Seekho Bigdata

Data is the New Oil

DAY 1 — Scenario-Based PySpark Questions

Topic Focus: Basic Transformations, Aggregations & Business Logic

Karthik Kondpak
9989454737

DAY 1 — Scenario-Based PySpark Questions

TopicFocus:BasicTransformations,Aggregations & Business Logic

Question 1: Total Sales Per State

Scenario

You are working as a Data Engineer for an Indian e-commerce company (like Flipkart).

You receive daily sales data containing **order_id, state, amount**.

The business team wants to know:

Total sales amount per Indian state

Sample Data

order_id	state	amount
101	Maharashtra	12000
102	Karnataka	9000
103	Maharashtra	15000
104	Tamil Nadu	8000
105	Karnataka	7000

Expected Output

state	total_sales
Maharashtra	27000
Karnataka	16000
Tamil Nadu	8000

PySpark Solution

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import sum

spark = SparkSession.builder.getOrCreate()

data = [
    (101, "Maharashtra", 12000),
    (102, "Karnataka", 9000),
    (103, "Maharashtra", 15000),
    (104, "Tamil Nadu", 8000),
    (105, "Karnataka", 7000)
]

columns = ["order_id", "state", "amount"]

df = spark.createDataFrame(data, columns)

result =
df.groupBy("state").agg(sum("amount").alias("total_sales"))

result.show()
result.show()

```

Question 2: Find Customers with Multiple Orders

Scenario

You are analyzing customer behavior for an Indian retail chain.

Management wants to identify **customers who placed more than 1 order**.

Sample Data

customer_id	customer_name	order_id
1	Rahul	201
1	Rahul	202
2	Priya	203
3	Amit	204
3	Amit	205

Expected Output

customer_id	customer_name	order_count
1	Rahul	2
3	Amit	2

PySpark Solution

```
from pyspark.sql.functions import count

result = (
    df.groupBy("customer_id", "customer_name")
      .agg(count("order_id").alias("order_count"))
      .filter("order_count > 1")
)

result.show()
```

Question 3: Identify High-Value Orders

Scenario

For GST audit purposes, the finance team wants to track **orders above ₹10,000**.

Sample Data

order_id

	customer	amount
301	Anil	8500
302	Sunita	12500
303	Ramesh	22000
304	Neha	6000

Expected Output

order_id	customer	amount

302	Sunita	12500
303	Ramesh	22000

PySpark Solution

```
result = df.filter(df.amount > 10000)
result.show()
```

Question 4: Count Orders Per City

Scenario

You work for a food delivery startup in India (Zomato/Swiggy).

The operations team wants **order count per city**.

Sample Data

order_id	city
401	Bengaluru
402	Bengaluru
403	Mumbai
404	Delhi
405	Mumbai

Expected Output

city	total_orders
Bengaluru	2

Mumbai	2
Delhi	1

PySpark Solution

```
from pyspark.sql.functions import count

result = df.groupBy("city").agg(count("*").alias("total_orders"))

result.show()
```

Interview Notes

- count("*") counts rows
- Used in **dashboard metrics**

Question 5: Add Discount Column Based on Amount

Scenario

An Indian fashion retailer applies:

- **10% discount if amount ≥ ₹10,000**
- Otherwise **no discount**

Sample Data

order_id	amount
501	8000
502	12000
503	15000

Expected Output

order_id	amount	discount
501	8000	0
502	12000	1200
503	15000	1500

PySpark Solution

```
from pyspark.sql.functions import when, col

result = df.withColumn(
    "discount",
    when(col("amount") >= 10000, col("amount") *
0.10)
    .otherwise(0)
)

result.show()
```

Seekho Bigdata

Data is the New Oil

Let’s build your Data Engineering journey together!

📩 Call us directly at: 9989454737

🌐 https://seekhobigdata.com/