What Are SQL Join Hints?

SQL **join hints** are optimization directives that guide Spark's Catalyst optimizer to choose a specific join strategy. They are useful when you know the data characteristics and want to improve performance.

✅ Broadcast Join Hint

Syntax:

```
SELECT /*+ BROADCAST(table_alias) */ ...
```

Description:

Tells Spark to **broadcast** the specified table to all worker nodes.

Works best when the **broadcasted table is small** and the other table is large.

Avoids **shuffle** by sending the small table to all partitions.

✅ Example:

```
SELECT /*+ BROADCAST(c) */
  t.id, t.countrycode, t.amount, c.countryname
FROM transactions t
JOIN countries c
ON t.countrycode = c.countrycode;
```

⚙️ How It Works:

Step Action

1   Spark scans and filters both tables.

2   The countries table is **broadcasted** to all executors.

3   Spark performs a **Broadcast Hash Join**.

4   No shuffle or sorting is required, which improves performance.

📈 When to Use:

When **one side of the join is small** (e.g., a reference or dimension table).

When **minimizing shuffle** is critical for performance.