Seekho Bigdata
Data is the New Oil

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 13 — Delta Lake, ACID & Slowly Changing Data

Karthik Kondpak
9989454737

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 13 — Delta Lake, ACID & Slowly Changing Data (Production Must-Know)

Concepts Covered Today

- What is Delta Lake and why it exists
- ACID transactions on data lakes
- Delta vs Parquet
- MERGE (UPSERT)
- SCD Type-1 & Type-2 using Delta
- Time Travel & data recovery

Scenario

You are building a **customer master table** for an Indian fintech / e-commerce company.

Source: Daily customer updates (CDC)

Target: customers_delta (Delta table)

Columns:

- customer_id
- name
- city
- phone
- updated_at

Why Delta Lake?

❌ Problems with Plain Parquet

- No updates/deletes
- No transactions
- Data corruption risk

✅ Delta Lake Solves

- ACID transactions
- Schema enforcement & evolution
- Upserts (MERGE)
- Time travel

Create Delta Table

PySpark Code

```
customers_df.write \
    .format("delta") \
    .mode("overwrite") \
    .save("/delta/customers")
```

ACID Properties (MUST MEMORIZE)

Property	Meaning in Delta
Atomicity	All or nothing write
Consistency	Schema & constraints enforced
Isolation	Concurrent writes safe
Durability	Data survives failures

Question 1: MERGE (UPSERT) — MOST ASKED

🔹 Scenario

Daily customer updates contain new & existing records.

PySpark MERGE

```
from delta.tables import DeltaTable

delta_tbl = DeltaTable.forPath(spark,
"/delta/customers")

delta_tbl.alias("t").merge(
    updates_df.alias("s"),
    "t.customer_id = s.customer_id"
).whenMatchedUpdateAll() \
 .whenNotMatchedInsertAll() \
 .execute()
```

Interview Explanation

- Handles inserts & updates in one operation
- Fully ACID-compliant

Question 2: SCD Type-1 using Delta

🔹 Scenario

Customer city correction (overwrite history).

🧪 Logic

Use MERGE → overwrite existing record.

Question 3: SCD Type-2 using Delta (MOST IMPORTANT)

🔹 Scenario

Maintain customer address history.

Columns added:

- effective_from
- effective_to

- is_current

🧪 High-Level Logic

```
1.Expire current record (is_current = false)
2.Insert new record with is_current = true
```

Interviewers focus on **logic explanation**, not full code.

Question 4: Time Travel

🔹 Scenario

Rollback table to yesterday's version.

🧪 PySpark Code

```
spark.read.format("delta") \
    .option("versionAsOf", 5) \
    .load("/delta/customers")
```

Question 5: Vacuum (Cleanup Old Files)

🔹 **Interview Question**

Why do we need VACUUM?

🧪 **Answer**

- Removes old data files
- Frees storage

```
spark.sql("VACUUM delta.`/delta/customers` RETAIN 168 HOURS")
```

Let's build your Data Engineering journey together!

📩 Call us directly at: 9989454737

https://seekhobigdata.com/