Wipro
Azure Data Engineer Interview Questions 2025

1. Databricks & PySpark

Q) How to flatten nested JSON in Spark?

Answer:

Use from_json and explode or selectExpr with dot notation:

Example:

```
from pyspark.sql.functions import explode, col
df = spark.read.json("file.json")
flattened_df = df.selectExpr("id", "nested.field1 as field1", "nested.field2 as field2")
```

Q) How to handle missing and null values in PySpark?

Answer:

Drop nulls: df.dropna()
Fill nulls: df.fillna({'col1': 'default', 'col2': 0})
Filter nulls: df.filter(df.col1.isNotNull())

Q) Write a UDF to return only the first name from full name.

Answer:

```
Def first_name(fullname):
   return fullname.split()[0]
first_name_udf = udf(first_name, StringType())
df = df.withColumn("first_name", first_name_udf(df.full_name))
```

1. **Compute 7-day rolling average for stock closing prices**.

Answer:

```
from pyspark.sql.window import Window
from pyspark.sql.functions import avg

window = Window.partitionBy("company").orderBy("date").rowsBetween(-6, 0)
df = df.withColumn("7_day_avg", avg("closing_price").over(window))
```

Q) Difference between repartition and coalesce.

Answer:

Repartition: Shuffles data across the cluster, can increase or decrease partitions, ensures even data distribution, slower due to shuffle.

Coalesce: Reduces partitions without full shuffle, faster, used mainly to decrease partitions efficiently.

Q) Explain Spark architecture and stages, jobs, tasks.

Answer:

Spark has a Driver that coordinates work and Executors that run tasks. A Job is triggered by an action, divided into Stages (tasks without shuffle), and each stage consists of Tasks (one per data partition).

Q) How many Spark jobs are created when using inferSchema vs manual schema?

Answer:

Using inferSchema creates **2** jobs (schema inference + data read), whereas a manual schema creates 1 job (direct read), so manual is faster.

Q) How to optimize queries in PySpark/Spark SQL?

Answer:

Optimize by caching repeated data, reducing shuffles, using broadcast joins for small tables, partitioning/bucketing data, and avoiding wide transformations when possible.

Q) What is AQE (Adaptive Query Execution)?

Answer:

AQE dynamically optimizes query execution at runtime by merging shuffle partitions, changing join strategies, and improving overall performance.

Q) Explain Liquid clustering.

Answer:

Liquid clustering organizes Delta Lake data based on **Z-order columns** to **speed up selective queries** and improve read performance

Q) Difference between Databricks and Synapse.

Answer:

Databricks is Spark-based for ETL, streaming, and advanced analytics. Synapse is SQL-based for data warehousing and BI, better for large-scale SQL queries and integration with Power BI.

Q) Difference between External table and Managed table in Synapse.

Answer:

Managed Table: Synapse manages data and metadata; dropping table deletes both.
External Table: Only metadata is in Synapse; data stays in storage.

Q) Types of tables in Synapse (SCD Type 1 vs Type 2).

Answer:

SCD Type 1: Overwrites old data, no history.
SCD Type 2: Keeps history by adding new rows for changes.

Q) Where do you apply quality checks for CSV files?

Answer:

At the landing/ingestion layer before loading into Bronze/Silver layer: check schema, nulls, duplicates, and data types

Q) How do you handle error records in a pipeline?

Answer:

Redirect errors to error/staging tables or DLQ (Dead Letter Queue), log for analysis, and optionally alert teams.

Q) Difference between ADLS Gen1 and Gen2.

Answer:

Gen2 supports hierarchical namespace, better performance, security, and cost; Gen1 is legacy.

Q) How to set up disaster recovery for ADF / Databricks.

Answer:

Use geo-redundant storage, backup pipelines/notebooks, multi-region deployment, and automated failover.

Q) How to configure Unity Catalog in Azure Databricks.

Answer:

Create Premium workspace, assign Access Connector to storage, create metastore, configure schemas/tables, and assign role-based access

Q) Difference between manager vs lead access.

Answer:

Manager usually has full access to all data, while Lead or others have restricted access based on columns, tables, or schemas.

Q) How to set schema-level access for different roles.

Answer:

Use GRANT statements in Unity Catalog for role-based privileges at schema or table level.

Q) Where can you get the access token in Databricks?

Answer:

In User Settings → Access Tokens tab (if available) or via Entra ID / PAT for API authentication.

Q) Can Unity Catalog use multiple metastores?

Answer:

Yes, Unity Catalog supports multiple metastores, one per workspace or environment.

Q) Write SQL query to find duplicate rows.

Answer:

```
SELECT col1, col2, COUNT(*)
FROM table_name
GROUP BY col1, col2
HAVING COUNT(*) > 1;
```

Q) Find the third highest value without using analytical functions.

Answer:

```
SELECT MAX(salary)
FROM table_name
```

WHERE salary < (SELECT MAX(salary) FROM table_name WHERE salary < (SELECT MAX(salary) FROM table_name));

Q) Top 10 active users query in PySpark.

Answer:

```
from pyspark.sql.functions import count
df.groupBy("user_id").agg(count("*").alias("activity_count"))\
 .orderBy("activity_count", ascending=False).limit(10)
```

Q) Difference between DataFrame SQL and T-SQL.

Answer:

DataFrame SQL follows Spark SQL syntax, optimized for distributed computing; T-SQL is SQL Server syntax, optimized for relational DBs.

Q) How Z-ordering works in Delta tables

Answer:

Z-ordering clusters data based on column values to improve read performance for selective queries by reducing scan volume.