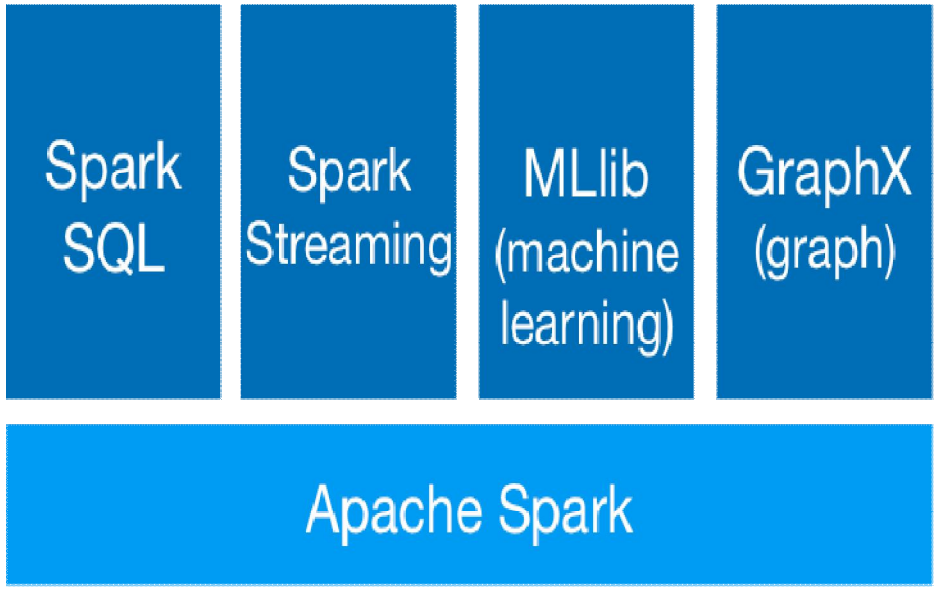
SPARK SQL

Spark SQL

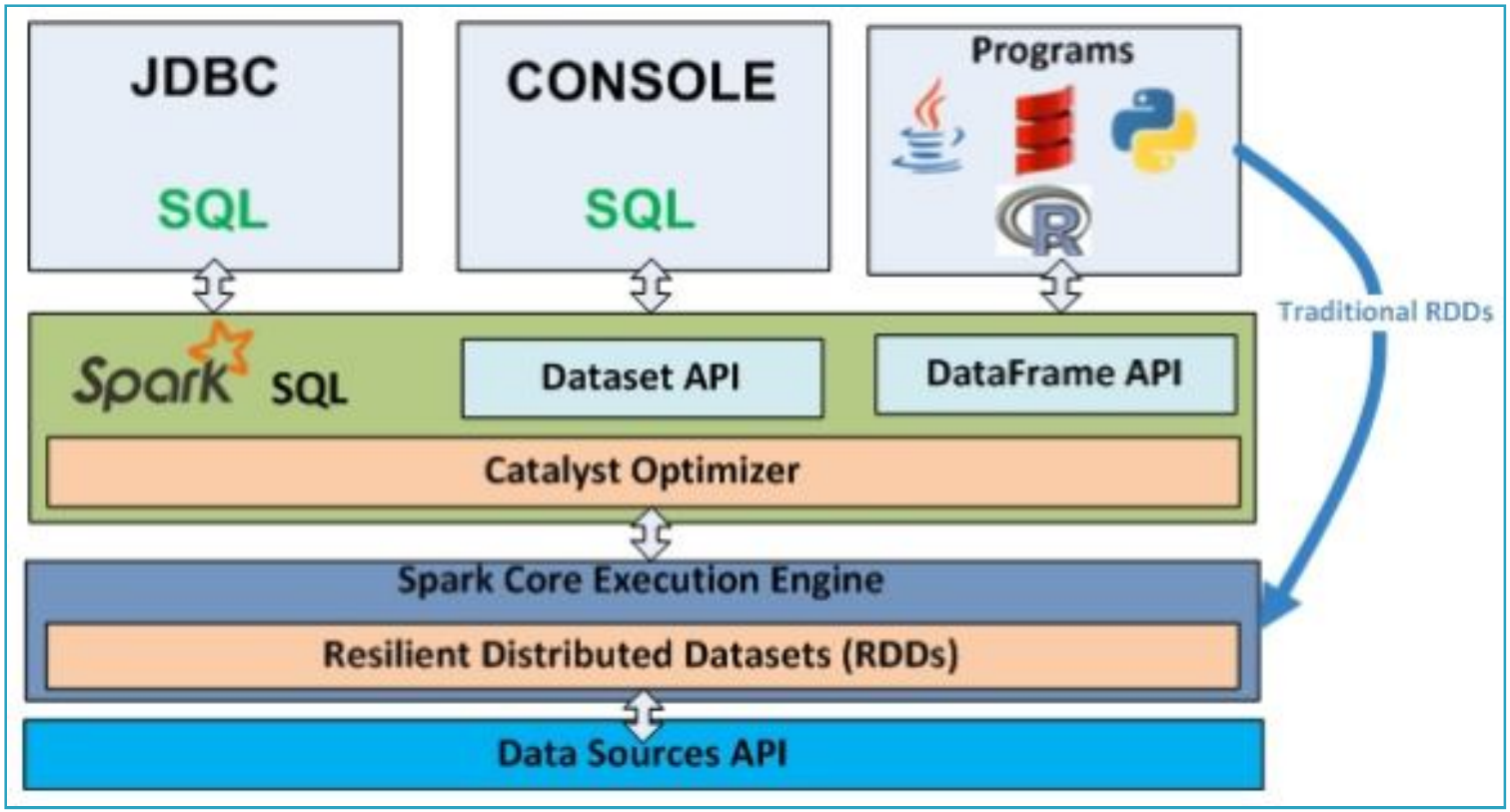
- Spark SQL is a Spark module for structured data processing.
- Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD.



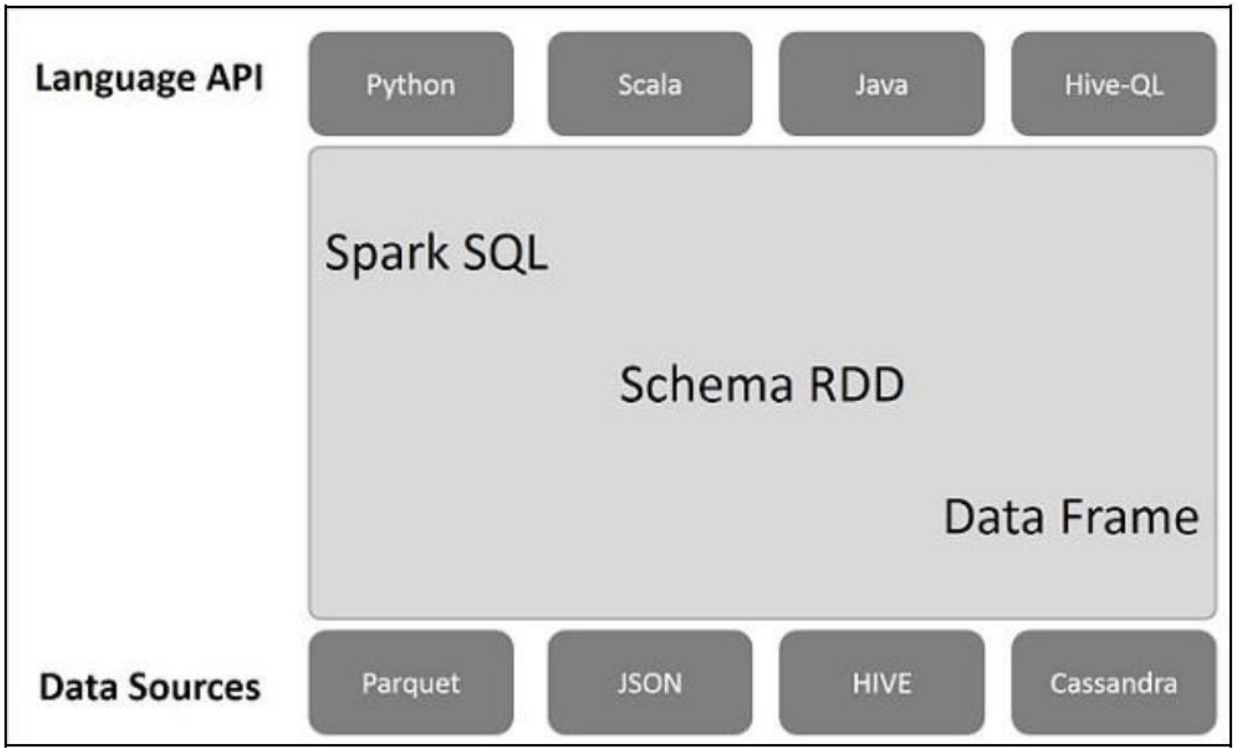
Spark SQL

- Spark SQL was first released in Spark 1.0 (May, 2014).
- Initialy committed by Michael Armbrust & Reynold Xin from Databricks.
- Spark introduces a programming module for structured data processing called Spark SQL.
- It provides a programming abstraction called DataFrame and can act as distributed SQL query engine.

Spark SQL Overview



Spark SQL Overview



Spark SQL Architecture

- Language API:
  - Spark is compatible with different languages and Spark SQL.
  - It is also, supported by these languages- API (python, scala, java, HiveQL).
- Schema RDD:
  - Spark Core is designed with special data structure called RDD.
  - Generally, Spark SQL works on schemas, tables, and records.
  - Therefore, we can use the Schema RDD as temporary table.
  - We can call this Schema RDD as Data Frame.
- Data Sources:
  - Usually the Data source for spark-core is a text file, Avro file, etc. However, the Data

    Sources for Spark SQL is different.
  - Those are Parquet file, JSON document, HIVE tables, and Cassandra database.

Features of Spark SQL

1. Integrated:
   - Seamlessly mix SQL queries with Spark programs.
   - Spark SQL lets you query structured data as a distributed dataset (RDD) in Spark, with integrated APIs in Python, Scala and Java.
   - This tight integration makes it easy to run SQL queries alongside complex analytic algorithms.
2. Unified Data Access:
   - Load and query data from a variety of sources.
   - Schema-RDDs provide a single interface for efficiently working with structured data, including Apache Hive tables, parquet files and JSON files.

Features of Spark SQL

3. Hive Compatibility:
   - Run unmodified Hive queries on existing warehouses.
   - Spark SQL reuses the Hive frontend and MetaStore, giving you full compatibility with existing Hive data, queries, and UDFs.
   - Simply install it alongside Hive.


```
SELECT COUNT(*) FROM
hiveTable WHERE
hive_udf(data)
```

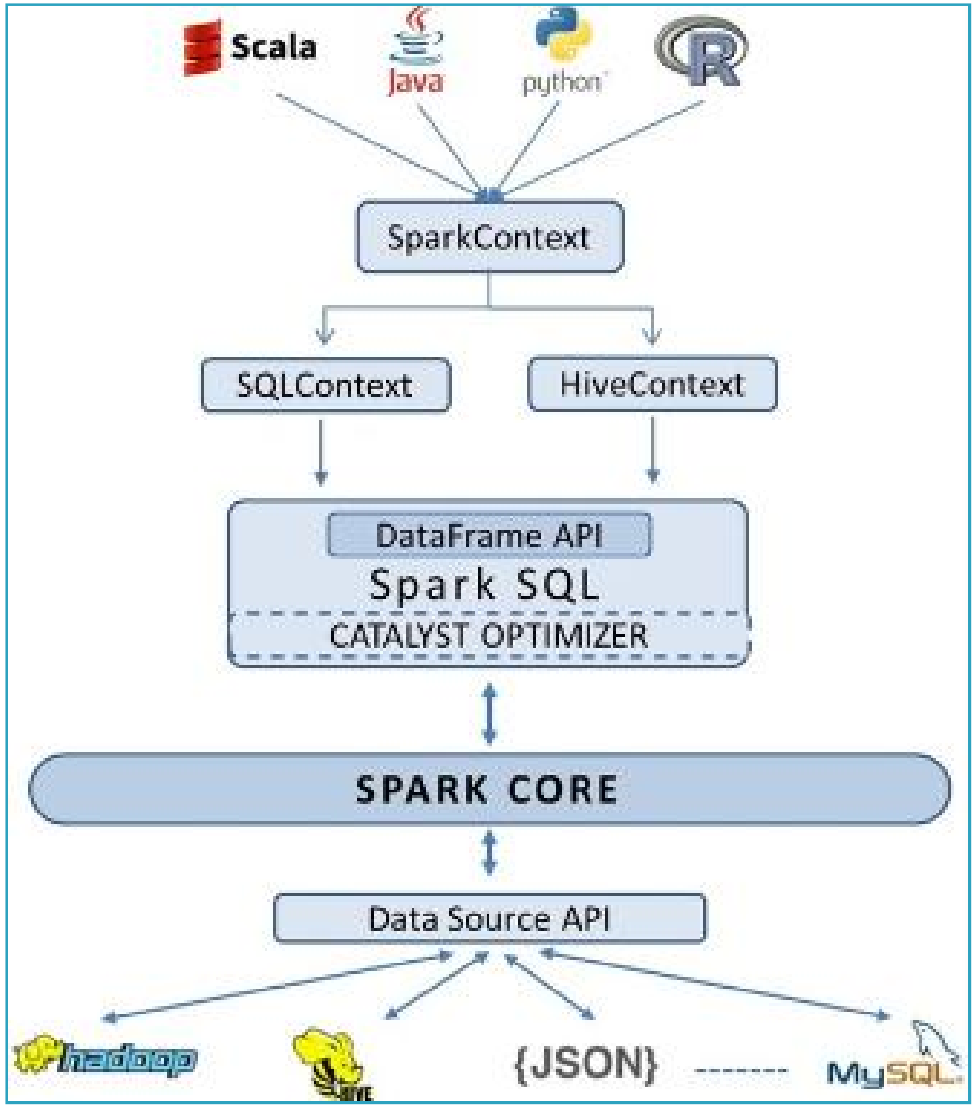
Features of Spark SQL

4. Standard Connectivity:
   - Connect through JDBC or ODBC.
   - Spark SQL includes a server mode with industry standard JDBC and ODBC connectivity.

Features of Spark SQL

5. Scalability:
   - Use the same engine for both interactive and long queries.
   - Spark SQL takes advantage of the RDD model to support mid-query fault tolerance, letting it scale to large jobs too.
   - Do not worry about using a different engine for historical data.

Spark SQL Architecture



Dataset and DataFrame

- Adistributed collection of data, which is organized into named columns.
- Conceptually, it is equivalent to relational tables with good optimization techniques.
- ADataFrame can be constructed from an array of different sources such as Hive tables, Structured Data files, external databases, or existing RDDs.
- This API was designed for modern Big Data and data science applications taking inspiration from DataFrame in RProgramming and Pandas in Python.

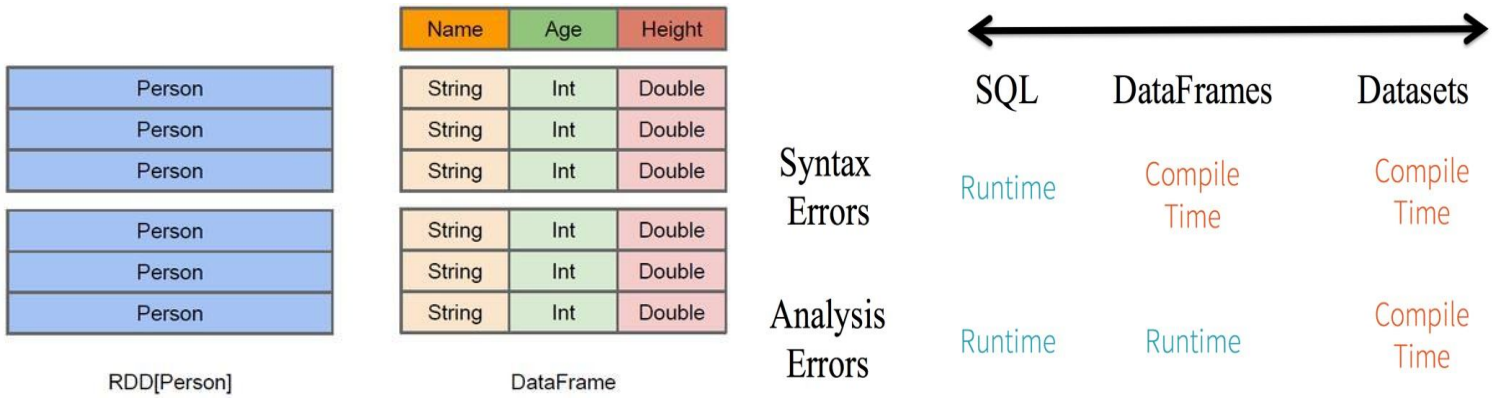
Dataset and DataFrame

DataFrame

Data is organized into named columns, like a table in a relational database

Dataset: a distributed collection of data

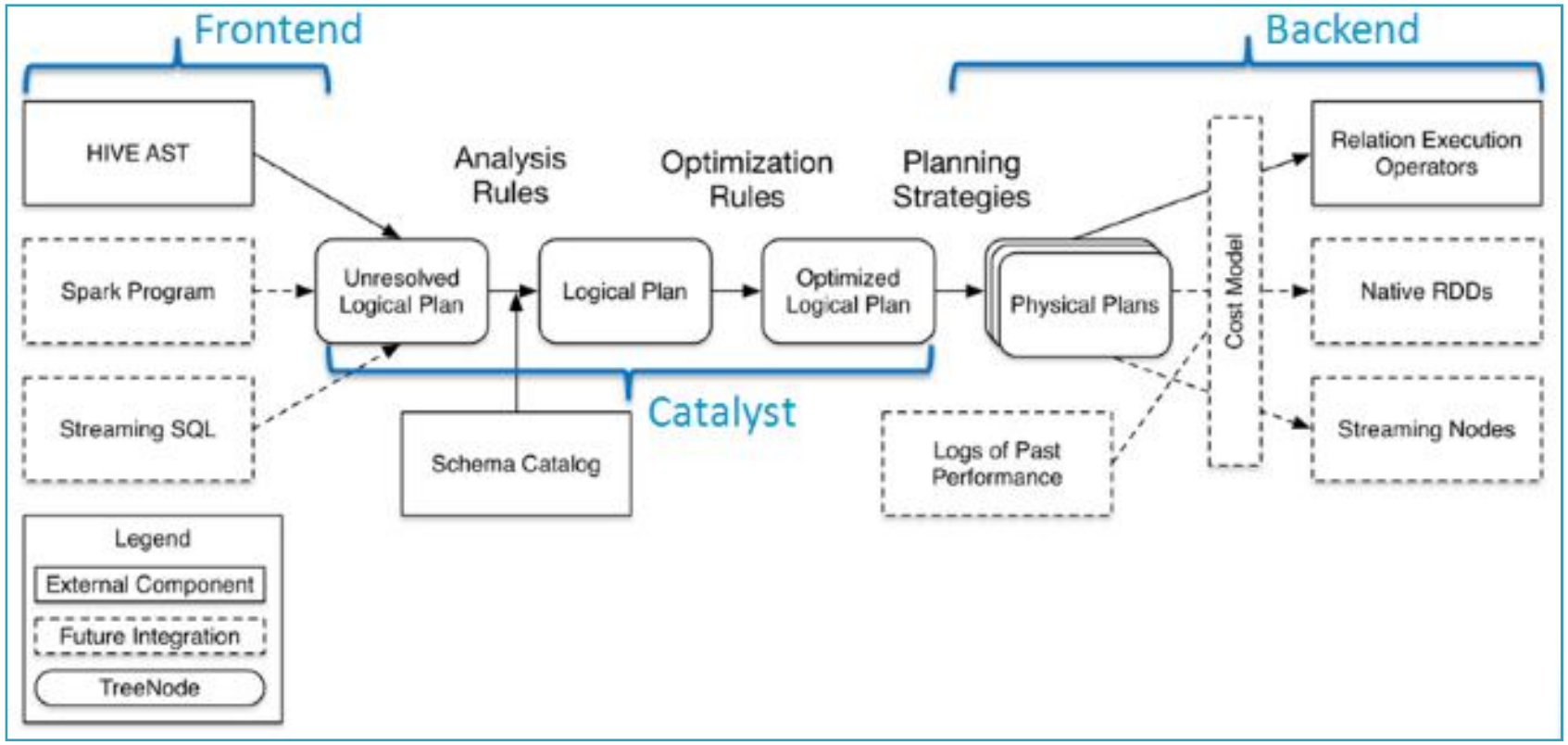
- A new interface added in Spark 1.6
- Static-typing and runtime type-safety



Features of DataFrame

- Ability to process the data in the size of Kilobytes to Petabytes on a single node cluster to large cluster.
- Supports different data formats (Avro, csv, elastic search, and Cassandra) and storage systems (HDFS, HIVE tables, mysql, etc).
- State of art optimization and code generation through the Spark SQL Catalyst optimizer (tree transformation framework).
- Can be easily integrated with all Big Data tools and frameworks via Spark-Core.
- Provides API for Python, Java, Scala, and R Programming.

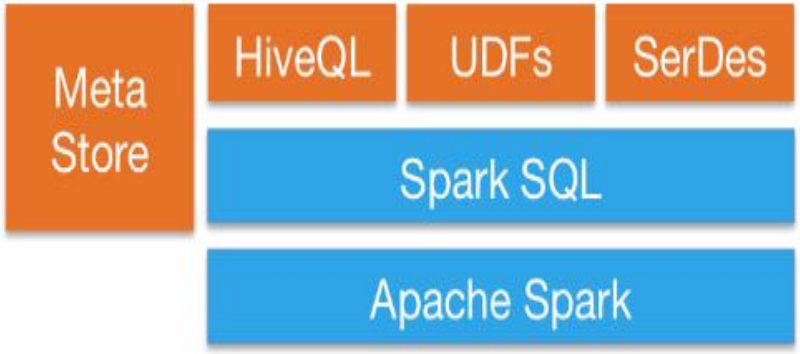
Spark SQL Internal Architecture



SPARK SQL HIVE INTEGRATION

Hive Compatibility

- Spark SQL reuses the Hive frontend and metastore, giving you full compatibility with existing Hive data, queries, and UDFs.



Spark SQL can use existing Hive metastores, SerDes, and UDFs.

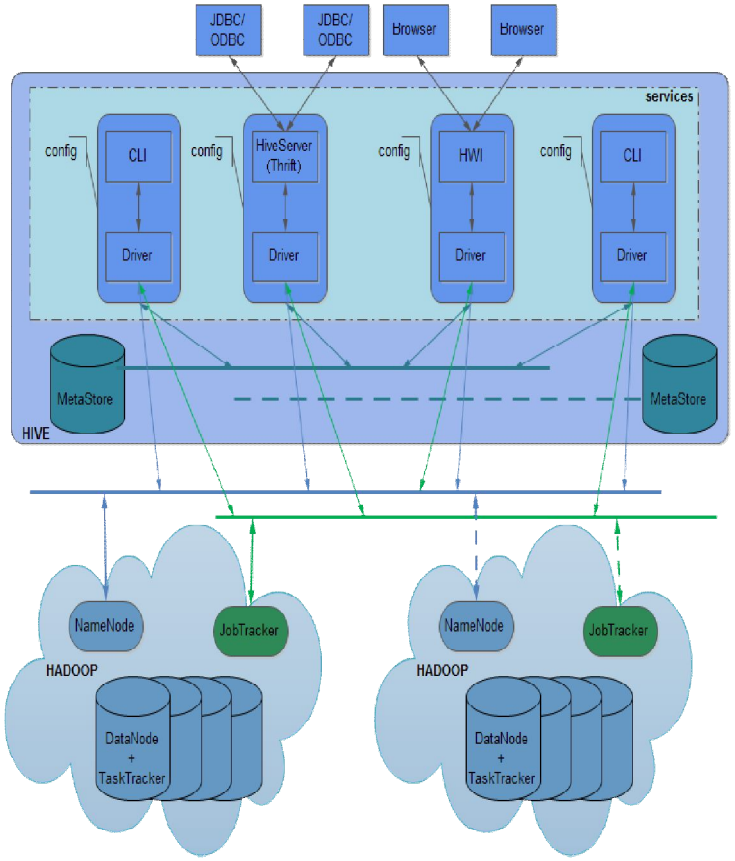
Hive

- A database/data warehouse on top of Hadoop
  - Rich data types
  - Efficient implementations of SQL on top of map reduce
- Support Analysis of large datasets stored in Hadoop's HDFS and compatible file

  systems
  - Such as Amazon S3 filesystem.
- Provides an SQL-like language called HiveQL with schema.

Hive Architecture

- User issues SQL query
- Hive parses and plans query
- Query converted to Map-Reduce
- Map-Reduce runs by Hadoop



User-Defined Functions

- UDF: Plug in your own processing code and invoke it from a Hive query
  - UDF (Plain UDF)
    - Input: single row, Output: single row
  - UDAF (User-Defined Aggregate Function)
    - Input: multiple rows, Output: single row
    - e.g. COUNT and MAX
  - UDTF (User-Defined Table-generating Function)
    - Input: single row, Output: multiple rows

Thank You!