Experiment No. 04

Aim: To create Hive Database and descriptive analytics basic statistics.

Theory:

HIVE

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). The Hadoop ecosystem contains different sub-projects (tools) such as Sqoop, Pig, and Hive that are used to help Hadoop modules.

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.

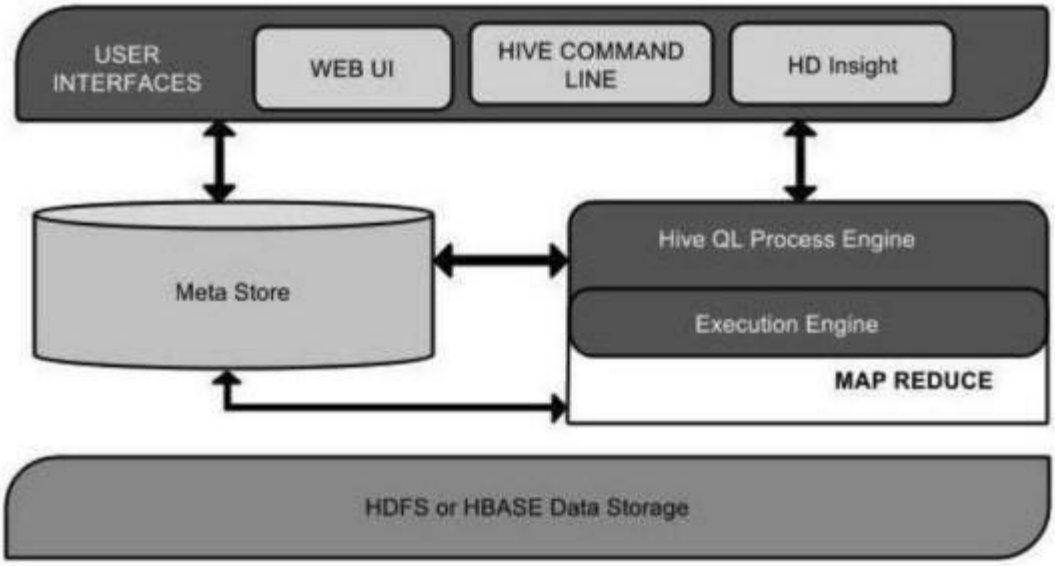
It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce. Hive provides the functionality of reading, writing, and managing large datasets residing in distributed storage. It runs SQL-like queries called HQL (Hive query language) which get internally converted to MapReduce jobs.

Using Hive, we can skip the requirement of the traditional approach of writing complex MapReduce programs. Hive supports Data Definition Language (DDL), Data Manipulation Language (DML), and User Defined Functions (UDF).

Features of Hive

- It stores schema in a database and processes 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

Architecture of Hive



Limitations of Hive commands:

1.Hive doesn't support sub queries.

2.Hive surely supports over-writing, but unfortunately, it doesn't support deletion and updates.

3.Hive is not designed for OLTP, but it is used for it.

Basic Hive Commands

1.Create: This will create the new database in the Hive.

create database database_name;

2.Show: show command will show all the databases residing in the Hive.

show databases;

3.Use: The command to use the database.

use database_name;

4.Drop: The drop will remove a table from Hive

Drop table_name;

5.Create table: This command creates the table.

create table table_name (column names and types);

6.Alter: Alter command will help you rename the table or table columns.

alter table table_name rename to new_table_name;

For renaming column name, replace table name with column name.

7. Describe: Describe command will help you with the information about the schema of the table.

describe table name;

8.LOAD, INSERT: The Load operation is used to move the data into corresponding Hive table.

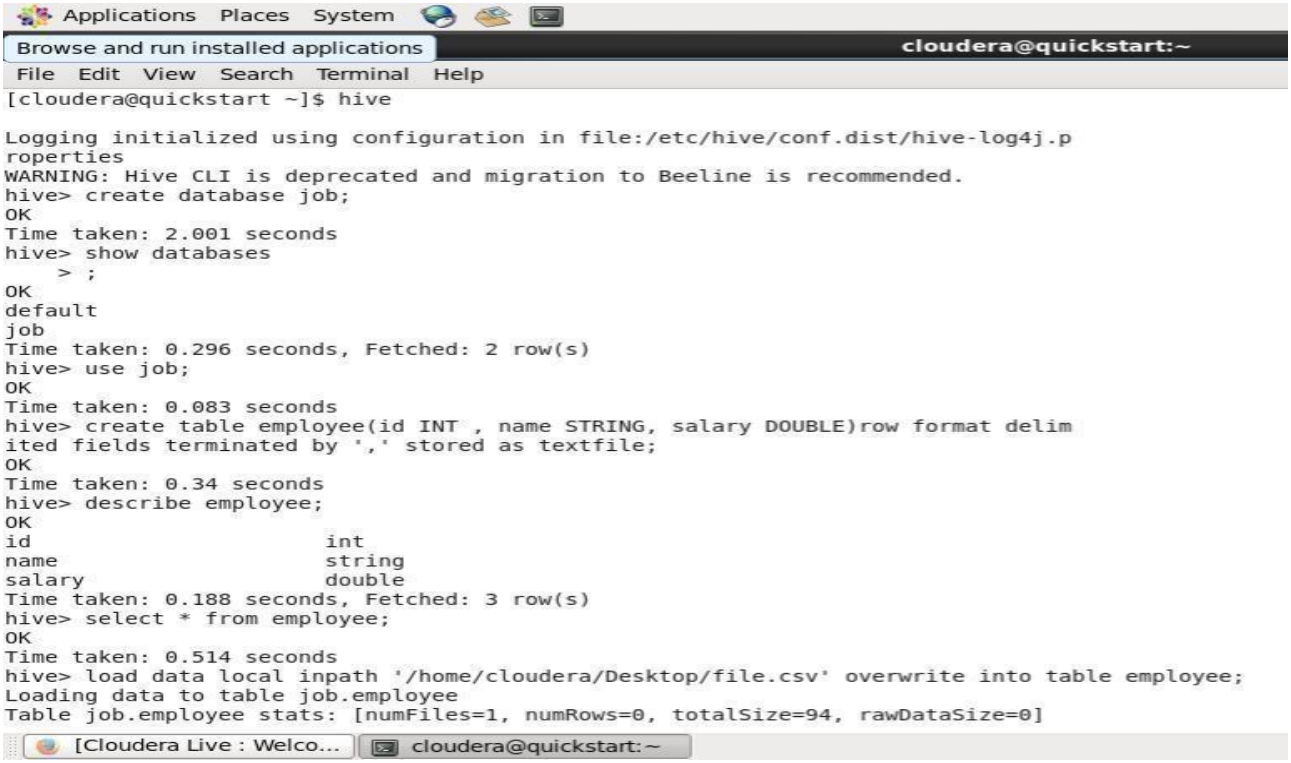
LOAD data inpath into table [tablename];

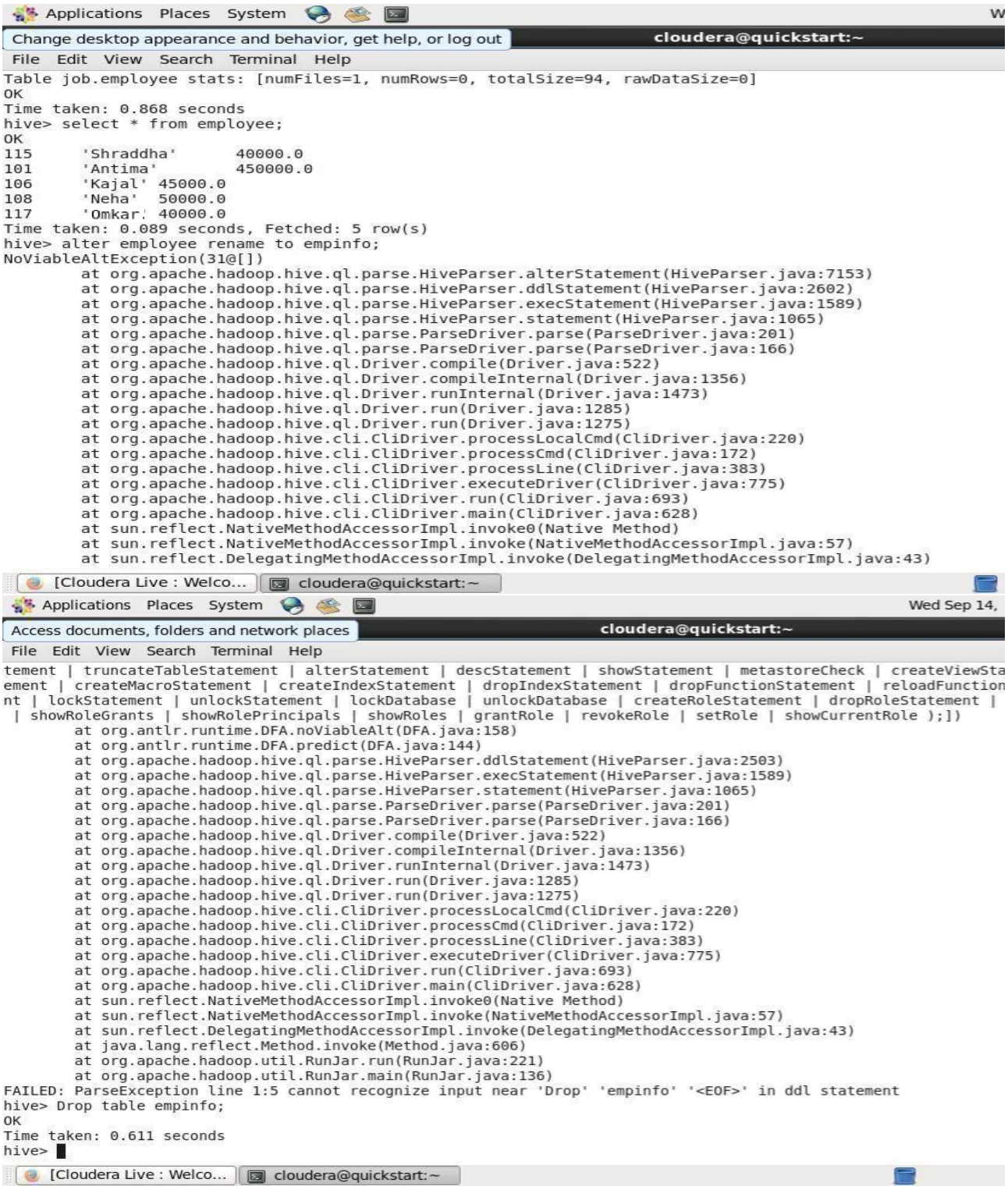
Sr. No.	Description	Commands
1	**Start HIVE**	$ hive
	o/p	Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
2	**Create database**	hive> create database retail;
	o/p	OK Time taken: 0.709 seconds
3	**Show databases**	hive> show databases
	o/p	OK Time taken: 0.284 seconds, Fetched: 2 row(s)
4	**Use database**	hive> use retail;
	o/p	OK Time taken: 0.087 seconds
5	**Create table**	hive> create table txnrecords(txnno INT, txndate STRING, custno INT,amount DOUBLE,category STRING, product STRING, state STRING,spendby STRING)row format delimited fields terminated by ',' stored as textfile;

	o/p	OK Time taken: 0.418 seconds
6	**Describe table**	hive> use retail; hive> describe txnrecords;
	o/p	OK txnno int txndate string custno int amount double category string product string state string spendby string Time taken: 0.165 second s, Fetched: 8 ow(s)
7	**Create table**	hive> use retail; hive> create table student (id INT, name STRING) row format delimited fields terminated by ',’;
	o/p	OK Time taken: 0.081 seconds
8	**Loading data**	hive> load data local inpath '/home/cloudera/Desktop/BE_A.csv' overwrite into table student;
	o/p	Loading data to table retail.student Table retail.student stats: [numFiles=1, numRows=0, totalSize=26, rawDataSize=0]
		OK Time taken: 0.887 seconds
9	**Fetching rows all**	hive> select * from student;

	o/p	OK name NULL 1 XYZABC 2 LMN Time taken: 0.516 seconds, Fetched: 4 row(s)

Output:





Conclusion:

Thus, the hive database was created and basic descriptive statistical analysis was performed on it.