SB
Seekho Bigdata
Data is the New Oil

Column Pruning

Karthik Kondpak
9989454737

Day 2 — Spark Optimization Topic

2. Column Pruning

Column PruningisaSpark optimizationwhere Spark reads onlythe columns you actually need instead of scanning the entire dataset.

It works automatically when using Parquet, ORC, and Delta because they are columnar formats.

Why Column Pruning Matters

- Reduces I/O (Spark reads fewer bytes)
- Reduces memory usage
- Reduces shuffle size
- Improves CPU performance
- Can boost speed by 3× to 20× for wide tables

If your table has 200 columns and you need only 5, Spark should read only those 5.

How It Works

WithoutColumnPruning
Spark reads all columns → keeps only required ones later.

```
df = spark.read.parquet("/mnt/data/orders")
selected = df.select("order_id", "amount")
```

If the format is not columnar (CSV, JSON), Spark has to read the full file.

With Column Pruning

Spark instructs the Parquet/Delta reader to read only needed columns.

```
df = spark.read.parquet("/mnt/data/orders") \
               .select("order_id", "amount")
```

The physical plan will show:

```
ReadSchema: struct<order_id:int, amount:double>
```

This means Spark did not read unnecessary columns.

Check in Spark UI (or explain plan)

Run:

```
df.explain(True)
```

Look for:

```
PushedProjection: [order_id, amount]
```

or

```
ReadSchema: order_id, amount
```

This confirms column pruning is applied.

Scenario Example

Dataset:/delta/zomato_orders

order_id	user_id	restaurant_id	city	food_items	gst	delivery_fee	payment_type	timestamp

Total columns: 9

You need only: order_id, city, payment_type.

Query

```
df =
spark.r ead.fo rmat( "delt a").lo ad("/de lta/zo mato_ order s
")

result = df.select("order_id", "city",
"payment_type")
```

What happens with pruning?

- Delta reads only 3 of 9 columns.
- 66% I/O reduction.
- Less shuffle during joins/aggregations.
- Dramatic gain for large tables (1B+ rows).

When Column Pruning Does NOT Work

- Reading CSV or JSON (full file must be parsed)
- When using Python UDFs (Spark loses column-level optimization)
- When selecting columns after a transformation that breaks lineage (for example, explode + nested columns)
- When using selectExpr with non-deterministic expressions

Tips to Ensure Column Pruning Works

1. Use columnar formats (Parquet, ORC, Delta)
2. Push select() as early as possible
3. Avoid using UDFs before projection
4. Verify using explain(True)
5. Avoid using select("*") unless required

Seekho Bigdata

Data is the New Oil

Let’s build your Data Engineering journey together!

📩 Call us directly at: 9989454737

🌐 https://seekhobigdata.com/