Introduction to Hive

Tushar B. Kute,
http://tusharkute.com

What is hive?

- Hive is a data warehouse infrastructure tool to process structured data in Hadoop.
- It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.
- Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive.
- It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

What is hive?

- Initially Hive is developed by Facebook and Amazon, Netflix and It delivers standard SQL functionality for analytics.
- Traditional SQL queries are written in the MapReduce Java API to execute SQL Application and SQL queries over distributed data.
- Hive provides portability as most data warehousing applications functions with SQL-based query languages like NoSQL.

Hive is not-

- A relational database
- A design for OnLine Transaction Processing (OLTP)
- A language for real-time queries and row-level updates

Characteristics of Hive

- Databases and tables are built before loading the data.
- Hive as data warehouse is built to manage and query only structured data which is residing under tables.
- At the time of handling structured data, MapReduce lacks optimization and usability function such as UDFs whereas Hive framework have optimization and usability.

Characteristics of Hive

- Programming in Hadoop deals directly with the files. So, Hive can partition the data with directory structures to improve performance on certain queries.
- Hive is compatible for the various file formats which are TEXTFILE, SEQUENCEFILE, ORC, RCFILE, etc.
- Hive uses derby database in single user metadata storage and it uses MYSQL for multiple user Metadata or shared Metadata.

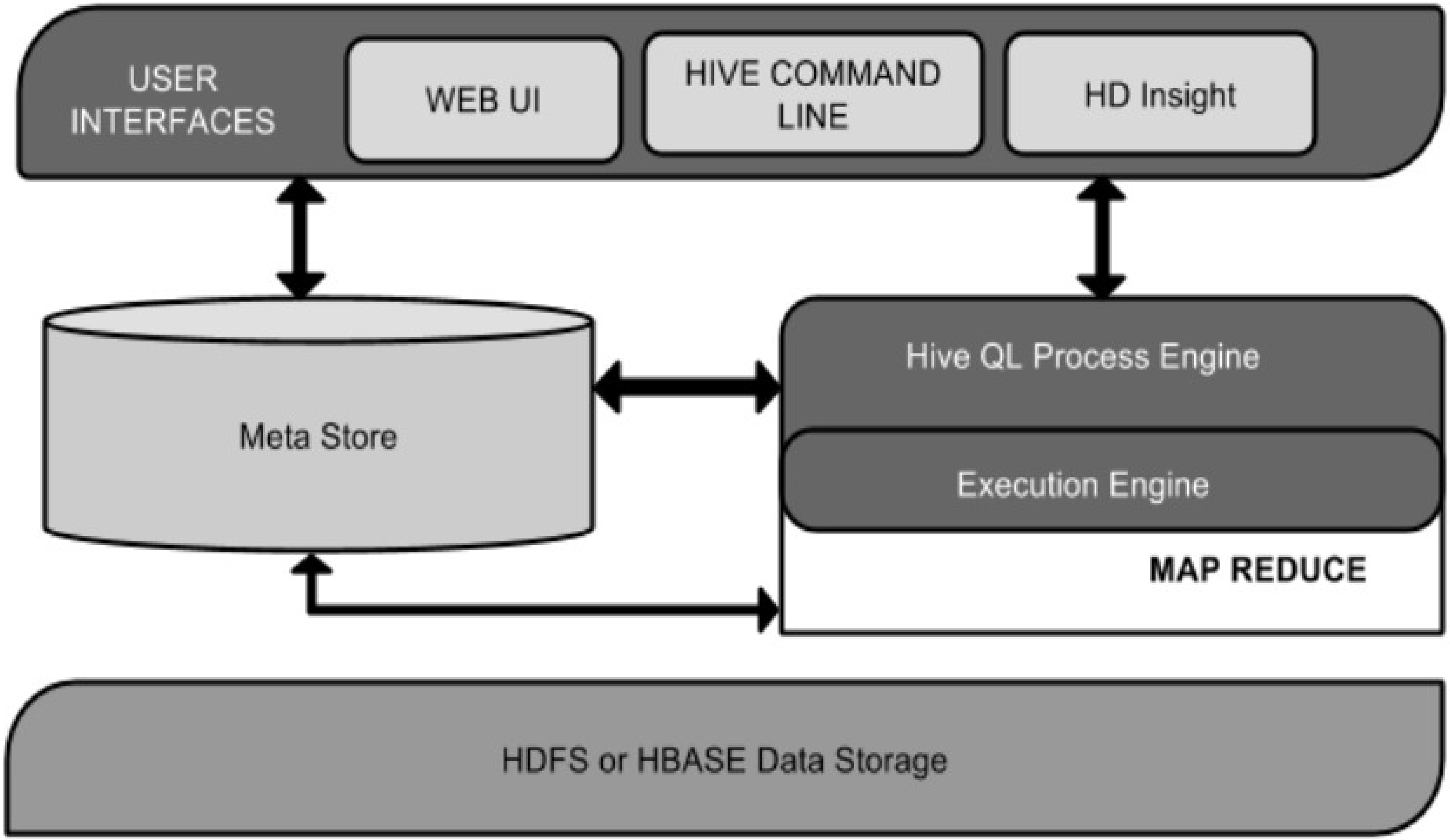
Features of Hive

- It provides indexes, including bitmap indexes to accelerate the queries. Index type containing compaction and bitmap index as of 0.10.
- Metadata storage in a RDBMS, reduces the time to function semantic checks during query execution.
- Built in user-defined functions (UDFs) to manipulation of strings, dates, and other data-mining tools. Hive is reinforced to extend the UDF set to deal with the use-cases not reinforced by predefined functions.

Features of Hive

- DEFLATE, BWT, snappy, etc are the algorithms to operation on compressed data which is stored in Hadoop Ecosystem.
- It stores schemas in a database and processes the data into the Hadoop File Distributed File System (HDFS).
- It is built for Online Analytical Processing (OLAP).
- It delivers various types of querying language which are frequently known as Hive Query Language (HVL or HiveQL).

Hive Architecture



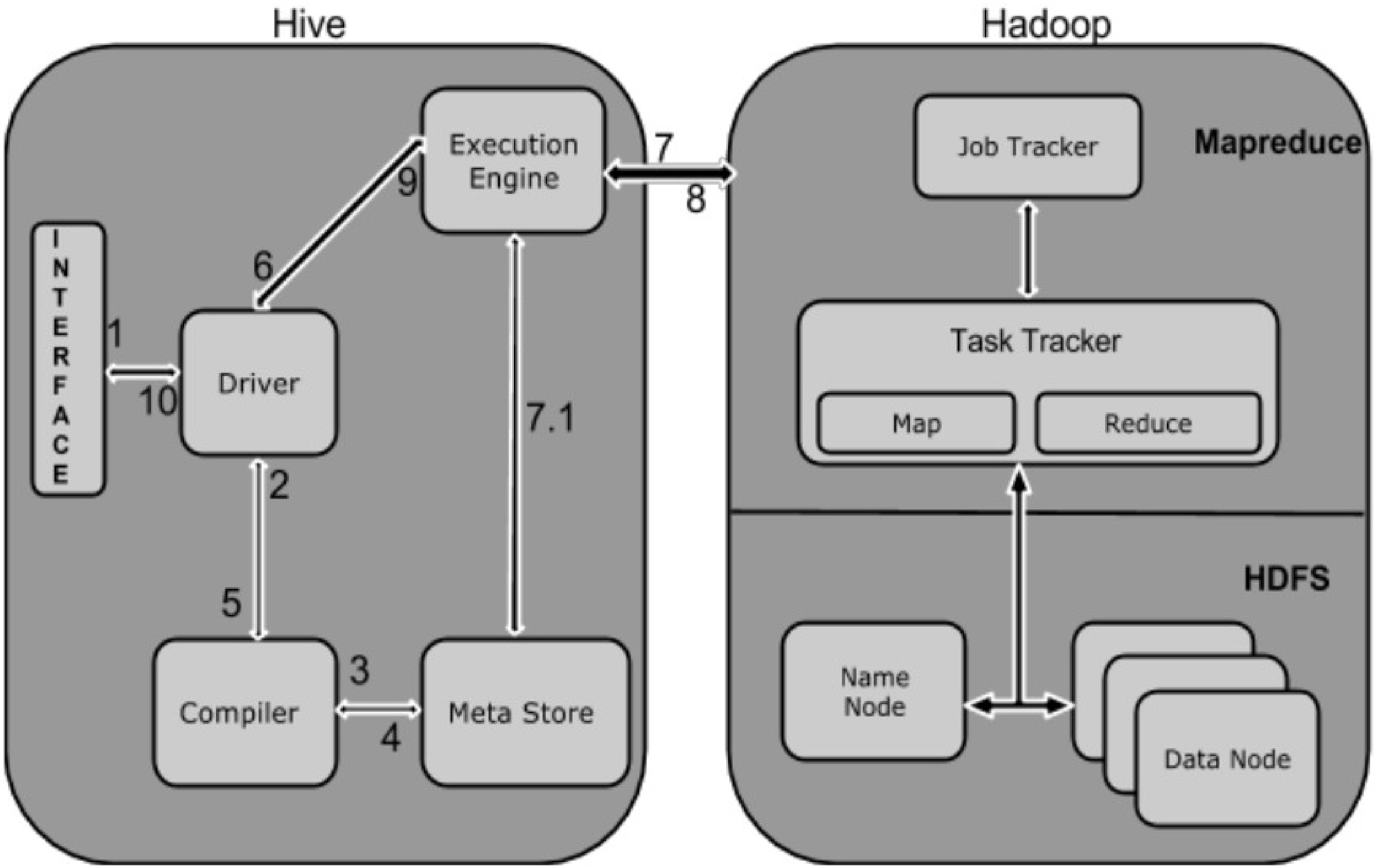
Hive Architecture

- **User Interface**
  - Hive is a data warehouse infrastructure software that can create interaction between user and HDFS. The user interfaces that Hive supports are Hive Web UI, Hive command line, and Hive HD Insight (In Windows server).
- **Meta Store**
  - Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping.

Hive Architecture

- **HiveQL Process Engine**
  - HiveQL is similar to SQL for querying on schema info on the Metastore. It is one of the replacements of traditional approach for MapReduce program. Instead of writing MapReduce program in Java, we can write a query for MapReduce job and process it.
- **Execution Engine**
  - The conjunction part of HiveQL process Engine and MapReduce is Hive Execution Engine. Execution engine processes the query and generates results as same as MapReduce results. It uses the flavor of MapReduce.
- **HDFS or HBASE**
  - Hadoop distributed file system or HBASE are the data storage techniques to store data into file system.

Working of Hive



Execution of Hive

1. **Execute Query**

   The Hive interface such as Command Line or Web UI sends query to Driver (any database driver such as JDBC, ODBC, etc.) to execute.

2. **Get Plan**

   The driver takes the help of query compiler that parses the query to check the syntax and query plan or the requirement of query.

3. **Get Metadata**

   The compiler sends metadata request to Metastore (any database).

4. **Send Metadata**

   Metastore sends metadata as a response to the compiler.

Execution of Hive

5 Send Plan

The compiler checks the requirement and resends the plan to the driver. Up to here, the parsing and compiling of a query is complete.

6 Execute Plan

The driver sends the execute plan to the execution engine.

7 Execute Job

Internally, the process of execution job is a MapReduce job. The execution engine sends the job to JobTracker, which is in Name node and it assigns this job to TaskTracker, which is in Data node. Here, the query executes MapReduce job.

Execution of Hive

7.1 Metadata Ops

Meanwhile in execution, the execution engine can execute metadata operations with Metastore.

8 Fetch Result

The execution engine receives the results from Data nodes.

9 Send Results

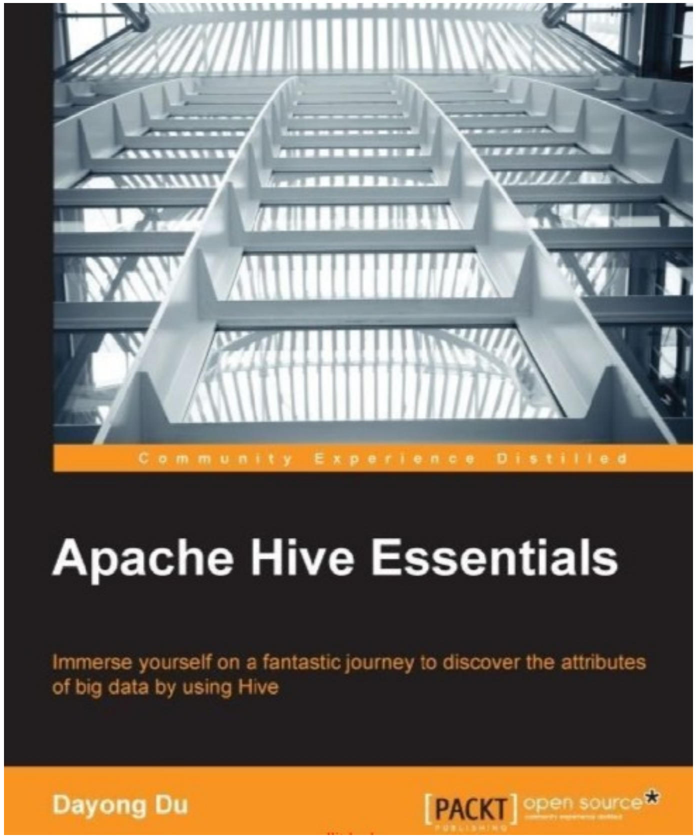
The execution engine sends those resultant values to the driver.

10 Send Results

The driver sends the results to Hive Interfaces.

References

Data Warehouse and Query Language for Hadoop
Programming
Hive
O'REILLY®
Jason Rutherglen,
Dean Wampler &
Edward Capriolo


Community Experience Distilled
Apache Hive Essentials
Immerse yourself on a fantastic journey to discover the attributes of big data by using Hive
Dayong Du
[PACKT] open source*

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com