Seekho Bigdata
Data is the New Oil

Predicate Pushdown

Karthik Kondpak
9989454737

Day 1 — Spark Optimization Topic 1.

Predicate Pushdown

PredicatePushdown isoneof themost powerful andbasic Sparkoptimizations.

It tells Spark to push your filters down to the data source (Parquet, ORC, Delta, JDBC) so Spark reads only the required rows instead of scanning the full file.

Why Predicate Pushdown Matters

- Reduces data scanned
- Reduces memory usage
- Reduces shuffle
- Speeds up queries 2× to 10×
- Works best for columnar formats (Parquet, ORC, Delta)

How It Works

WithoutPredicatePushdown

Spark reads entire dataset → applies filter later.

```
df = spark.read.parquet("/delta/sales")
result = df.filter("country = 'India'")
```

With Predicate Pushdown

Spark sends filter to Parquet reader → reads only the matching rows.

```
df = spark.read \
    .option("spark.sql.parquet.filterPushdown", "true") \
    .parquet("/delta/sales")

result = df.filter(col("country") == "India")
```

Check in Spark UI

Go to:SQL →PhysicalPlan

You should see:

```
PushedFilters: [EqualTo(country, India)]
```

This confirms pushdown is applied.

Real Indian Scenario Example

Dataset: /delta/transactions

txn_id	state	amount	payment
101	Maharashtra	1200	UPI
102	Karnataka	1500	Card
103	Tamil Nadu	700	Cash
104	Maharashtra	900	UPI

Goal: Get all transactions from Maharashtra

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

```
df = spark.read.parquet("/delta/transactions")
mh_df = df.filter(col("state") == "Maharashtra")
```

SB

Seekho Bigdata

Data is the New Oil

Let’s build your Data Engineering journey together!

📩 Call us directly at: 9989454737

🌐 https://seekhobigdata.com/