prominent Academy

PySpark Scenario-Based Interview Questions!

Capgemini

www.prominentacademy.in

+91 98604 38743

🧪**Question** : You call an external HTTP service in a mapPartitions. How do you safely retry failed rows without dropping the partition?

Code:

```python
for r in rows:
    for attempt in range(3):
        try:
            yield enrich(r)
            break
        except:
            time.sleep(2)
    else:
        log_error(r)
```

Retries within partitions avoid whole-partition failure.

🧪**Question** : Demonstrate safe use of a broadcast variable containing static lookup data across UDF calls.

Code:

```python
lookup = sc.broadcast(dict_data)
def map_func(r):
    return (r.id, lookup.value.get(r.key))
df.rdd.map(map_func)
```

🧪**Question** : You want micro-batches triggered every minute but ensure each batch only has 5 seconds to process. How to configure?

Code:

```python
df.writeStream \
  .trigger(processingTime="1 minute") \
  .option("maxTriggerDelay", "5000") \
  .option("spark.streaming.stopGracefullyOnShutdown", "true") \
  .start()
```

This ensures batches start every minute and timeout quickly if overloaded.

🧪**Question** : You want to wait until watermark advances before emitting aggregates. How?

Code:

```python
df.withWatermark("event_time", "1 hour") \
  .groupBy(window("event_time", "10 minutes")) \
  .writeStream \
  .trigger(once=True) \
  .start()
```

This emits when new watermark completes window.

🧪**Question** : How do you enable Adaptive Query Execution (AQE) and explain its benefits in iterative joins?

Code:

```python
spark.conf.set("spark.sql.adaptive.enabled", "true")
spark.conf.set("spark.sql.adaptive.shuffle.targetPostShuffleInputSize", "64MB")
```

AQE monitors runtime statistics, coalesces shuffle partitions, and optimizes join strategies dynamically.

🧪**Question** : You want your streaming job (Auto Loader) to process whenever new files arrive — near-real-time — without tight polling.

Code: Use trigger:

```python
spark.readStream.format("cloudFiles") \
  .option("cloudFiles.format","json") \
  .load(input_path) \
  .writeStream \
  .trigger(availableNow=True) \
  .start()
```

This processes all currently available files and exits — scheduler can re-launch.

🧪**Question** : You want to group user click events into session windows with 10-minute inactivity timeout. How do you implement this in PySpark?

Code:

```python
from pyspark.sql.functions import session_window
df.groupBy(session_window("event_time", "10 minutes"), "user_id") \
  .agg(count("*").alias("event_count"))
```

🧪**Question** : You want your streaming job to fail rather than silently skip when Kafka offsets are missing (e.g., due to retention gap). What option to set?

Code:

```python
df = spark.readStream.format("kafka") \
  .option("failOnDataLoss", "true") \
  .load()
```

This ensures Spark throws an error on offset gap .

🧪**Question** : Your Python UDF fails on GPU-enabled cluster due to missing libraries. How to gracefully fall back?

Approach:

- Detect GPU availability at runtime:
- Avoid crashing the entire job if GPU UDF errors out.

```python
if spark.conf.get("spark.rapids.sql.enabled") == "true":
    df = df.withColumn("gpu_feat", gpu_udf(...))
else:
    df = df.withColumn("cpu_feat", cpu_udf(...))
```

#AzureSynapse #DataEngineering
#InterviewPreparation #JobReady
#MockInterviews #Deloitte #CareerSuccess
#ProminentAcademy

prominent Academy

❌Think your skills are enough?
Think again—these Data engineer scenario-based questions could cost you your data engineering job.

In a recent interview at many big MNC's, one of our students faced scenario-based questions related to data engineering, and many candidates struggled to answer them correctly. These questions are designed to test your real-world knowledge and ability to solve complex data engineering problems.

Unfortunately, many students failed to answer these questions confidently. The truth is, preparation is key, and that's where Prominent Academy comes in!
We specialize in preparing you for spark and data engineering interviews by:
✅ Offering scenario-based mock interviews
✅ Providing hands-on training with data engineering features
✅ Optimizing your resume & LinkedIn profile
✅ Giving personalized interview coaching to ensure you're job-ready
Don't leave your future to chance!

📞Call us at **+91 98604 38743** and get the interview prep you need to succeed