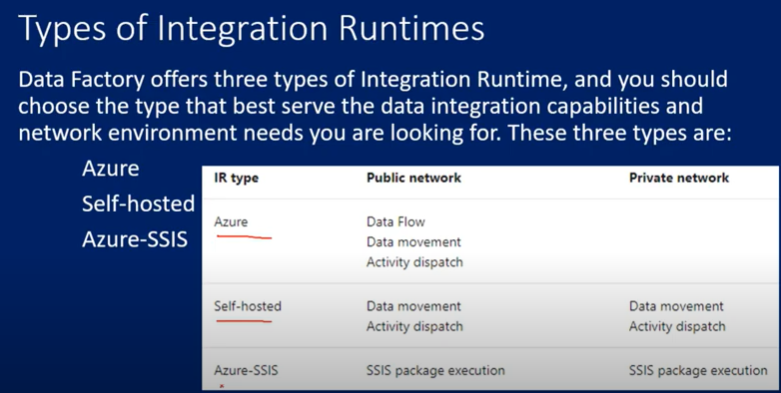
**There are three types of Runtimes. There are**

Azure

Self-Hosted

Azure -SSIS





1. What is the limit on the number of Integration runtimes?

There is no limit based on your requirement you can use.

1. What is a trigger?

Triggers determines when a pipeline execution needs to be kicked off. There are 3 types of Triggers

Triggers are another way that you can execute a pipeline run. Triggers represent a unit of processing that determines when a pipeline execution needs to be kicked off. Currently, Data Factory supports three types of triggers:

**1. Scheduled trigger:** A trigger that invokes a pipeline on a wall-clock schedule

**2. Thumbing Trigger:** Tumbling window triggers are a type of trigger that fires at a periodic time interval from a specified start time, while retaining state. Tumbling windows are a series of fixed-sized, non-overlapping, and contiguous time intervals.

**3. Event based Trigger:** An event-based trigger runs pipelines in response to an event, such as the arrival of a file, or the deletion of a file, in Azure Blob Storage

1. What is control Flow?

Control flow rules decides the rule of navigation from one activity to another activity.  ADF control flow activities allow building complex, iterative processing logic within pipelines.

Control flow is an orchestration of pipeline activities that includes chaining activities in a sequence, branching, defining parameters at the pipeline level, and passing arguments while invoking the pipeline on-demand or from a trigger. It also includes custom-state passing and looping containers, that is, for-each iterators.

1. Can I Pass parameters to a pipeline run?

YES, we can pass parameters to Pipeline.

1. Can i define default values for the pipeline parameters?

YES, we can define default values to the Pipeline parameters.

1. Is the self-hosted integration runtime available for the data flows?

There is no direct option so we need to copy data to staging area first. From staging to target we have a data flows. Directly we can’t copy data from on-premises to cloud directly

1. What is Delta Load Flow?

This is nothing but an incremental data flow. In a data integration solution, incrementally (or delta) loading data after an initial full data load is a widely used scenario.

1. What is connection string in ADF?

Nothing but linked Service

1. What is Data set?

Datasets represent data structures within the data stores, which simply points the data you want to use in your activities as inputs or outputs. Datasets identify data within different data stores, such as tables, files, folders, and documents

1. What is Activity in ADF?

Activity represents a processing step in a pipeline. The activities in a pipeline define actions to perform on your data for example you might use a copy activity to copy data from one data store to another data store. There are three types of activities in ADF.

## Data movement activities:

Data movement activities to move data between supported sources and sink data stores.

## Data transformation activities:

## Data transformation activities to transform data using compute services. A transformation activity executes in a computing environment such as Azure Databricks or Azure HDInsight. It provides links to articles with detailed information on each transformation activity.

## Control flow activities:

## Control flow rules decides the rule of navigation from one activity to another activity.  ADF control flow activities allow building complex, iterative processing logic within pipelines

## 21. What is Pipeline in ADF?

## Pipeline is a logical grouping of activities that performs a unit of work. Together, the activities in a pipeline perform a task

## Pipeline is a collection of activities such as data extraction or processing data. Activities can be operated sequentially or in parallel.

1. Which activity you will use to copy files or Table data from on premise to Azure Data lake/blob Storage or Sql server?

I will use copy activity and self-hosted run time

1. How to call one pipeline in another?

The Execute Pipeline activity allows a Data Factory pipeline to invoke another pipeline.

1. How to execute SSIS package in ADF which activity to use?

Using Azure SSIS runtime engine we can achieve it.

1. What is a lookup Activity?

Lookup Activity reads and returns the content of a configuration file or table. It also returns the result of executing a query or stored procedure. The output from Lookup activity can be used in a subsequent activity.

1. What is Metadata Activity Does in ADF?

You can use the Get Metadata activity to retrieve the metadata of any data in Azure Data Factory. You can use this activity in the following scenarios:

Validate the metadata of any data.

Trigger a pipeline when data is ready/available.

The following functionality is available in the control flow:

You can use the output from the Get Metadata activity in conditional expressions to perform validation.

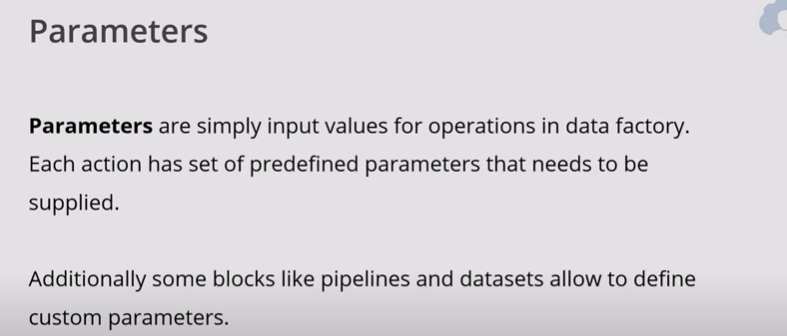
You can trigger a pipeline when a condition is satisfied via Do Until looping.

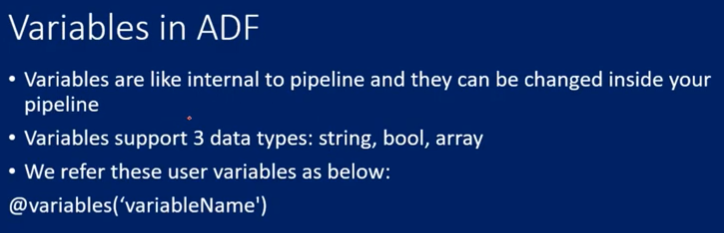
1. How to execute stored procedure in ADF?
2. Add the stored procedure activity to the pipeline, it’s found under “General”.
3. Next, connect the two activities, the stored procedure will only execute following a successful completion of the get metadata activity.
4. What is Set Variable?

Use the Set Variable activity to set the value of an existing variable of type String, Bool, or Array defined in a Data Factory pipeline.

1. Difference b/w parameters and Variables:

**P**arameters are external to the Pipeline where as Variables are internal to the Pipeline hence variable never ask to pass the value while executing the pipeline, We change the values of variable required times by using set variables whereas we can’t change the value of Parameters after passing.





1. What is difference Between Set Variable and Append Variable in ADF?

In Set variable we can use String, Boolean and Array data types whereas in Append Variable we can use only Array data types. Once we execute pipeline if any value in set variable it will remove the existing value and store new value where as in Append variable it will keep existing value and add append new value.

1. What is IF, Switch, until and ForEach in ADF?

If Condition activity provides the same functionality that an if statement provides in programming languages. It executes a set of activities when the condition evaluates to true and another set of activities when the condition evaluates to false.

The Switch activity provides the same functionality that a switch statement provides in programming languages. It evaluates a set of activities corresponding to a case that matches the condition evaluation.

The until activity provides the same functionality that a do-until looping structure provides in programming languages. It executes a set of activities in a loop until the condition associated with the activity evaluates to true. You can specify a timeout value for the until activity in Data Factory.

The ForEach Activity defines a repeating control flow in your pipeline. This activity is used to iterate over a collection and executes specified activities in a loop. The loop implementation of this activity is similar to For each looping structure in programming languages.

1. What is use of WEB and WEBHOOK Activities?

**Web:** Web activity can be used to call a custom REST endpoint from a Data Factory pipeline.

**WebHook:** Using webhook activity, you can call an endpoint and pass a callback URL.

.

1. How to send mail when pipeline fails or Individual Activity fails?

By using Logic Apps

1. What is Logic App in Azure?

Connect your business-critical apps and services with Azure Logic Apps, automating your workflows without writing a single line of code.

Create business processes and workflows visually

Integrate with SaaS and enterprise applications

Unlock value from on-premises and cloud applications

Automate EAI, B2B/EDI, and business processes

Take advantage of the Microsoft Cloud to enhance your integration solutions

1. What is Lift and Shift SSIS package in ADF?

SSIS Integration Runtime (IR)

1. What is the use of Validation Activity in ADF?

You can use a Validation in a pipeline to ensure the pipeline only continues execution once it has validated the attached dataset reference exists, that it meets the specified criteria, or timeout has been reached.

1. What is Databricks in ADF?

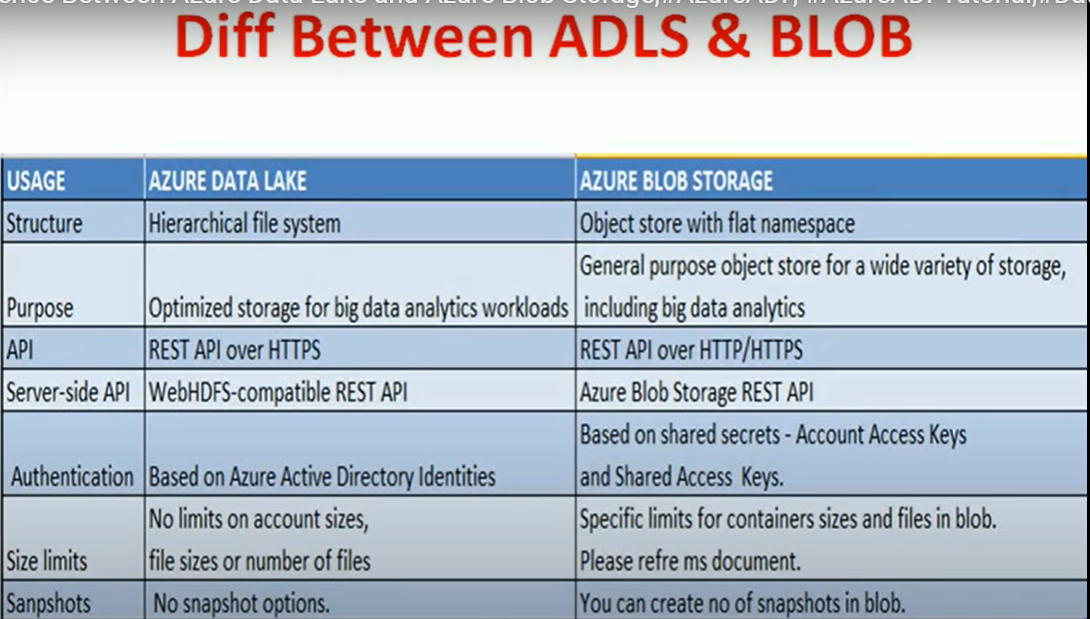
Azure Databricks is an Apache Spark-based analytics platform optimized for the Microsoft Azure cloud services platform.

Databricks is integrated with Azure to provide one-click setup, streamlined workflows, and an interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

Azure Databricks is a fast, easy, and collaborative Apache Spark-based analytics service.

1. What is the difference between Blob Storage and Data Lake Storage?

In Azure Data lake storage we can have empty folders whereas not possible in azure data blob storage.



1. What is the use of ARM templates in ADF?

ARM Templates are a way to declare the objects you want, the types, names and properties in a JSON file which can be checked into source control and managed like any other code file. ARM Templates are what really gives us the ability to roll out Azure “Infrastructure as code”..

1. Name some data flow Transformation in ADF?

Aggregate, Alter Row, Conditional Split, Join, Lookup, New Branch, select, sink, Sort, Surrogate, Union, filter, Exists.

1. What is the difference between Lookup and Joins in Dataflow?

If you want to find rows matching in source 2 based on source 1 input and if you know there will be only one match for every input row, then I would suggest to use Lookup operation. An example would be you Order Details table and you want to find the matching Order Id and Customer Number, then Lookup is a better option.

1. Which Transformations allow to add new column in Dataflow?

Use the derived column transformation to generate new columns in your data flow or to modify existing fields.

1. What is the use of select Transformation in Dataflow activity?

Use the select transformation to rename, drop, or reorder columns. This transformation doesn't alter row data, but chooses which columns are propagated downstream.

In a select transformation, users can specify fixed mappings, use patterns to do rule-based mapping, or enable auto mapping. Fixed and rule-based mappings can both be used within the same select transformation. If a column doesn't match one of the defined mappings, it will be dropped.

1. How to add surrogate key in DIM Table using Dataflow?

Use the surrogate key transformation to add an incrementing key value to each row of data. This is useful when designing dimension tables in a star schema analytical data model. In a star schema, each member in your dimension tables requires a unique key that is a non-business key.

1. What is pivot and Unpivot Transformations in dataflow?

Use Unpivot in ADF mapping data flow as a way to turn a normalized dataset into a more normalized version by expanding values from multiple columns in a single record into multiple records with the same values in a single column.

Use the pivot transformation to create multiple columns from the unique row values of a single column. Pivot is an aggregation transformation where you select group by columns and generate pivot columns using [aggregate functions](https://docs.microsoft.com/en-us/azure/data-factory/data-flow-expression-functions#aggregate-functions).

1. How to achieve incremental Load in ADF?

<http://microsoft-bitools.blogspot.com/2019/05/azure-incremental-load-using-adf-data.html>

1. What is Condition Spilt in Dataflow?

The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.

1. What is Aggregate Transformations in dataflow?

The Aggregate transformation defines aggregations of columns in your data streams. Using the Expression Builder, you can define different types of aggregations such as SUM, MIN, MAX, and COUNT grouped by existing or computed columns.

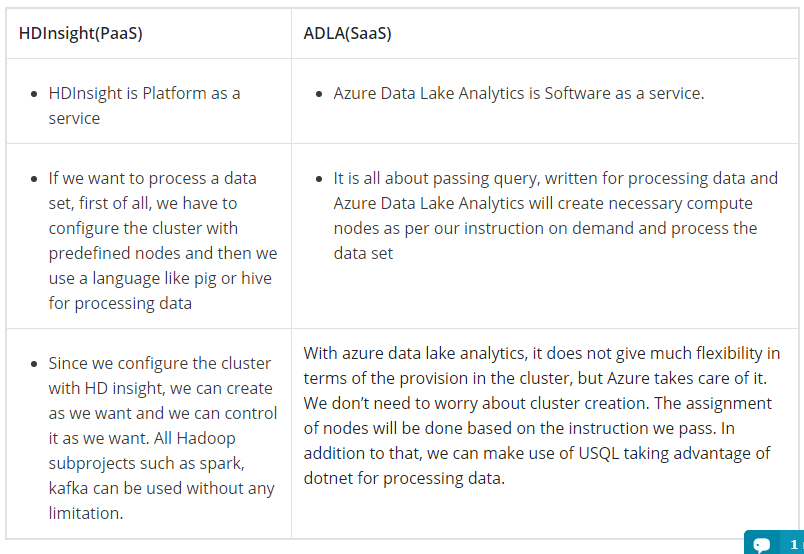
1. What is the use of Exists Transformations?

The exists transformation is a row filtering transformation that checks whether your data exists in another source or stream. The output stream includes all rows in the left stream that either exist or don't exist in the right stream. The exists transformation is similar to SQL WHERE EXISTS and SQL WHERE NOT EXISTS.

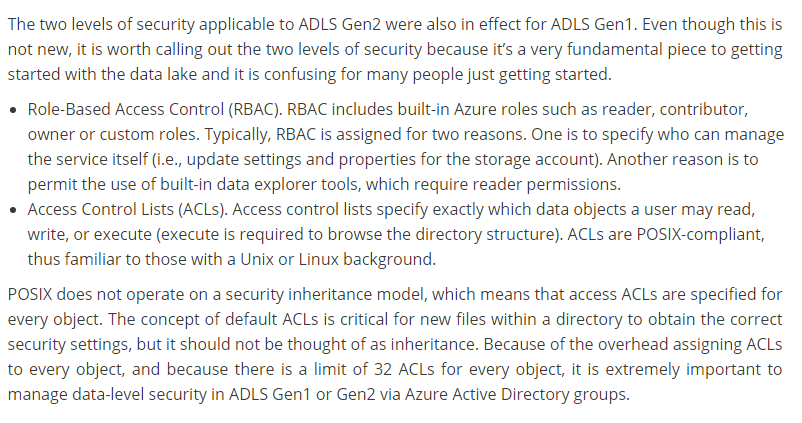
1. What is partition property inside Source and Sink in Dataflow Activity?

It is divided the multiple partitions based on the partition types i.e. Round Robin, Hash and Dynamic etc. It will increase the ADF performance.

1. What is the difference between HDInsight and Data lake storage?



1. Explain the two levels of security in ADLS Gen2?



1. How do i gracefully handle Null values in activity output?

By using Derived column i.e. iifNull function

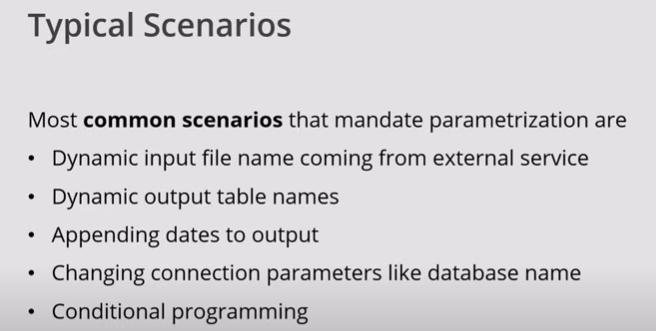
1. Can an activity output property be consumed in another activity?

Yes

**TCS Interview questions:**

1. How to use parameters?

You can provide the **parameter value** to use manually, through triggers, or through the **execute pipeline** activity. Then, that **parameter** can be passed into the **pipeline** and used in an activity. Activities can **pass parameters** into datasets and linked services



1. How to create schema in ADF in Sink data base?

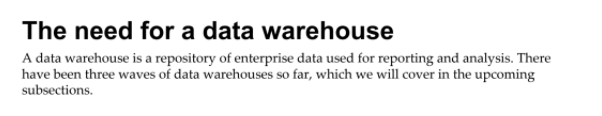
Yes - it takes a bit of configuration, but you can accomplish this with Azure Data Factory Data Flow (ADFDF).

1. Create a DataSet pointing to your CSV location (I'm assuming Azure Blob Storage).
   1. Initially, select a specific CSV file.
   2. On the Schema tab, click "Import schema". It is OK that this will change later, but the DataSet must have a schema at design time.
   3. On the Parameters tab, create a parameter for the blobName.
   4. On the Connection tab, reference that parameter in the "File" box. You will set its value in the pipeline at runtime. [This overrides the initial value used to define the schema].
2. Create a DataSet for the SQLDW table.
   1. Select "Create new table"
   2. Add the schema and table names [this should be configurable/overrideable later via DataSet parameters if needed]
   3. The Schema tab will show no Schema.
3. Create a DataFlow to move the data from CSV to SQLDW.
   1. SOURCE: select the DataSet created in step 1.
      1. On the Source Settings tab: Make sure "Allow schema drift" is checked and "Validate schema" is unchecked [These are the default settings].
      2. CHECK "Infer drifted column types", which is NOT the default.
   2. SINK: select the DataSet created in step 2.
      1. On the Sink tab: Make sure "Allow schema drift" is checked and "Validate schema" is unchecked [These are the default settings].
      2. On the Settings tab, change "Table action" to "Recreate table". This should infer the new schema and drop and create the columns based on what it finds.
      3. On the Mappings tab: make sure "Auto Mapping" is enabled [should be by default]
4. In the Pipeline:
   1. Create a parameter for "blobName"
   2. Select the Data Flow activity:
      1. On the Settings tab: set the source parameter for blobName to the pipeline parameter you just created.
      2. **SQLDW** specific: you will need to provide a Blob Storage Linked Service and location for Polybase.
5. How will you handle streaming data?

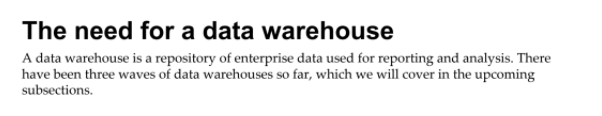
The HDInsight Streaming Activity in a Data Factory [pipeline](https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities) executes Hadoop Streaming programs on [your own](https://docs.microsoft.com/en-us/azure/data-factory/compute-linked-services#azure-hdinsight-linked-service) or [on-demand](https://docs.microsoft.com/en-us/azure/data-factory/compute-linked-services#azure-hdinsight-on-demand-linked-service) HDInsight cluster. This article builds on the [data transformation activities](https://docs.microsoft.com/en-us/azure/data-factory/transform-data) article, which presents a general overview of data transformation and the supported transformation activities.

<https://docs.microsoft.com/en-us/azure/data-factory/transform-data-using-hadoop-streaming>

What is Datawherehouse?







Resource group:

Resource group help us what are the services their under resource group.

