# PRACTICAL NO – 4

**Aim:** Implementing Database Operations on Hive.

**THEORY:**

Hive defines a simple SQL-like query language to querying and managing large datasets called Hive-QL ( HQL ). It's easy to use if you're familiar with SQL Language. Hive allows programmers who are familiar with the language to write the custom MapReduce framework to perform more sophisticated analysis.

**Uses of Hive:**

1. The Apache Hive distributedstorage.
2. Hive provides tools to enable easy data extract/transform/load(ETL)
3. It provides the structure on a variety of dataformats.
4. By using Hive, we can access files stored in Hadoop Distributed File System (HDFS is used to querying and managing large datasets residing in) or in other data storage systems such as Apache HBase.

**Limitations of Hive:**

• Hive is not designed for Online transaction processing (OLTP ), it is only used for the Online AnalyticalProcessing.

• Hive supports overwriting or apprehending data, but not updates anddeletes.

• In Hive, sub queries are not supported.

**Why Hive is used inspite of Pig?**

The following are the reasons why Hive is used in spite of Pig's availability:

- Hive-QL is a declarative language line SQL, PigLatin is a data flowlanguage.
- Pig: a data-flow language and environment for exploring very largedatasets.
- Hive: a distributed data warehouse.

**Components of Hive:**

**Metastore :**

Hive stores the schema of the Hive tables in a Hive Metastore. Metastore is used to hold all the information about the tables and partitions that are in the warehouse. By default, the metastore is run in the same process as the Hive service and the default Metastore is DerBy Database.

**SerDe :**

Serializer, Deserializer gives instructions to hive on how to process a record.

**Hive Commands :**

**Data Definition Language (DDL )**

DDL statements are used to build and modify the tables and other objects in the database.

*Example :*

CREATE, DROP, TRUNCATE, ALTER, SHOW, DESCRIBE Statements.

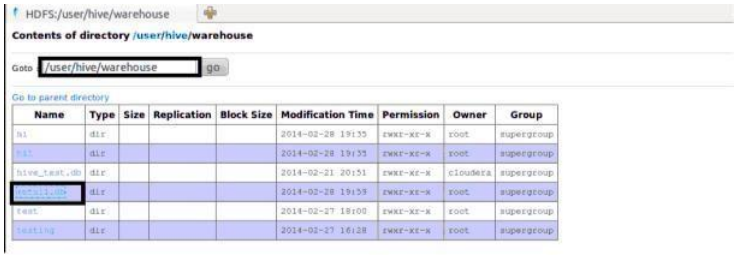
Go to Hive shell by giving the command sudo hive and enter the command **'create database<data base name>'** to create the new database in the Hive.

```
hive> create database retail;
OK
Time taken: 5.275 seconds
hive>
```

To list out the databases in Hive warehouse, enter the command '**show databases'.**

```
hive> show databases;
OK
default
retail
Time taken: 0.228 seconds
hive>
```

The database creates in a default location of the Hive warehouse. In Cloudera, Hive database store in a /user/hive/warehouse.




The command to use the database is **USE <data base name>**

```
hive> use retail;
OK
Time taken: 0.023 seconds
hive>
```

Copy the input data to HDFS from local by using the copy From Local command.

txns1.txt

```
00000000,06-26-2011,4007024,040.33,Exercise & Fitness,Cardio Machine Accessories,Clarksville,Tennessee,credit
00000001,05-26-2011,4006742,198.44,Exercise & Fitness,Weightlifting Gloves,Long Beach,California,credit
00000002,06-01-2011,4009775,005.58,Exercise & Fitness,Weightlifting Machine Accessories,Anaheim,California,credit
00000003,06-05-2011,4002199,198.19,Gymnastics,Gymnastics Rings,Milwaukee,Wisconsin,credit
00000004,12-17-2011,4002613,098.81,Team Sports,Field Hockey,Nashville ,Tennessee,credit
00000005,02-14-2011,4007591,193.63,Outdoor Recreation,Camping & Backpacking & Hiking,Chicago,Illinois,credit
00000006,10-28-2011,4002190,027.89,Puzzles,Jigsaw Puzzles,Charleston,South Carolina,credit
00000007,07-14-2011,4002964,096.01,Outdoor Play Equipment,Sandboxes,Columbus,Ohio,credit
00000008,01-17-2011,4007361,010.44,Winter Sports,Snowmobiling,Des Moines,Iowa,credit
00000009,05-17-2011,4004798,152.46,Jumping,Bungee Jumping,St. Petersburg,Florida,credit
```

```
cloudera@cloudera-vm:~$ hadoop dfs -copyFromLocal Desktop/blog/txns1.txt hdfs:/
cloudera@cloudera-vm:~$
```

When we create a table in hive, it creates in the default location of the hive warehouse. – "/user/hive/warehouse", after creation of the table we can move the data from HDFS to hive table.

The following command creates a table with in location of "/user/hive/warehouse/retail.db"

*Note :* retail.db is the database created in the Hive warehouse.

```
hive> create table txnrecords(txnno INT, txndate STRING, custno INT, amount DOUBLE,category STRING, product STRING, city STRIN
G, state STRING, spendby STRING) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 1.163 seconds
hive>
```

```
hive> describe txnrecords;
OK
txnno   int
txndate string
custno  int
amount  double
category        string
product string
city    string
state   string
spendby string
Time taken: 0.122 seconds
hive>
```

**Data Manipulation Language (DML )**

DML statements are used to retrieve, store, modify, delete, insert and update data in the database.

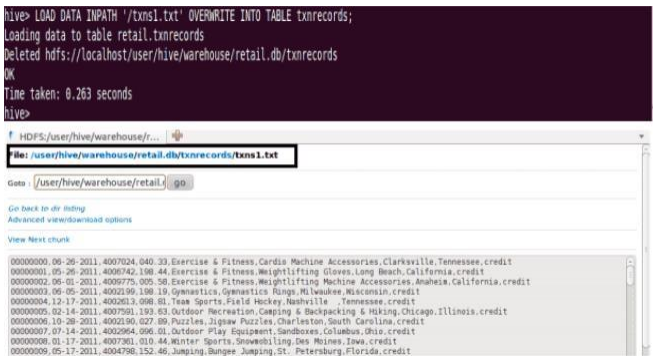
*Example :*

LOAD, INSERT Statements.

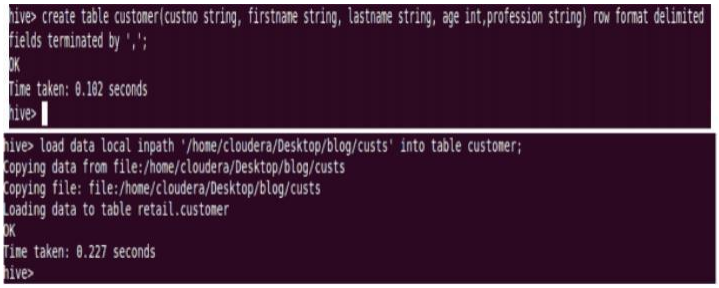
Syntax :

LOAD data <LOCAL> inpath <file path> into table [tablename]

The Load operation is used to move the data into corresponding Hive table. If the keyword **local** is specified, then in the load command will give the local file system path. If the keyword local is not specified we have to use the HDFS path of the file.



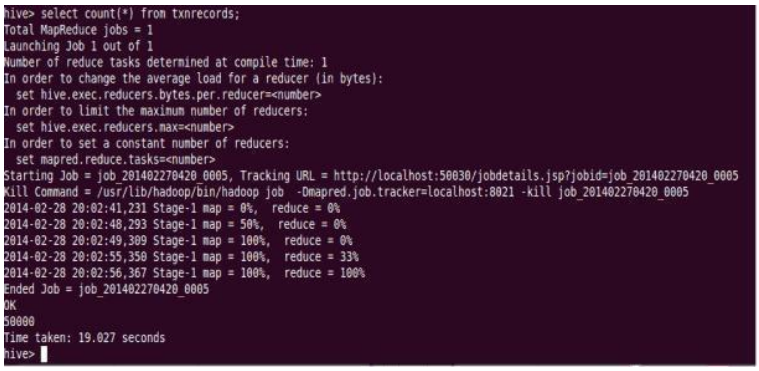

*Here are some examples for the LOAD data LOCAL command*




After loading the data into the Hive table we can apply the Data Manipulation Statements or aggregate functions retrieve the data.

*Example to count number of records:*

Count aggregate function is used count the total number of the records in a table.




**'create external' Table :**

The **create external** keyword is used to create a table and provides a location where the table will create, so that Hive does not use a default location for this table. An **EXTERNAL** table points to any HDFS location for its storage, rather than defaultstorage.

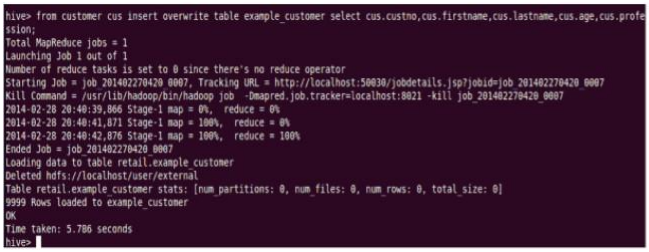
```
hive> create external table example_customer(custno string, firstname string, lastname string, age int,profession string) row
format delimited fields terminated by ',' LOCATION '/user/external';
OK
Time taken: 0.059 seconds
hive>
```


**Insert Command:**

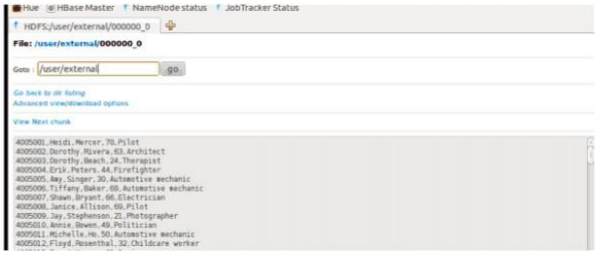
The **insert** command is used to load the data Hive table. Inserts can be done to a table or a partition.

• INSERT OVERWRITE is used to overwrite the existing data in the table orpartition.

• INSERT INTO is used to append the data into existing data in a table. (Note: INSERT INTO syntax is work from the version 0.8)



```
hive> from customer cus insert overwrite table example_customer select cus.custno,cus.firstname,cus.lastname,cus.age,cus.profe
ssion;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201402270420_0007, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201402270420_0007
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201402270420_0007
2014-02-28 20:40:39,866 Stage-1 map = 0%,  reduce = 0%
2014-02-28 20:40:41,871 Stage-1 map = 100%,  reduce = 0%
2014-02-28 20:40:42,876 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201402270420_0007
Loading data to table retail.example_customer
Deleted hdfs://localhost/user/external
Table retail.example_customer stats: [num_partitions: 0, num_files: 0, num_rows: 0, total_size: 0]
9999 Rows loaded to example_customer
OK
Time taken: 5.786 seconds
hive>
```




**Example for 'Partitioned By' and 'Clustered By' Command :**

'**Partitioned by**' is used to divided the table into the Partition and can be divided in to buckets by using the '**Clustered By**' command.

```
hive> create table txnrecsByCat(txnno INT, txndate STRING, custno INT, amount DOUBLE,product STRING, city STRING, state STRING
, spendby STRING) partitioned by (category STRING) clustered by (state) INTO 10 buckets row format delimited fields terminated
 by ',' stored as textfile;
OK
Time taken: 0.101 seconds
hive>
```

```
hive> from txnrecords txn INSERT OVERWRITE TABLE record PARTITION(category)select txn.txnno,txn.txndate,txn.custno,txn.amount,
txn.product,txn.city,txn.state,txn.spendby, txn.category;
FAILED: Error in semantic analysis: Dynamic partition strict mode requires at least one static partition column. To turn this
off set hive.exec.dynamic.partition.mode=nonstrict
```

When we insert the data Hive throwing errors, the dynamic partition mode is strict and dynamic partition not enabled (by Jeff at dresshead website). So we need to set the following parameters in Hiveshell.