# SparkSQL

2020年5月19日 22:20

Source: SIGMOD
Title: Spark SQL: Relational Data Processing in Spark
Publish year: 2015

## Contribution

1. Presented a new module (**SparkSQL**) in Apache Spark providing rich integration with relational processing.
    a. Automatic optimization, complex pipelines that mix relational and complex analytics.
    b. Support a wide range of features tailored to large-scale data analysis.
        i. Semi-structured data, query federation, and machine learning data types.
    c. Based on Catalyst optimizer and DataFrames data type.
2. Integrate **DataFrame** API into SparkSQL.
3. Use **Catalyst** to optimize SparkSQL.

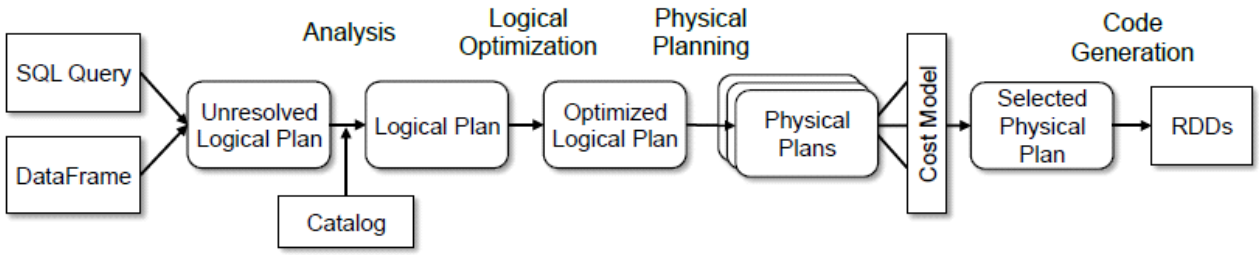
## Summary

1. SparkSQL is a SQL engine for distributed data processing.
2. It uses DataFrame as abstract data type which is like table in SQL.
3. An optimizer called Catalyst make SparkSQL faster.

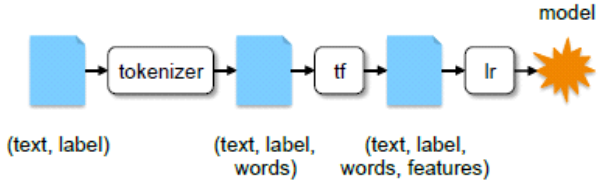
## Work evaluation

**Theoretical analysis**

1. Aggregated with compiler principle, SparkSQL is very similar.



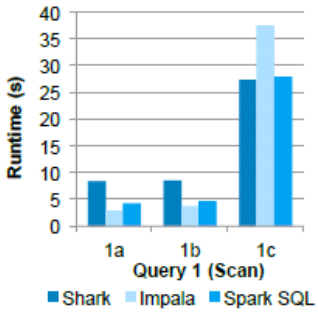

    a. Lines of code: analysis (1000), logical optimization (800), physical planning (500), code generation (700).
2. Easy to expand for rules and optimization.
3. Support schema inference for semi-structured data, integration with Spark's machine learning library (pipeline), and query federation to external databases.

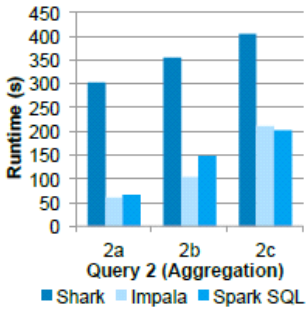


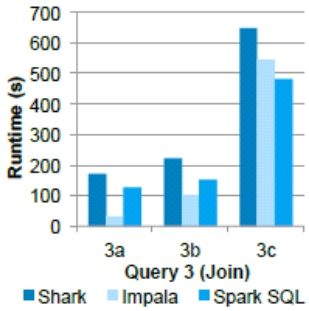

    ○ TF: Hashing TF. LR: Logistic Regression.

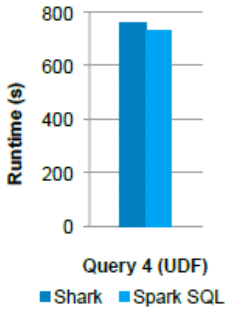
**Metrics**

1. Benchmark:



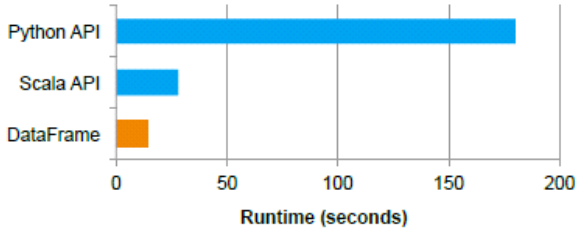




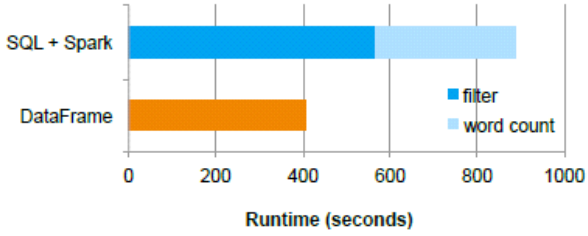



Faster for large dataset and good at aggregation and join (relation).

2. DataFrame performance rather than other language API:




3. Pipeline performance of 2-stage:




**Data set**
AMPLab big data benchmark.

**Methods comparation**
Faster than Impala in large dataset and complex aggregation.

Source code / data
http://spark.apache.org

Future work

1. Generalized online aggregation.
2. Computational genomics for large where condition.

Next paper