Spark- Sql Optimiser

System Overview

1. **Reads SQL metadata from an Excel file**.
2. **Analyzes the source tables (Class 2)**:
    - Identifies join keys.
    - Detects **data skew** in columns.
3. **Generates histograms and classifies columns(Class 3) into Pdf**:
    - Skewed columns
    - Non-skewed columns
4. **Optimizes SQL queries (Class 1)**:
    - Uses **CTEs** and **grouped skewed data**.
    - Applies join optimizations with string concatenation of skewed data.
    - Outputs optimized SQL.

Class Responsibility Mapping

Class	Responsibility
Class 1/ ETLProfilerFromExcel	Query Optimizer: Uses data skew info to rewrite the SQL with optimized joins.
Class 2/SparkRankProfiler	Skew Analyser: Detects skew based on the hash logic from the base tables
Class 3/PercentilePlotter	Skew Plotter: Generates histograms for columns skewness into pdf for user to visualise
Custom SQL Parser (In dev)	Parses raw SQL for join conditions, types, source columns, subqueries.

Architecture Diagram Description