Seekho Bigdata
Data is the New Oil

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 23 — Small Files Problem, File Compaction & Write Optimization

Karthik Kondpak
9989454737

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 23 — Small Files Problem, File Compaction & Write Optimization 🔥

Concepts Covered Today

- What is the Small Files Problem
- Why small files kill performance
- File compaction strategies
- Write optimization techniques
- Delta Lake OPTIMIZE & Z-ORDER
- Real Indian production scenarios

What is the Small Files Problem?

The **Small Files Problem** occurs when a data lake contains **thousands or millions of tiny files** instead of fewer large files.

Each file creates overhead in **HDFS / cloud storage + Spark planning**.

Scenario

You work for an **Indian fintech company** processing:

- UPI transactions every minute
- Structured Streaming writes to Delta

Result after few days:

- Millions of files (~5–50 KB each)
- Queries become very slow

Why Small Files Are Dangerous

- High metadata overhead
- Slow job startup time
- Too many tasks created
- Driver memory pressure
- Inefficient IO

Interview Question: How Small Files Are Created?

Correct Reasons

- Too many partitions
- Streaming micro-batches
- Frequent appends

- Default partitioning

Write Path Internals

```
Partitions → Tasks → Files
```

One task usually writes **one output file**.

Write Optimization — Reduce Files at Source

◆ Control Partitions Before Write

```
df.coalesce(10) \
  .write \
  .mode("append") \
  .parquet("/data/upi")
```

Repartition vs Coalesce

Aspect	repartition	coalesce
Shuffle	Yes	No
File control	Yes	Yes

Best use	Balance data	Reduce files

Structured Streaming & Small Files

◆ Use Trigger Once (Batch-like)

```
.writeStream \
.trigger(once=True)
```

Delta Lake Compaction

◆ OPTIMIZE Command

```
OPTIMIZE delta.`/delta/upi_transactions`
```

➡ Combines small files into larger files.

Z-ORDER Optimization

```
OPTIMIZE delta.`/delta/upi_transactions`
ZORDER BY (customer_id, txn_date)
```

✅ Benefits

- Faster selective queries
- Better data skipping

Seekho Bigdata

Data is the New Oil

Let’s build your Data Engineering journey together!

📩 Call us directly at: 9989454737

🌐 https://seekhobigdata.com/