Seekho Bigdata
Data is the New Oil

Spark Optimization Topic

Karthik Kondpak
9989454737

Day 12 — Spark Optimization Topic

🔥 Window Functions — The Core of Analytical Workloads

Window functions are one of the **most important features in Spark**.

They allow you to perform **analytics across groups of rows** *without* collapsing them—unlike GROUP BY.

If you are doing:

✔ Time series

✔ Customer analytics

✔ Sessionization

✔ Ranking

✔ Running totals

...you **must know window functions deeply**.

Why Window Functions Matter

- They perform calculations **over a window of rows**.
- They **do not reduce** the number of rows (unlike aggregate).
- They work with **partition + order**.
- They power 70% of analytical pipelines (banking, ecommerce, telecom).

Key Components

A window has 3 parts:

`PARTITION BY → defines group`

`ORDER BY → defines sequence`

`FRAME → defines row range`

Most Important Window Functions

Ranking Functions

- `row_number()`

- rank()
- dense_rank()

Value Functions

- lag()
- lead()
- first_v alue()
- last_va lue()

Aggregation with Window

- sum()
- avg()
- count()

Scenario Example

📂 Dataset: Daily Sales of BigBasket India

date	city	amount

2024-01-01	Bengaluru	12000
2024-01-02	Bengaluru	15000
2024-01-03	Bengaluru	11000
2024-01-01	Delhi	10000
2024-01-02	Delhi	17000

⭐ Example 1 — Running Total (Cumulative Sum)

```
from pyspark.sql.window import Window
import pyspark.sql.functions as F


w = Window.partitionBy("city").orderBy("date")

df2 = df.withColumn("running_total",
F.sum("amount").over(w))
```

✔Bengaluru: 12000 → 27000 → 38000

✔Delhi: 10000 → 27000

⭐ Example 2 — Previous Day Sales (LAG)

```
df2 = df.withColumn(
    "prev_day_amount",
    F.lag("amount", 1).over(w)
)
```

⭐ Example 3 — Ranking Cities by Daily Sales

```
rank_w =
Window.partitionBy("date").orderBy(F.desc("amount"))


df3 = df.withColumn("rank", F.rank().over(rank_w))
```

Frame Types Explained (Very Important)

ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW

Running totals

ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING

Sliding window

RANGE BETWEEN

Used for numerical ranges (less common)

Let's build your Data Engineering journey together!

Call us directly at: 9989454737

https://seekhobigdata.com/