Seekho Bigdata
Data is the New Oil

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 15 — Streaming + Delta Lake (CDC & Real-Time MERGE)

Karthik Kondpak
9989454737

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 15 — Streaming + Delta Lake (CDC & Real-Time MERGE)

Concepts Covered Today

- CDC (Change Data Capture) basics
- Streaming → Delta Lake integration
- MERGE in streaming pipelines
- Exactly-once semantics with Delta
- Idempotent streaming design

Scenario

You work for an **Indian fintech / banking platform**.

Source:

- `Kafka topic: customer_cdc`
- Operations: INSERT, UPDATE, DELETE

Target:

- `Delta table: customers_delta`

Goal:

- Apply **real-time CDC changes** into Delta Lake
- Maintain **latest customer state**

What is CDC?

CDC is a mechanism to **capture and propagate incremental changes** (insert/update/delete)

from source systems to downstream systems **in real time**.

Examples:

- MySQL → Kafka → Spark → Delta
- Oracle → Debezium → Kafka → Delta

⚙ Step 1: Read CDC Stream from Kafka

```
cdc_stream_df = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", "broker:9092")
\
    .option("subscribe", "customer_cdc") \
    .load()
```

🔄 Step 2: Parse CDC Payload

Assume CDC message contains:

- customer_id
- name
- city
- op (I/U/D)
- event_time

```
parsed_df = cdc_stream_df.select(
    from_json(col("value").cast("string"),
schema).alias("data")
```

```
).select("data.*")
```

Question 1: Why NOT Direct MERGE in writeStream?

❌ **Interview Trap**

You cannot call `MERGE` directly inside `writeStream.`

✅ **Correct Approach**

Use **foreachBatch** for transactional logic.

🚀 Question 2: Streaming MERGE using foreachBatch (MOST ASKED)

```
from delta.tables import DeltaTable

def upsert_to_delta(batch_df, batch_id):
    delta_tbl = DeltaTable.forPath(spark,
"/delta/customers")

    delta_tbl.alias("t").merge(
```

```
        batch_df.alias("s"),
        "t.customer_id = s.customer_id"
    ).whenMatchedUpdateAll() \
     .whenNotMatchedInsertAll() \
     .execute()

query = parsed_df.writeStream \
    .foreachBatch(upsert_to_delta) \
    .option("checkpointLocation",
"/chk/customer_cdc") \
    .start()
```

Why foreachBatch Works

- Each micro-batch is a static DataFrame
- Full Delta ACID support
- Exactly-once semantics with checkpointing

Question 3: Handling DELETE Events

🔹 Scenario

CDC event with `op = 'D'`.

MERGE Logic

```
.whenMatchedDelete(condition="s.op = 'D'")
```

Let's build your Data Engineering journey together!

📩 Call us directly at: 9989454737

https://seekhobigdata.com/