**What is Integration Runtime?**

**Ans)**

An integration runtime, also known as IR, is a managed service that executes data integration pipelines in Azure Data Factory. It provides a scalable and reliable environment for running pipelines, and it can be used to connect to a variety of data sources.

**Types of Integration Runtimes:**

Azure Data Factory provides three types of integration runtimes:

1. **Azure-managed integration runtime:** This is the default integration runtime that is provided by Azure Data Factory. It is a fully managed service that is scalable and reliable. It can be used to connect to a variety of data sources, including Azure Blob Storage, Azure Data Lake Storage, and Azure SQL Database.
2. **Self-hosted integration runtime:** This integration runtime is installed and managed on your own infrastructure. It is a good option if you need to connect to data sources that are not supported by the Azure-managed integration runtime.
3. **Azure Synapse Analytics integration runtime:** This integration runtime is specifically designed for Azure Synapse Analytics. It can be used to connect to Azure Synapse Analytics and run pipelines that use Synapse Analytics features, such as Spark, SQL, and Hadoop.
4. **Pipeline:** A pipeline is a logical grouping of activities that together perform a task.

**Example:** Pipeline has a group of Activities that ingests CSV data from Azure Blob container andcopy to another container in JSON format.

1. **Azure Data Factory**: Azure Data Factory is a cloud-based data integration service provided by Microsoft. It allows you to create, schedule, and orchestrate data pipelines that facilitate the movement and transformation of data from various sources to destinations.
2. **Linked services:** Linked services are much like connection strings, which define the connection information needed for the service to connect to external resources. In simple words Linked service is a connection string between storages.
3. **Dataset**: Dataset is a named view of data that simply points or references the data you want to use in your activities as inputs and outputs. Datasets identify data within different data stores, such as tables, files, folders, and documents.
4. **Trigger:** Triggers represent a unit of processing that determines when a pipeline execution needs to be start.

**Types of Triggers**

**a. Event Trigger:**

* An event trigger starts the data pipeline when a specific event occurs.
* The event can be a file arrival in a storage account, a message in a queue, or other supported event sources.
* Essentially, this trigger listens for events and automatically starts the data processing when it detects the specified event. Think of it as an automated response system.

**b. Schedule Trigger:**

* A schedule trigger allows you to run a pipeline on a pre-defined time schedule. You can set the trigger to run at specific intervals, such as hourly, daily, weekly, or custom time periods. It's like setting an alarm clock for your data workflows.
* When a scheduled trigger runs, it first checks the data that it has already processed. If the trigger finds that there is no new data to process, then it does not run the pipeline. If the trigger finds that there is new data to process, then it runs the pipeline.
* **Important :-** This type of Trigger can deal the file current or recent files but not very old/past file nor far future files.

**c. Tumbling window trigger:**

* A trigger that operates on a periodic interval, while also retaining state. This means that the trigger will remember the data that it has processed, and **it will only process new data** that has been added since the last time the trigger ran. For example, you could create a tumbling window trigger that runs a pipeline every hour, and that only processes data that has been added to the source data in the last hour.
* **Important :- One of the advantage of Tumbling window Trigger is that – It can execute the past date files as well**.
* For example : files which are dumped in a container on date Jan-23-2023 it can run those files as well even though it’s been long time from Jan-2023 till date August-2023.

**Difference between Tumbling, Schedule and Event Trigger:**

1. **Tumbling trigger:**

* A tumbling trigger is a type of trigger that fires at a specific interval, such as every hour, day, or week. The trigger fires once for each interval, regardless of whether any data has changed.
* Can trigger files of - **Past**, **Current** and F**uture** **time**.
* Can create Dependencies to run the triggers
* **To execute :** enter this in pipeline parameter :- **@trigger().outputs.windowStartTime**
* **Max concurrency –** This means it will execute in parallel. (If we give value as 50 – then 50 windows will create and execute parallelly.

**2) Scheduled trigger:**

* A scheduled trigger is a type of trigger that fires at a specific time, such as 10:00 AM every day. The trigger fires once at the specified time, regardless of whether any data has changed.
* Can trigger only files of **Current** **time** and F**uture** **time**.
* There are no Dependencies to create in Scheduled triggers

**3) Event based trigger:**

* It is a type of trigger that is used to start a pipeline in response to an event.
* This event can be a creation of file, updating a database table, or any other event that can be detected by Azure Data Factory.
* To execute : enter this in pipeline parameter :- **@triggerBody().fileName**

**Triggers in Azure Data Factory**

* Triggers- You can execute your pipeline.
* It determines when a pipeline execution needs to be kicked off.
* Pipeline and triggers have a many-to-many relationship (Except for the tumbling window trigger)
* Multiple triggers can kick off a single pipeline or single trigger can kickoff multiple pipelines.

**Types of Triggers in ADF:**

1. **Schedule triggers:** A trigger that invokes a pipeline on a wall clock schedule.
2. **Tumbling window trigger:** A trigger that operates on a periodic interval while also retaining state.
3. **Event based triggers:** A trigger that responds to an event
4. **Schedule triggers**

**Create Pipeline**

**Copy Data Activity**

1. In copy data activity select the source file and sink file

**Add Trigger (New add/edit)**

1. Select a trigger if you have or else create new.
2. If creating/selecting new, set the configurations
3. Now click on okay

**Debug.**

# **2. Use Event triggers to copy files coming in different times**

**Create Pipeline**

**Copy Data Activity**

1. In copy data activity select the source file and sink file
2. Create a parameter inside dataset and then after point the parameter as @dataset().Filename

**Triggers ---> Pipelines ---> Dataset**

**Flow is like Triggers gives parameters to Pipelines and Pipelines to Dataset.**

1. Create parameters in the pipeline and then after point the parameter as @pipeline().parameters.Filename

**Add Trigger (New add/edit)**

1. Select a trigger if you have or else create new.
2. If creating/selecting new, set the configurations
3. Now click on okay
4. After clicking okay, you will get the parameter value box which need to be set to @triggerBody().filename

**Debug and Publish.**

# **3. Implement tumbling triggers in ADF**

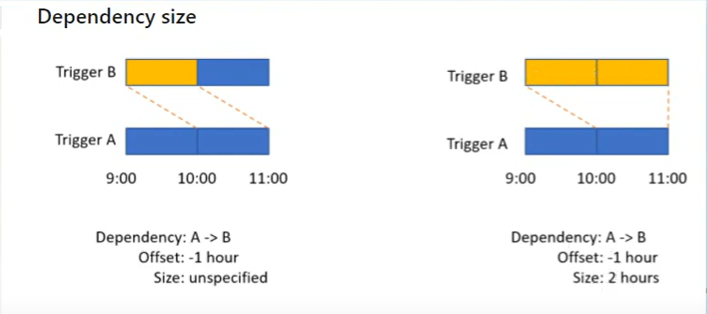
**Create Pipeline**

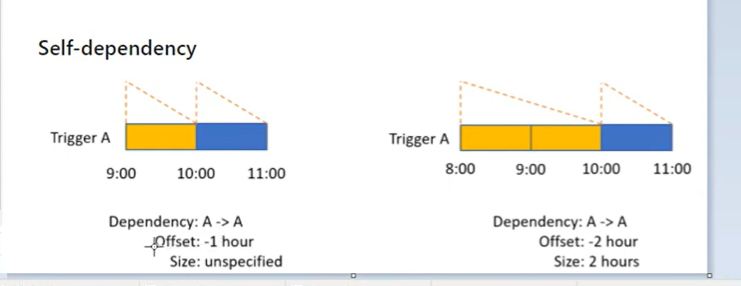
**Copy Data Activity**

1. In copy data activity select the source file and sink file
2. Create parameters in the pipeline as “starttimestamp” and then after point the parameter as @trigger().outputs.windowStartTime
3. In source query section add @concat(‘select \* from t\_salesdata where modified\_ts>=convert(datetime,’’’ ,pipeline().parameters.starttimestamp,’’’)’)

**Add Trigger (New add/edit)**

1. Select a trigger if you have or else create new.
2. If creating/selecting new, set the configurations
3. Now click on okay





**General Activities**

1. **Append Variable Activity:**

This activity assigns a value to the array variable. For example, you could use the Append Variable Activity to add a new item to a list of products.

1. **Delete Activity:**

This activity deletes a file or folder from a storage account.

**Eg:** You can use the Delete Activity to delete a file that is no longer needed. For example, you could use the Delete Activity to delete a log file after it has been processed.

1. **Execute Pipeline Activity:**

This activity allows you to call Azure Data Factory pipelines.

**Eg:** you could use the Execute Pipeline Activity to call a pipeline that cleans and transforms your data.

1. **Execute SSIS Package Activity:** This activity executes an SSIS package in a managed environment. You can use the Execute SSIS Package Activity to execute an SSIS package that performs complex data transformations.

**Eg:** you could use the Execute SSIS Package Activity to load data from a flat file into a database.

1. **Fail Activity:** This activity stops the pipeline and logs an error message.

**Eg:** you could use the Fail Activity to stop the pipeline if a file cannot be found.

1. **Get Metadata Activity:**

* It is used to get metadata of files/folders.
* It retrieves metadata information about a specified data source or file, such as size, schema, or last modified date.

**Eg:** you could use the Get Metadata Activity to get the size of a file.

1. **Lookup Activity:**

* This activity retrieves data from a dataset.
* The Lookup Activity can read data from a variety of sources, including Azure Storage, Azure SQL Database, and Azure Cosmos DB. The Lookup Activity can also read data from a file system or a database table.

**Eg:** you could use the Lookup Activity to retrieve a list of customers from a database.

1. **Stored Procedure Activity:**

This activity executes a stored procedure that performs database operations.

**Eg:** you could use the Stored Procedure Activity to update a customer record in a database.

1. **Script Activity:**

This activity executes a script in a supported language such as Python or JavaScript.

**Eg:** you could use the Script Activity to perform custom data processing.

1. **Set Variable Activity:**

This activity assigns a value to the variable.

**Eg:** you could use the Set Variable Activity to store the output of a calculation.

1. **Validation Activity:**

This activity validates the input and output of an activity.

**Eg:** you could use the Validation Activity to check that the input data is valid before it is processed.

1. **Web Activity:**

This activity makes a web request to an external endpoint.

**Eg:** you could use the Web Activity to send an email when a pipeline completes.

1. **Webhook Activity:**

This activity triggers a webhook when a pipeline completes.

**Eg:** you could use the Webhook Activity to trigger a notification in a chat application when a pipeline completes.

1. **Wait Activity:**

This activity waits for a specified amount of time before continuing.

For example, you could use the Wait Activity to wait for a database to be ready before continuing with a pipeline.

**Eg:** Wait('30 seconds')

**Iteration & conditionals**

1. **Filter Activity:**

This activity allows you to apply different filters on your input dataset.

**Eg:** you could use the Filter Activity to filter out all products that are not in stock.

1. **For Each Activity:**

* This activity provides the functionality of **for each loop** that executes for multiple iterations.
* The ForEach activity allows you to loop over an array or a collection of items and execute a set of activities for each item in the collection.
* It's commonly used for batch processing or iterating over a list of files, tables, or other data elements.

**Eg:** you could use the For Each Activity to iterate over a list of products and calculate their total price.

1. **If Condition Activity**:

It provides the same functionality as If statement, it executes the set of expressions based on if the condition evaluates to true or false.

**Eg:** you could use the If Condition Activity to decide whether to process a file based on its size.

1. **Switch Activity**:

* The Switch Activity is a control flow activity in Azure Data Factory.
* Switch Activity is used to execute a different set of activities based on the value of an expression.
* The Switch Activity is like a traffic light. It checks the value of an expression and then "routes" the pipeline to different activities depending on the value of the expression.

**Eg:** The Switch Activity can be used to route a pipeline to different activities depending on the file extension. For example, if the file has a **.csv** extension, the Switch Activity will route the pipeline to an activity that loads the data into a CSV format. If the file has a .txt extension, the Switch Activity will route the pipeline to an activity that loads the data into a text format. And if the file has a .json extension, the Switch Activity will route the pipeline to an activity that loads the data into a JSON format.

1. **The Until Activity:**

* This activity is used to execute a set of activities in a loop until a condition is met.
* The Until Activity has an expression that is evaluated after each iteration of the loop.
* If the expression evaluates to true, the Until Activity will execute the activities in the loop again. If the expression evaluates to false, the Until Activity will stop executing the loop.

**Move and Transform**

1. **Copy Activity:**

* This activity is used to move data between different data stores.
* It copies data from a source to a destination.
* It supports various sources and sinks, such as Azure Blob Storage, Azure Data Lake Storage, SQL databases, and more.

1. **Data Flow Activity:**

The Data Flow Activity in Azure Data Factory is used to transform data in real time such as :

Filtering data, Aggregation, Joining data, Splitting data, Formatting data.

1. **Filtering data:** This is the process of removing rows or columns from a dataset that do not meet certain criteria. For example, you could filter a dataset to only include rows where the customer's age is greater than 18.
2. **Aggregation:** This is the process of summarizing data by grouping it together and calculating a total, average, or other statistic.

**Eg:** you could aggregate a dataset to calculate the total number of customers, the average order value, or the most popular product.

1. **Joining data:** This is the process of combining two or more datasets based on a common field.

**Eg:** you could join a customer dataset with an order dataset to get information about the customers who have placed orders.

1. **Splitting data:** This is the process of dividing a dataset into two or more smaller datasets based on a common field.

**Eg:** you could split a dataset of customers into two groups: customers who have placed an order in the past year and customers who have not placed an order in the past year.

1. **Formatting data:** This is the process of changing the format of data to make it easier to read or use.

**Eg:** you could format a date field to display the date in a specific format, such as "MM/DD/YYYY".

**Azure Function**

**Azure Function Activity:**

Azure Functions are serverless functions that can be triggered by events, such as HTTP requests or file changes. The Azure Function Activity can be used to perform a variety of tasks, such as:

* Processing data
* Generating reports
* Sending emails
* Integrating with other systems

**Azure Data Explorer**

**Azure Data Explorer command:**

The Azure Data Explorer Activity is used to execute queries against an Azure Data Explorer cluster.

The Azure Data Explorer Activity is used to execute queries against an Azure Data Explorer cluster.

Azure Data Explorer is a fully managed, real-time analytics service that can be used to analyze streaming and historical data.

The Azure Data Explorer Activity can be used to perform a variety of tasks, such as:

* Querying data from Azure Data Explorer
* Creating and managing Azure Data Explorer tables
* Loading data into Azure Data Explorer
* Deleting data from Azure Data Explorer

**HD Insight**

**1. Hive Activity:**

The Hive Activity in Azure Data Factory is used to run Hive queries on an Azure HDInsight cluster. Hive is a data warehouse infrastructure that uses Hadoop to process and store large datasets. The Hive Activity can be used to perform a variety of data processing tasks, such as:

* ETL (extract, transform, and load)
* Data warehousing
* Data analysis

**2. MapReduce Activity:**

The MapReduce Activity is used to run MapReduce jobs on an Azure HDInsight cluster. MapReduce is a programming model and an associated implementation for processing and generating large data sets.

The MapReduce Activity can be used to perform a variety of data processing tasks, such as:

* Sorting
* Joining
* Summarizing

**3. Pig Activity:** Pig Activity is used to run Pig scripts on an Azure HDInsight cluster. Pig is a high-level language for data analysis that uses Pig Latin to process large datasets. The Pig Activity can be used to perform a variety of data processing tasks, such as:

* Cleaning data
* Combining data
* Analyzing data

**4. Spark Activity:** Spark Activity is used to run Spark jobs on an Azure HDInsight cluster. Spark is a unified analytics engine for large-scale data processing. The Spark Activity can be used to perform a variety of data processing tasks, such as:

* Machine learning
* Real-time analytics
* Streaming analytics

**5. Streaming Activity:** Streaming Activity is used to process streaming data on an Azure HDInsight cluster. Streaming data is data that is continuously generated and processed. The Streaming Activity can be used to perform a variety of data processing tasks, such as:

* Monitoring data
* Detecting anomalies
* Responding to events

**Databricks**

1. **Notebook Activity:**  NotebookActivity is used to run notebooks on a Databricks cluster.

Notebooks are documents that contain code, text, and visualizations. The Notebook Activity can be used to perform a variety of data processing tasks, such as:

* ETL (extract, transform, and load)
* Data warehousing
* Data analysis

1. **Jar Activity:** The Jar Activity is used to run Java jobs on a Databricks cluster.

Jar files are Java archive files that contain Java classes and resources. The Jar Activity can be used to perform a variety of data processing tasks, such as:

* Machine learning
* Real-time analytics
* Streaming analytics

1. **Python** **Activity:** The Python Activity is used to run Python scripts on a Databricks cluster. Python is a general-purpose programming language that is often used for data science and machine learning. The Python Activity can be used to perform a variety of data processing tasks, such as:

* Cleaning data
* Combining data
* Analyzing data

**Synapse**

1. **Notebook** **Activity:** This Activity is used to run notebooks on a Synapse Analytics workspace. Notebooks are documents that contain code, text, and visualizations. The Notebook Activity can be used to perform a variety of data processing tasks, such as:

* ETL (extract, transform, and load)
* Data warehousing
* Data analysis

1. **Spark Job Definition** **Activity:** This Activity is used to run Spark jobs on a Synapse Analytics workspace. Spark is a unified analytics engine for large-scale data processing. The Spark Job Definition Activity can be used to perform a variety of data processing tasks, such as:

* Machine learning
* Real-time analytics
* Streaming analytics

**Batch Service**

**Custom:** The "custom" option under Batch Service in Azure Data Factory allows you to use your own custom batch service to run your pipelines. This can be useful if you have a specific need that is not met by the default batch services that are provided by Azure Data Factory.

**Power Query**

**Power Query Activity:** It can be used to perform a variety of data transformations.

It can be used to:

* Clean data
* Combine data from multiple sources
* Create derived columns
* Apply business logic to data
* Prepare data for analysis

The Power Query activity is a great way to automate data transformation tasks. It can be used to perform complex transformations that would be difficult or time-consuming to do manually.