Introduction to Hive

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

Big data and Hadoop

- The term ‘Big Data’ is used for collections of large datasets that include huge volume, high velocity, and a variety of data that is increasing day by day.
- Using traditional data management systems, it is difficult to process Big Data. Therefore, the Apache Software Foundation introduced a framework called Hadoop to solve Big Data management and processing challenges.

Hadoop

- Hadoop is an open-source framework to store and process Big Data in a distributed environment. It contains two modules, one is MapReduce and another is Hadoop Distributed File System (HDFS).
  - MapReduce: It is a parallel programming model for processing large amounts of structured, semi-structured, and unstructured data on large clusters of commodity hardware.
  - HDFS: Hadoop Distributed File System is a part of Hadoop framework, used to store and process the datasets. It provides a fault-tolerant file system to run on commodity hardware.

Hadoop Tools

- The Hadoop ecosystem contains different sub-projects (tools) such as Sqoop, Pig, and Hive that are used to help Hadoop modules.
  - **Sqoop**: It is used to import and export data to and fro between HDFS and RDBMS.
  - **Pig**: It is a procedural language platform used to develop a script for MapReduce operations.
  - **Hive**: It is a platform used to develop SQL type scripts to do MapReduce operations.

Ways to execute MapReduce

- The traditional approach using Java MapReduce program for structured, semi-structured, and unstructured data.
- The scripting approach for MapReduce to process structured and semi structured data using Pig.
- The Hive Query Language (HiveQL or HQL) for MapReduce to process structured data using Hive.

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.

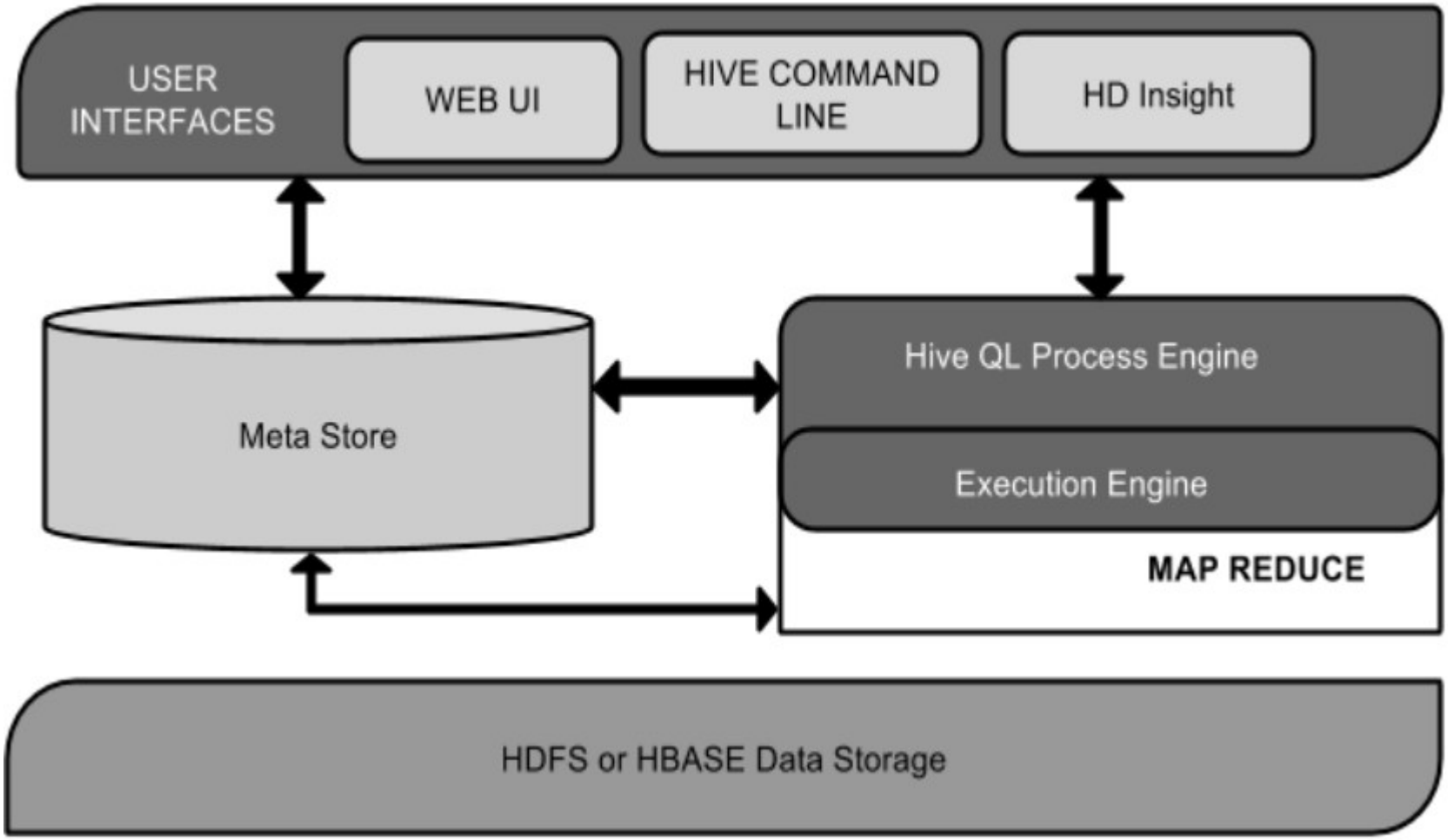
Hive is not-

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

Features of Hive

- It stores schema in a database and processed data into HDFS.
- It is designed for OLAP.
- It provides SQL type language for querying called HiveQL or HQL.
- It is familiar, fast, scalable, and extensible.

Hive Architecture



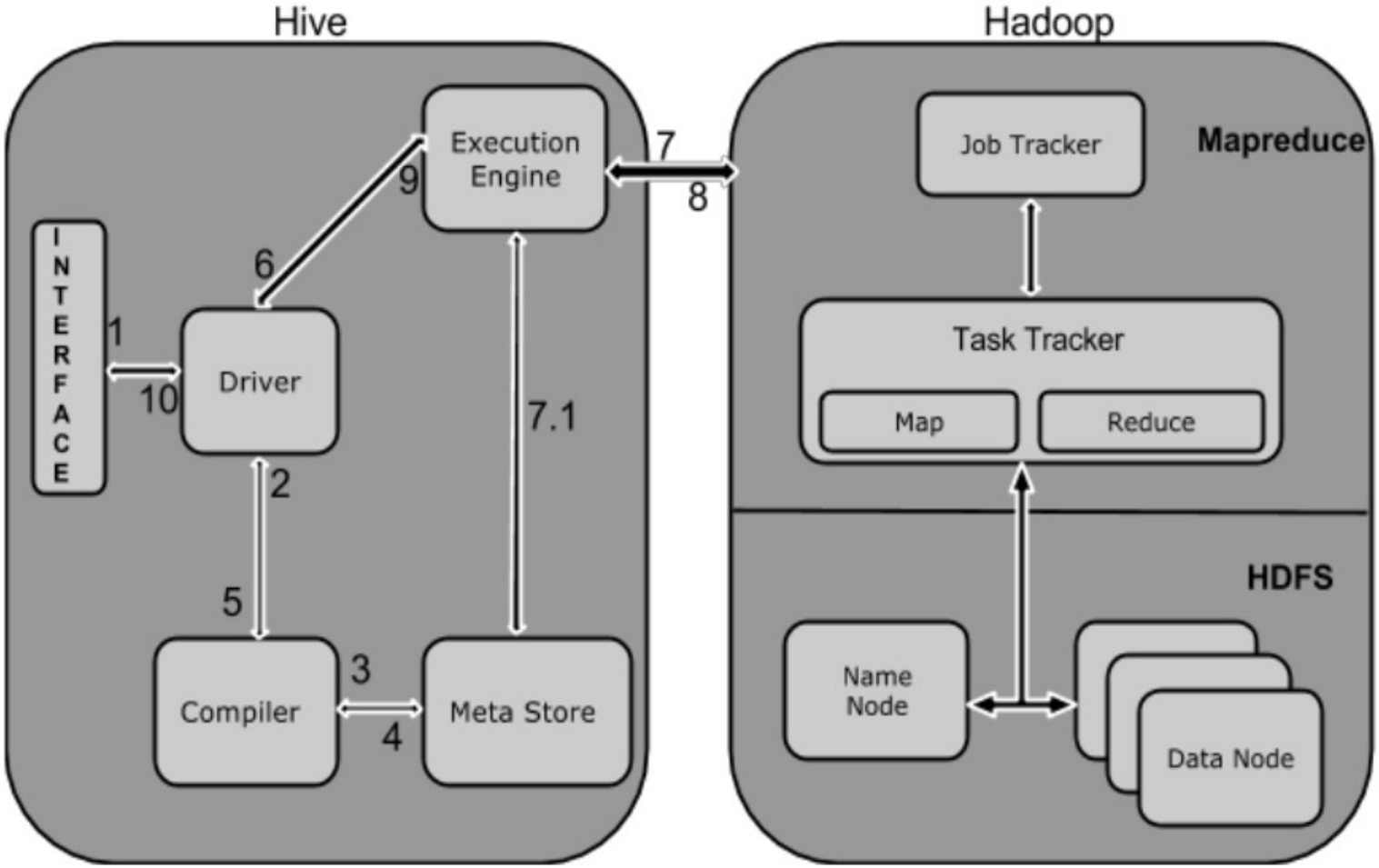
Hive Architecture

- **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

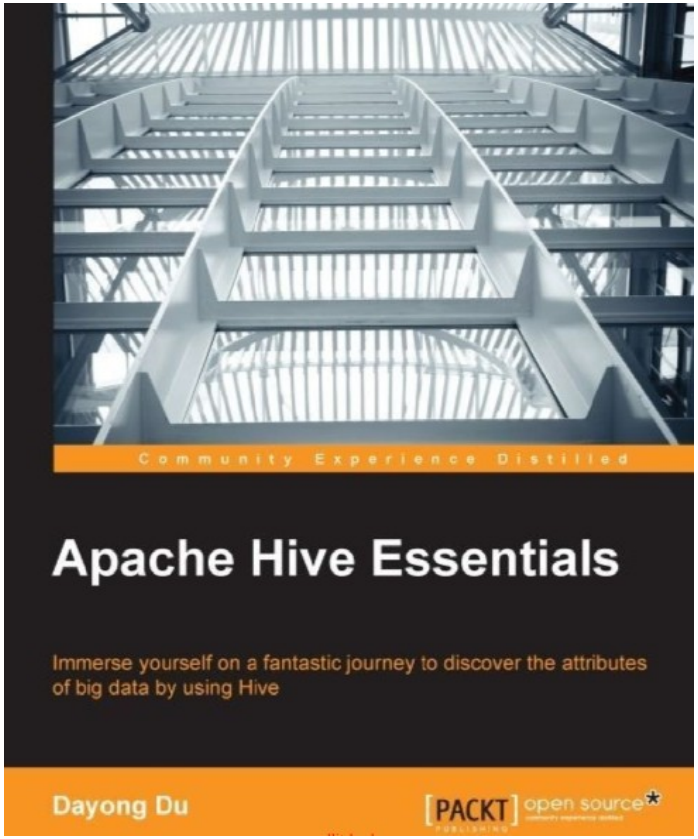
The execution engine sends those resultant values to the driver.

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

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

Blogs
http://digitallocha.blogspot.in
http://kyamputar.blogspot.in

tushar@tusharkute.com