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.

It is a platform used to develop SQL type scripts to do MapReduce operations.

Features of Hive

* Iet 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 Part:

1. User Interface
2. 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
3. Hive QL Processing 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.
4. 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.
5. HDFS or HBASE : Hadoop distributed file system or HBASE are the data storage techniques to store data into file system

Hive with MySQL:

In hive-config.sh : we need to provide the hadoop home

1. Put my-sql connector in lib of hive.
2. Create database in mysql named metastore.
3. SOURCE $HIVE\_HOME/scripts/metastore/upgrade/mysql/hive-schema-0.14.0.mysql.sql;
4. CREATE USER 'hiveuser'@'%' IDENTIFIED BY 'hivepassword';
5. GRANT all on \*.\* to 'hiveuser'@localhost identified by 'hivepassword';
6. Open hive site.xml
7. Provide mysql username/ password/ driver and connector name.

To check tables in mysql we use below command.

mysql> select \* from TBLS;

DataTypes :

1. Integral Types:

INT like 10, TINYINT like 10Y, SMALLINT like 10S, BIGINT like 10L

1. String Types : char, varchar
2. TimeStamp
3. Dates
4. Decimals – decimal(10,0)
5. Union Types : UNIONTYPE<int, double, array<string>, struct<a:int,b:string>>
6. Array - ARRAY<data\_type>
7. Maps are similar to Java Types :
   1. MAP<primitive\_type, data\_type>

Creating table :

CREATE TABLE IF NOT EXISTS employee ( eid int, name String, salary String, destination String)

COMMENT ‘Employee details’

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ‘\t’

LINES TERMINATED BY ‘\n’

STORED AS TEXTFILE;

Loading data:

LOAD DATA LOCAL INPATH '/home/user/sample.txt'

OVERWRITE INTO TABLE employee

Hive organizes tables into partitions. It is a way of dividing a table into related parts based on the values of partitioned columns such as date, city, and department. Using partition, it is easy to query a portion of the data.

Tables or partitions are sub-divided into **buckets,** to provide extra structure to the data that may be used for more efficient querying. Bucketing works based on the value of hash function of some column of a table.

ALTER TABLE employee

> ADD PARTITION (year=’2012’)

> location '/2012/part2012';

We can rename a partition as well as drop a partition.

There are built in function are also there.