**Review of Spark SQL**

# **Summary**

Spark SQL is a new module in Apache Spark that lets:

Programmer: leverage the benefits of relational processing.

User: call complex analytics libraries.

**What’s new:**

1. Offer much tighter integration between relational and procedural processing.
2. Include a highly extensible optimizer, Catalyst, that makes it easy to add composable rules, control code generation, and define extension points.

**Contributions**

1. Support relational processing both within Spark programs and on external data sources using a programmer-friendly API.
2. Provide high performance using established DBMS techniques.
3. Easily support new data sources, including semi-structured data and external databases amenable to query federation.
4. Enable extension with advanced analytics algorithms such as graph processing and machine learning.

**Comments**

Spark SQL makes it easy to add optimization rules, data sources and data types. Spark SQL also makes it significantly simpler and more efficient to write data pipelines that mix relational and procedural processing, while offering substantial speedups over previous engines.

1. Provides a DataFrame API that can perform relational operations on both external data sources and built-in distributed collections.
2. Introduces a novel extensible optimizer, Catalyst, to support the wide range of data sources and algorithms in big data.

**Limitations:** the engine does not understand the structure of the data in RDDs or the semantics of user functions.