Seekho Bigdata
Data is the New Oil

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 18 — Shuffle, Partitioning & Parallelism

Karthik Kondpak
9989454737

PySpark Scenario-Based Interview Questions (Complete Notes Series)

DAY 18 — Shuffle, Partitioning & Parallelism

Concepts Covered Today

- What is Shuffle & why it is expensive
- Default shuffle partitions
- Repartition vs Coalesce
- Partitioning strategies
- Parallelism tuning
- Real interview performance scenarios

What is Shuffle?

A **Shuffle** is the process of **redistributing data across executors** based on a key.

Occurs during:

- `groupBy`
- `join`
- `distinc t`
- `orderBy`

Shuffle involves **disk I/O + network transfer** → very expensive.

Scenario

You are processing **UPI transaction data** across India.

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

➡ `groupBy(state)`triggers a **shuffle**.

Default Shuffle Partitions (VERY COMMON QUESTION)

```
spark.conf.get("spark.sql.shuffle.partitions")
```

🔹 Default Value

```
200
```

Bad for small data, insufficient for very large data.

How to Tune Shuffle Partitions

```
spark.conf.set("spark.sql.shuffle.partitions", 50)
```

Rule of Thumb

- Small data → reduce partitions
- Large data → increase partitions

Repartition vs Coalesce

Feature	repartition	coalesce
Shuffle	Yes	No
Increase partitions		
Decrease partitions		
Use case	Balance data	Reduce files

Example

```
df = df.repartition(100)   # full shuffle

df = df.coalesce(10)       # no shuffle
```

Parallelism Explained

🔹 Definition

Parallelism = Number of tasks running simultaneously.

Controlled by:

- Number of partitions
- Number of executor cores

Does increasing partitions always improve performance?

✅ Correct Answer

No. Too many small partitions increase task scheduling overhead.

Partitioning Strategies

🔹 Hash Partitioning (Default)

```
df.repartition("customer_id")
```

🔹 Range Partitioning

```
df.repartitionByRange("order_date")
```

How to Identify Shuffle in Spark UI?

- New stage creation
- Shuffle Read / Write metrics
- Tasks stuck at last stage

Why One Task Runs Very Long?

✅ Correct Explanation

- Data skew

- Uneven partitions
- Hot keys

Let's build your Data Engineering journey together!

📩 Call us directly at: 9989454737

https://seekhobigdata.com/