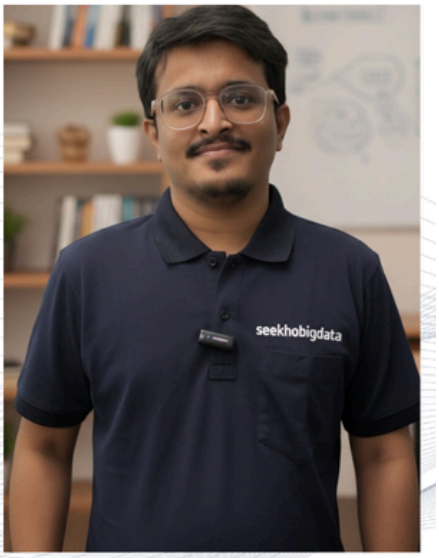
Seekho Bigdata
Data is the New Oil

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 21 — Data Skew & Skew Join Optimization (Salting, AQE)



Karthik Kondpal
9989454737

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 21 — Data Skew & Skew Join Optimization (Salting, AQE) 🔥🔥

Concepts Covered Today

- What is data skew
- How skew impacts Spark execution
- Skewed joins & aggregations
- Salting technique
- Adaptive Query Execution (AQE)
- Real Indian production scenarios

What is Data Skew?

Data skew happens when **a small number of keys hold a very large portion of data**,

causing **uneven partition sizes**.

Result: One task runs forever while others finish quickly.

Scenario

UPI transactions across India:

- 60% transactions from **UP / Maharashtra**
- Remaining 40% from other states

```
upi_df.groupBy("state").sum("amount")
```

➡ Partition handling **UP / MH** becomes massive → skew.

How Data Skew Appears in Spark UI

- One task much longer than others
- Huge Shuffle Read for single task
- Stage stuck at 99%

Why Data Skew is Dangerous

- Executor OOM errors
- Disk spill
- Network bottlenecks
- Wasted cluster resources

Interview Question 1: How to Detect Data Skew?

Best Answers

- Analyze key distribution
- Check Spark UI (Shuffle Read)
- Compare partition sizes

```
df.groupBy("key").count().orderBy("count",
ascending=False)
```

Salting Technique

🔹 Concept

Break a skewed key into **multiple artificial keys** to spread load.

Before Salting (Skewed)

```
orders.join(customers, "customer_id")
```

After Salting (Optimized)

```
from pyspark.sql.functions import rand, concat

salted_orders = orders.withColumn(
    "salt",
    (rand() * 10).cast("int")
)

salted_customers = customers.withColumn(
    "salt",
    lit(0)
)
```

```
salted_orders.join(
    salted_customers,
    ["customer_id", "salt"]
)
```

Salting increases data size — use carefully.

Adaptive Query Execution (AQE) — GAME CHANGER

🔹 What is AQE?

AQE dynamically optimizes queries **at runtime**, not compile time.

Introduced in **Spark 3.x**.

Enable AQE

```
spark.conf.set("spark.sql.adaptive.enabled", "true")
```

How AQE Fixes Skew Automatically

- Detects skewed partitions
- Splits large partitions
- Converts Sort-Merge Join → Broadcast Join (if possible)

```
spark.conf.set("spark.sql.adaptive.skewJoin.enabled", "true")
```

Is AQE enough to handle all skew problems?

✅ **Correct Answer**

No. Severe skew still needs **manual techniques like salting**.

Skew in Aggregations vs Joins

Area	Problem	Solution
Aggregation	Hot keys	Pre-aggregation
Join	Skewed keys	Salting / AQE

🛠️ Other Skew Handling Techniques

Filter skewed keys separately

Broadcast non-skewed table

Increase shuffle partitions

Use AQE + salting combo

Seekho Bigdata

Data is the New Oil

Let's build your Data Engineering journey together!

📩 Call us directly at: 9989454737

https://seekhobigdata.com/