**1. What is the ETL process?**

**Answer**:  
**"ETL stands for Extract, Transform, and Load. It’s the process of moving data from one place to another. First, we Extract data from a source like a database. Then, we Transform the data by cleaning or changing it into the right format. Finally, we Load the data into a destination, like a data warehouse, for analysis."**

**2. What are some common issues in ETL and how do you troubleshoot them?**

**Answer**:  
**"In ETL, common issues include things like primary key violations, data type mismatches, and missing values (nulls). To solve these problems, I use a systematic approach. For example, if I get an error, I backtrack through the ETL process to find out where the issue started. I check the data formats, fix any mismatches, and ensure there are no duplicate primary keys or missing values."**

**3. Can you explain how you resolved data quality issues?**

**Answer**:  
**"In data quality, I’ve handled issues like data inconsistencies, duplicates, and null values. For example, if two records are the same but look slightly different, I remove duplicates. If there are missing values, I either fill them with a default value or remove those rows, depending on the situation. I use tools like SSIS and Azure Data Factory to automate this process and ensure the data is clean before it’s loaded into the data warehouse."**

**4. What is a Data Warehouse, and why is it important?**

**Answer**:  
**"A Data Warehouse is a system where data from different sources is stored in one place. It helps businesses organize and analyze large amounts of data to make decisions. It’s important because it makes it easier to run reports and do analysis without slowing down other systems."**

**5. What is the difference between SCD Type 1 and Type 2?**

**Answer**:  
\*\*"SCD stands for Slowly Changing Dimensions, which refers to how we handle changes in data over time.

* **Type 1**: When data changes, we simply update the old value with the new one (no history is kept).
* **Type 2**: We keep both the old and new values, so we have a history of the changes."\*\*

**6. Can you explain how you use SSIS or Azure Data Factory in your work?**

**Answer**:  
**"I use SSIS and Azure Data Factory to automate ETL processes. SSIS helps me move and transform data from different sources into a data warehouse. I use Azure Data Factory for cloud-based data integration, where I build pipelines to move, clean, and load data into Azure-based systems. Both tools help make sure the data is clean, accurate, and ready for analysis."**

**7. How do you handle null values in data?**

**Answer**:  
**"Null values are empty or missing data. I handle them by either filling in a default value, using the average or median value (for numerical data), or removing rows that have too many missing values. It depends on what’s best for the project and how much data is missing."**

**8. What are some common challenges you face when working with data pipelines?**

**Answer**:  
**"Some challenges include handling data format issues, managing large volumes of data, and ensuring data quality. For example, I may run into problems with inconsistent data formats, like different date styles, or duplicate records. I make sure to identify these issues early and fix them by applying data transformation rules or using data cleaning tools."**

**9. How do you ensure data integrity during the ETL process?**

**Answer**:  
**"I ensure data integrity by validating the data before loading it into the system. For example, I check that data types match, there are no duplicates, and that the primary keys are unique. I also use checks and balances, like error handling, to make sure no incorrect data gets through."**

**10. Can you explain how you’ve used SCD Type 2 in a project?**

**Answer**:  
**"In a project, I used SCD Type 2 to track changes in customer data over time. When a customer’s information changed, I kept the old data and added a new record with the updated information. This way, we could always see the history of changes for each customer, which is important for tracking customer behavior and creating reports."**

**Azure Data Factory (ADF)**

1. **What is Azure Data Factory?**  
   **Answer**:  
   **"Azure Data Factory is a cloud-based tool for building data pipelines. It helps move data from different sources to destinations, clean and transform it along the way, and automate the process. It's great for handling both batch and real-time data."**
2. **What are the key components of Azure Data Factory?**  
   **Answer**:  
   \*\*"The main components are:

* **Pipelines**: The workflow that moves and transforms data.
* **Activities**: The steps in the pipeline, like copying or transforming data.
* **Datasets**: Definitions of data structures from sources and destinations.
* **Linked Services**: Connections to data sources or compute environments."\*\*

1. **How does Azure Data Factory handle data transformation?**  
   **Answer**:  
   **"ADF can use data flows for transformations. For advanced processing, it integrates with other tools like Databricks or SQL Server for custom transformations."**

**Azure Data Lake**

1. **What is Azure Data Lake?**  
   **Answer**:  
   **"Azure Data Lake is a storage solution for big data. It stores structured, semi-structured, and unstructured data, like logs, images, or videos. It is designed for high scalability and supports analytics tools like Databricks or Synapse."**
2. **Why would you use Azure Data Lake?**  
   **Answer**:  
   **"You would use it to store large amounts of data at a low cost and to process that data for analysis. It works well with tools for big data, like Databricks and Hadoop."**
3. **What’s the difference between Azure Blob Storage and Azure Data Lake?**  
   **Answer**:  
   **"Azure Blob Storage is for general-purpose object storage, while Azure Data Lake is optimized for big data analytics and supports hierarchical file structures, making it better for complex queries."**

**Azure Databricks**

1. **What is Azure Databricks?**  
   **Answer**:  
   **"Azure Databricks is a data analytics platform based on Apache Spark. It’s used for big data processing, machine learning, and data science. It integrates well with other Azure services like Data Lake and Synapse."**
2. **How do Azure Databricks and Azure Data Factory work together?**  
   **Answer**:  
   **"Azure Data Factory is used to schedule and manage data pipelines, while Databricks is used for advanced data transformations or machine learning within those pipelines. For example, ADF can trigger a Databricks notebook to process data from a data lake."**
3. **What are the main features of Azure Databricks?**  
   **Answer**:  
   \*\*"Some key features are:

* **Big data processing**: Handles large datasets with Apache Spark.
* **Collaborative workspace**: Multiple users can work together on notebooks.
* **Machine learning**: Supports libraries for building models.
* **Integration**: Works well with Azure tools like Data Lake and Blob Storage."\*\*

**Scenarios**

1. **How would you use Azure Data Lake and Databricks together?**  
   **Answer**:  
   **"I would store raw data in Azure Data Lake and then use Azure Databricks to clean, process, and analyze the data. For example, I could use Databricks to filter logs in Data Lake and prepare them for machine learning."**
2. **How does Azure Data Factory integrate with Azure Data Lake?**  
   **Answer**:  
   **"Azure Data Factory can extract data from Azure Data Lake, transform it using Data Flows or Databricks, and then load the processed data into another destination, like a database or a report system."**
3. **What is the role of Spark in Azure Databricks?**  
   **Answer**:  
   **"Spark is the core engine in Databricks. It’s used for processing large datasets quickly, running queries, and performing transformations. It also supports machine learning and streaming data."**

**General Questions**

1. **What are the advantages of using Azure for data engineering?**  
   **Answer**:  
   **"Azure offers scalable and flexible tools like Data Factory, Data Lake, and Databricks. These tools integrate well with each other, making it easy to handle big data workflows, store data, and analyze it using machine learning or reporting tools."**
2. **What are triggers in Azure Data Factory?**  
   **Answer**:  
   **"Triggers are used to start pipelines automatically in ADF. For example, you can run a pipeline on a schedule, like daily, or when a file is added to storage."**
3. **How does Azure Databricks handle real-time data?**  
   **Answer**:  
   **"Databricks uses Spark Streaming to process real-time data streams. For example, you can process live data from IoT devices or event logs."**