# BIG DATA ANALYTICS FRAMEWORK

## Unit IV

## Hive and HBase

# Apache Hive

- ✓ It is a data warehouse framework for querying and managing large datasets stored in Hadoop distributed filesystems (HDFS)

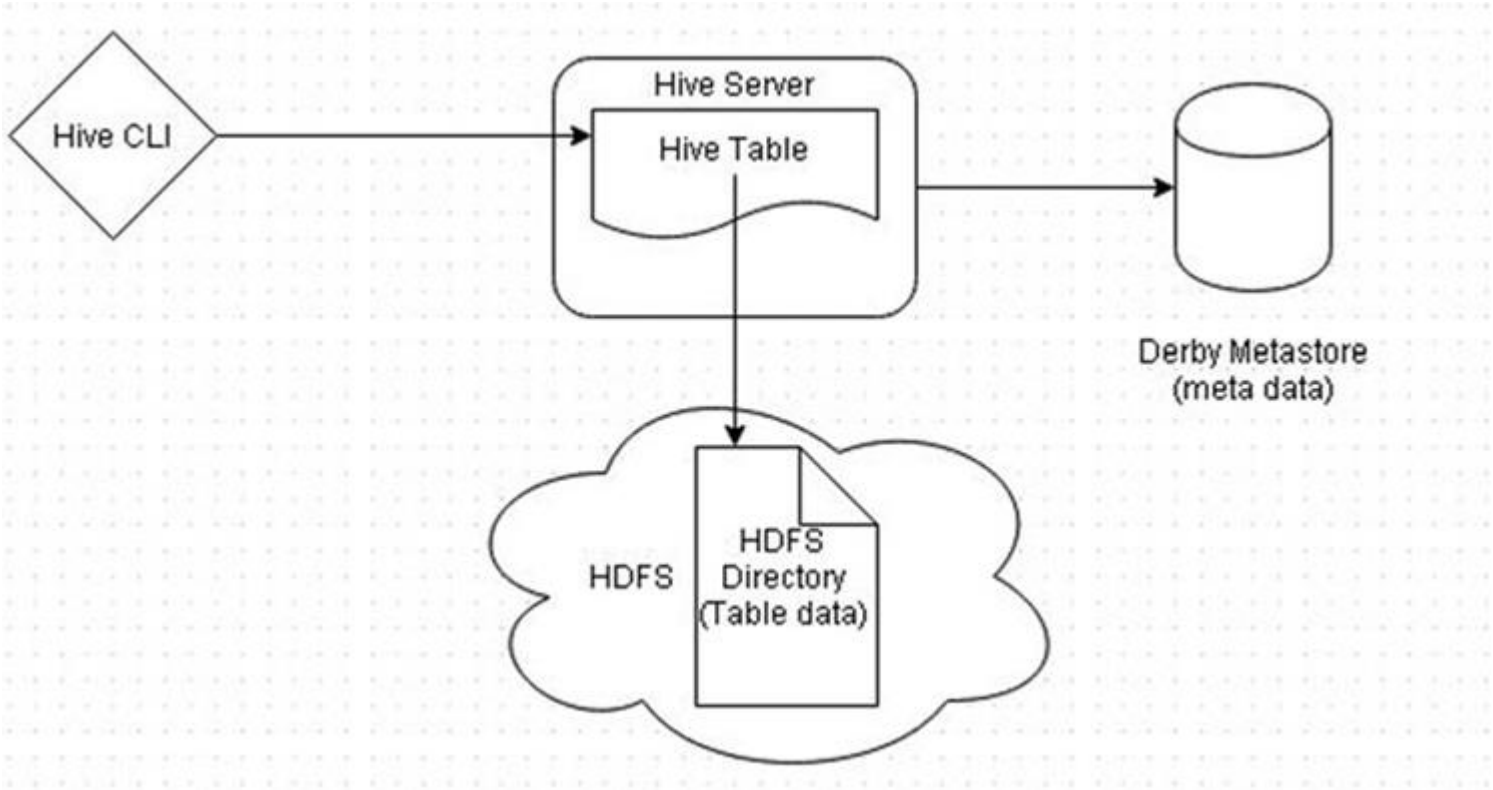
- ✓ Apache Hive is a data warehousing tool built on top of the Hadoop Distributed File System (HDFS) and MapReduce framework.

- ✓ It allows users to query and analyze large datasets stored in Hadoop without the need for complex programming.

# Apache Hive

- ✓ Hive also provides a SQL-like query language called HiveQL .
- ✓ The HiveQL queries may be run in the Hive CLI(Command Line Interface) shell .
- ✓ By default, Hive stores data in the HDFS
- ✓ Hive stores data in tables.
- ✓ A Hive table is an abstraction and the metadata for a Hive table is stored in an embedded Derby database called a Derby metastore.
- ✓ Other databases such as MySQL and Oracle
- ✓ Database could also be configured as the Hive metastore

# Apache Hive  architecture



# Setting the Environment

- ✓ The following software is required.
- ✓ Apache Hadoop
- ✓ Apache Hive
- ✓ J a v a 7

# Configuring Hive

- ✓ Apache Hive is configured in the hive-site.xml configuration file.
- ✓ Create the hive-site.xml file from the template file.

```
cp/cdh/hive-0.13.1-cdh5.3.6/conf/hive-default.xml.template /cdh/hive-0.13.1-cdh5.3.6/conf/hive-site.xml
```

- ✓ Open the hive-site.xml file in the vi editor.

```
vi /cdh/hive-0.13.1-cdh5.3.6/conf/hive-site.xml
```