Day 2: Azure Data Factory Basics

1. What are the key components of Azure Data Factory?

- **Answer:** The main components of Azure Data Factory are:
  - **Pipelines:** Groups of activities that perform a unit of work.
  - **Activities:** Tasks performed by the pipeline, such as data movement or transformation.
  - **Datasets:** Represents the data structures within the data stores that the activities work with.
  - **Linked Services:** Defines the connection information needed for Data Factory to connect to external data sources.
  - **Triggers:** Units of processing that determine when a pipeline execution should be kicked off.

2. How do you create a pipeline in Azure Data Factory?

- **Answer:** To create a pipeline in Azure Data Factory:
  1. Open the Azure portal and navigate to your Data Factory.
  2. In the Data Factory UI, go to the "Author & Monitor" section.
  3. Click on the "Create pipeline" button.
  4. Add activities to the pipeline by dragging and dropping them from the Activities pane.
  5. Configure the activities as needed.
  6. Save and publish the pipeline.

3. What is the purpose of Linked Services in Azure Data Factory?

- **Answer:** Linked Services in Azure Data Factory act as connection strings, defining the connection information needed for Data Factory to connect to external data sources. They are used to specify the credentials and connection details required to access different types of data stores, such as Azure Blob Storage, Azure SQL Database, and others.

4. What types of data stores can Azure Data Factory connect to?

- **Answer:** Azure Data Factory can connect to a wide range of data stores, including:
  - Azure services (e.g., Azure Blob Storage, Azure SQL Database, Azure Data Lake Storage)
  - On-premises data stores (e.g., SQL Server, Oracle, File System)
  - Cloud-based data stores (e.g., Amazon S3, Google Cloud Storage)
  - SaaS applications (e.g., Salesforce, Dynamics 365)

5. What is the Copy Activity in Azure Data Factory, and how is it used?

- **Answer:** The Copy Activity in Azure Data Factory is used to copy data from a source data store to a destination data store. It is commonly used in ETL operations. To use the Copy Activity:
    1. Define the source and destination datasets.
    2. Configure the source and destination properties in the Copy Activity.
    3. Specify any additional settings such as data mapping, logging, and error handling.
    4. Add the Copy Activity to a pipeline and run it.

6. Explain the concept of Integration Runtime (IR) in Azure Data Factory.

- **Answer:** Integration Runtime (IR) is the compute infrastructure used by Azure Data Factory to provide data integration capabilities across different network environments. There are three types of IR:
    - **Azure IR:** Used for data movement and transformation within Azure.
    - **Self-hosted IR:** Installed on an on-premises machine or a virtual machine in a virtual network to connect to on-premises data sources.
    - **Azure-SSIS IR:** Used for running SQL Server Integration Services (SSIS) packages in the cloud.

7. How do you implement an incremental data load in Azure Data Factory?

- **Answer:** To implement an incremental data load in Azure Data Factory:
    1. Identify the column that will be used to track changes (e.g., a timestamp or ID column).
    2. Store the last loaded value of this column in a control table or variable.
    3. In the pipeline, use the stored value to filter the source data for new or updated records.
    4. Load the incremental data into the destination data store.
    5. Update the stored value to reflect the latest loaded record.

8. How can you handle data transformation in Azure Data Factory?

- **Answer:** Data transformation in Azure Data Factory can be handled using:
    - **Mapping Data Flows:** Visual interface for designing data transformations.
    - **Data Flow Activities:** Perform transformations using SQL, Spark, or custom scripts.
    - **External Services:** Integrate with Azure Databricks or HDInsight for complex transformations.

9. What are Tumbling Window Triggers in Azure Data Factory?

- **Answer:** Tumbling Window Triggers are a type of trigger in Azure Data Factory that fire at periodic intervals. They are useful for processing data in fixed-size, non-overlapping time windows. Each trigger instance is independent, and the trigger will only execute if the previous instance has completed.

10. How do you monitor and troubleshoot pipeline failures in Azure Data Factory?

- **Answer:** Monitoring and troubleshooting pipeline failures in Azure Data Factory can be done using:
    - **Azure Monitor:** Provides a comprehensive view of pipeline runs, including success and failure metrics.
    - **Activity Runs:** Reviewing the details of individual activity runs to identify the root cause of failures.
    - **Logs and Alerts:** Configuring logging to capture detailed execution logs and setting up alerts to notify of failures.
    - **Retry Policies:** Implementing retry policies for transient failures.
    - **Debugging Tools:** Using the debug mode in the Data Factory UI to test and troubleshoot pipelines before deployment.