

Four Major Phases of Spark SQL Engine Execution:

1. Analysis

- **Input:** Unresolved Logical Plan
- **Action:**
 - Validates references to tables and columns using metadata from the **Catalog**.
 - Converts Unresolved Logical Plan to Resolved Logical Plan.

2. Logical Planning (Logical Optimization)

- **Input:** Resolved Logical Plan
- **Action:**
 - Applies various optimization rules.
 - Combines multiple filters, pushes predicates down, prunes unnecessary columns, etc.
 - Outputs the Optimized Logical Plan.

3. Physical Planning

- **Input:** Optimized Logical Plan
- **Action:**
 - Generates multiple candidate physical plans based on available strategies (like broadcast joins, sort-merge joins).
 - Uses a **Cost Model** to select the most efficient physical plan.
 - Result: Best Physical Plan.

4. Code Generation

- **Input:** Best Physical Plan
- **Action:**
 - Converts the plan into **Java Bytecode** using **WholeStageCodegen**.
 - Generates final **RDD** transformations and actions.

Final Output:

- Efficient and optimized Java Bytecode and RDDs
- Executes across Spark executors in a distributed fashion

[Code: SQL | DataFrame | Dataset]



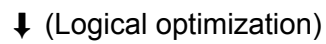
Catalyst SQL Engine



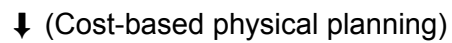
1. Unresolved Logical Plan



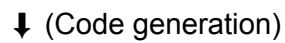
2. Resolved Logical Plan



3. Optimized Logical Plan



4. Best Physical Plan



Final RDDs + Java Bytecode Execution