Seekho Bigdata
Data is the New Oil

PySpark Scenario-Based Interview Questions

DAY 7 — Aggregations & GroupBy

Karthik Kondpak
9989454737

PySpark Scenario-Based Interview Questions

DAY 7 — Aggregations & GroupBy (Advanced Real-Time Scenarios)

Concepts Covered Today

- Basic & multi-column aggregations
- Conditional aggregation
- Distinct counts
- Percentage & ratio metrics
- GroupBy vs Window (interview trap)

Sample Data: `sales_df` (Indian Retail Scenario)

order_id	customer_id	state	category	amount	order_date
1	101	Karnataka	Mobile	20000	2024-01-01
2	101	Maharashtra	Laptop	55000	2024-01-10
3	102			18000	2024-01-15
4	103	Telangana	TV	40000	2024-02-01
5	103	Telangana	Mobile	22000	2024-02-12

✅ Question 1: Total Sales Per State

🔹 Scenario

Management wants to see **total revenue generated per Indian state**.

🧪 PySpark Solution

```
from pyspark.sql.functions import sum

state_sales_df = sales_df.groupBy("state") \
    .agg(sum("amount").alias("total_sales"))

state_sales_df.show()
```

📝 Explanation

- Simple aggregation, but very frequently asked
- Often extended with filters or joins in interviews

✅ Question 2: Category-wise Sales Per State (Multi-Level GroupBy)

🔹 Scenario

Business wants **category-wise revenue split per state**.

🧪 PySpark Solution

```
category_state_df = sales_df.groupBy("state", "category") \
    .agg(sum("amount").alias("category_sales"))

category_state_df.show()
```

📝 Explanation

- Demonstrates multi-dimensional reporting
- Very common in dashboards

✅ Question 3: Conditional Aggregation (Mobile Sales Only)

🔹 Scenario

Find **total mobile sales per state**, ignoring other categories.

🧪 PySpark Solution

```
from pyspark.sql.functions import when

mobile_sales_df = sales_df.groupBy("state") \
    .agg(sum(when(col("category") == "Mobile",
col("amount")).otherwise(0))
         .alias("mobile_sales"))

mobile_sales_df.show()
```

📝 Explanation

- Conditional aggregation is a **favorite interview topic**
- Avoids filtering and losing other data

✅ Question 4: Distinct Customer Count Per State

🔹 Scenario

Marketing wants to know **unique customers per state**.

🧪 PySpark Solution

```
from pyspark.sql.functions import countDistinct

customer_count_df = sales_df.groupBy("state") \
    .ag g(coun tDist inct( "custo mer_id" ).alia s("un ique_ c
ustomers"))

customer_count_df.show()
```

📝 Explanation

- Distinct counts are expensive → discuss performance
- Interviewers may ask optimization follow-ups

✅ Question 5: Percentage Contribution of Each Category

🔹 Scenario

Analytics team wants **category contribution (%) to total sales**.

🧪 PySpark Solution

```
from pyspark.sql.window import Window
from pyspark.sql.functions import sum

category_total_df = sales_df.groupBy("category") \
    .agg(sum("amount").alias("category_sales"))

window_spec = Window.partitionBy()

percentage_df = category_total_df.withColumn(
    "percentage_contribution",
    col("category_sales") * 100 /
sum("category_sales").over(window_spec)
)
```

```
percentage_df.show()
```

📝 Explanation

- Combines aggregation + window function
- Common analytics interview pattern

Let's build your Data Engineering journey together!

Call us directly at: 9989454737

https://seekhobigdata.com/