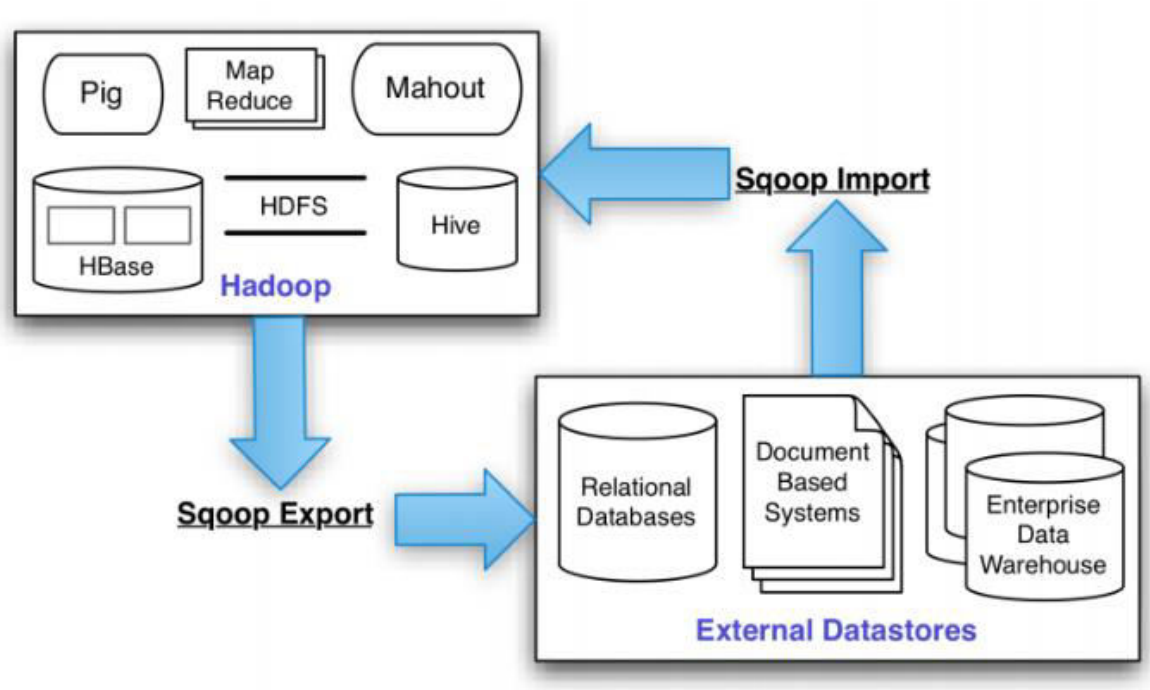
Documentation of Sqoop

-Prepared by Vignesh.R (15CSE107)

Abstract:

Sqoop is a command-line interface application for transferring data between relational databases and Hadoop. It supports incremental loads of a single table or a free form SQL query as well as saved jobs which can be run multiple times to import updates made to a database since the last import. Exports can be used to put data from Hadoop into a relational database. Sqoop got the name from sql+hadoop. Sqoop became a top-level Apache project in March 2012.



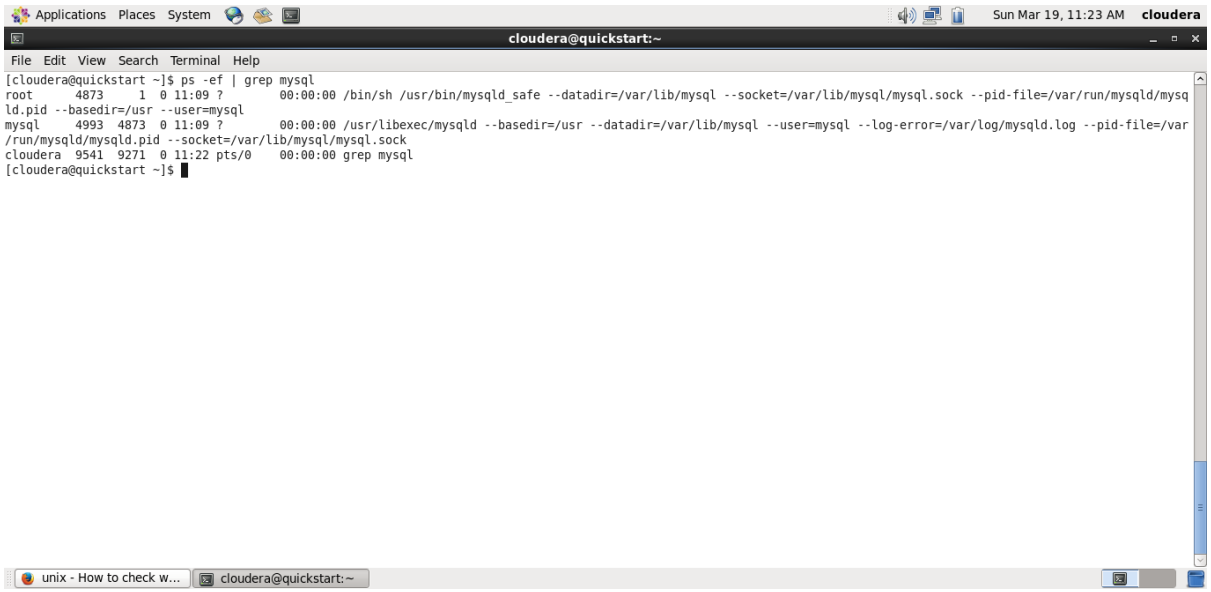
Sqoop is a tool designed to transfer data between Hadoop and relational databases. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.

Creating a table in SQL:-

Step 1:

Open a new terminal and check whether SQL is present in the cloudera.

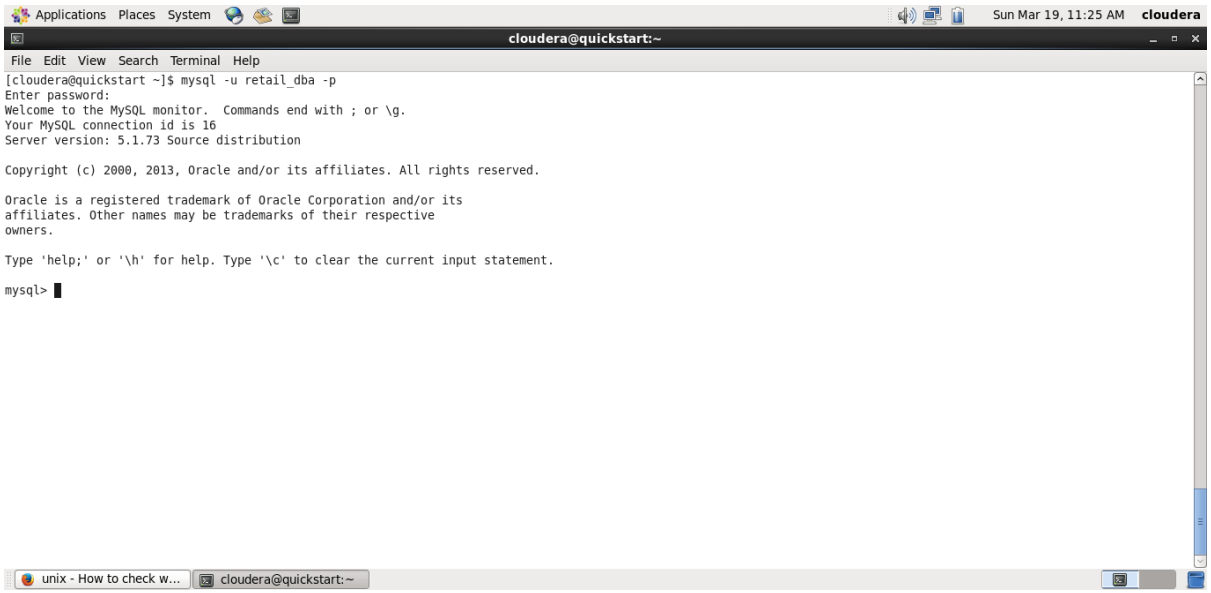
Command: ps -ef | grep mysql



Step 2:

A new database is created using the following command. Terminal will prompt for the password. Give the password as ‘cloudera’ which is not displayed on typing.

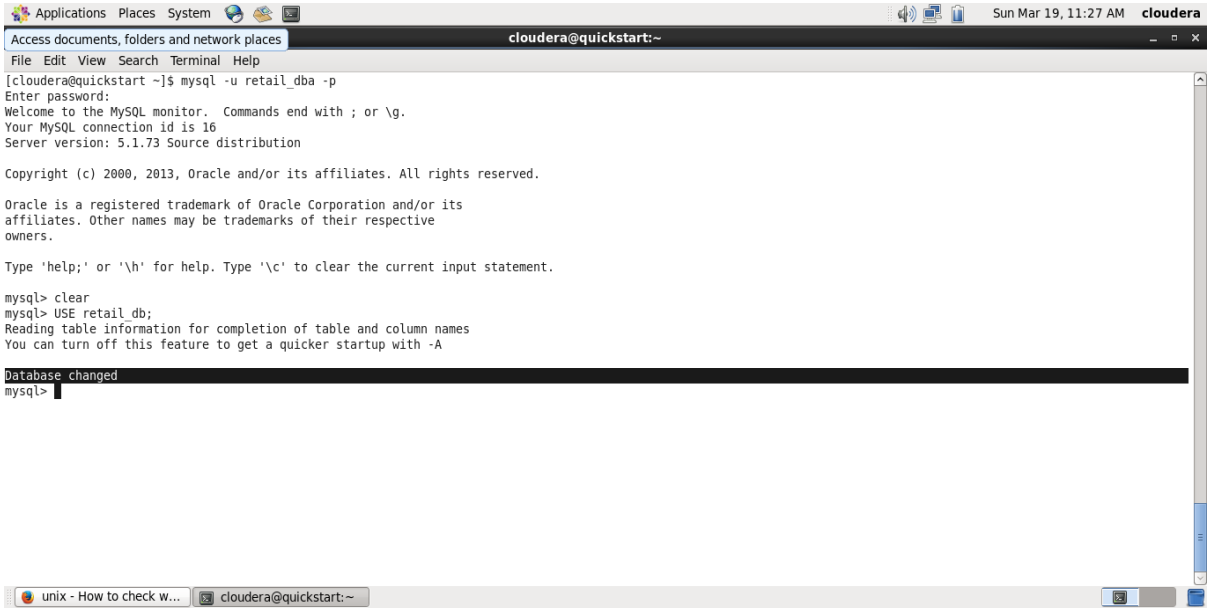
Command: mysql -u retail_dba –p



Step 3:

Then the database is changed using the following command.

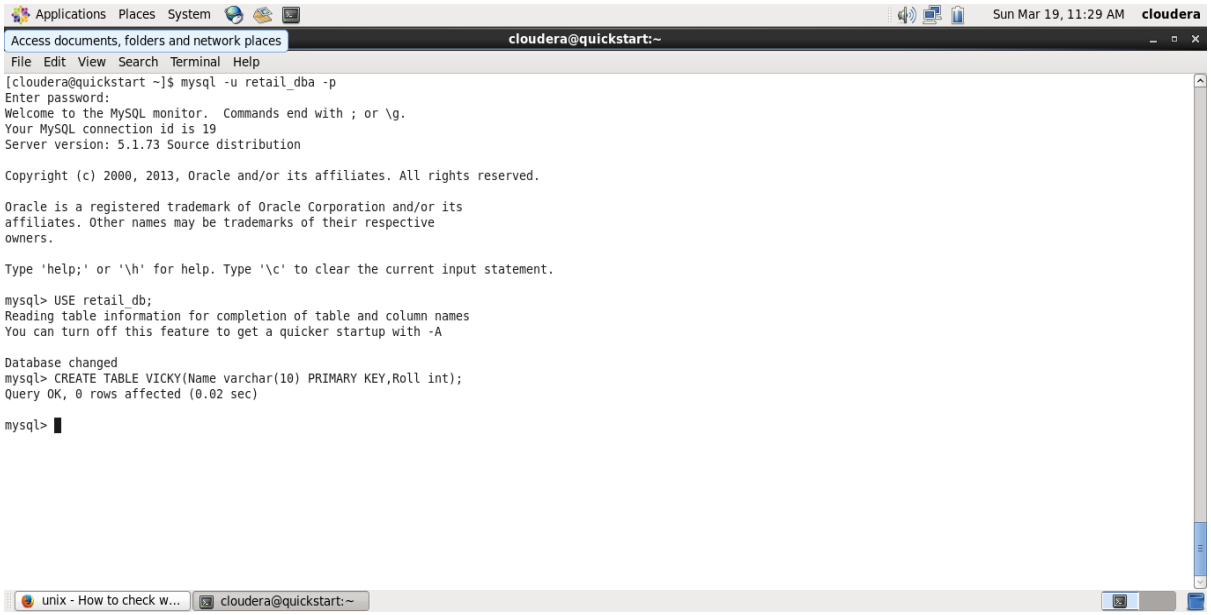
Command: USE retail_db;



Step 4:

Create a new table using normal SQL command. Primary key is mandatory for the table.

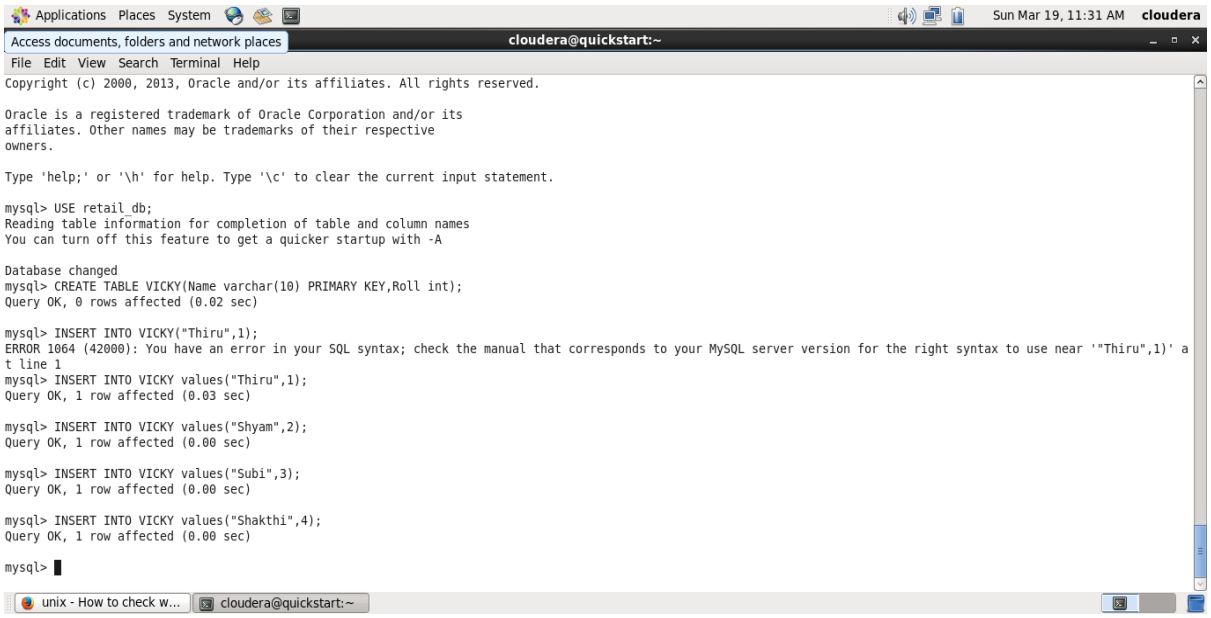
Command: CREATE TABLE <*table_name*> (*column1 datatype constraint*, *column2 datatype*, *column3 datatype*,....);



Step 5:

Insert some data in the created table.

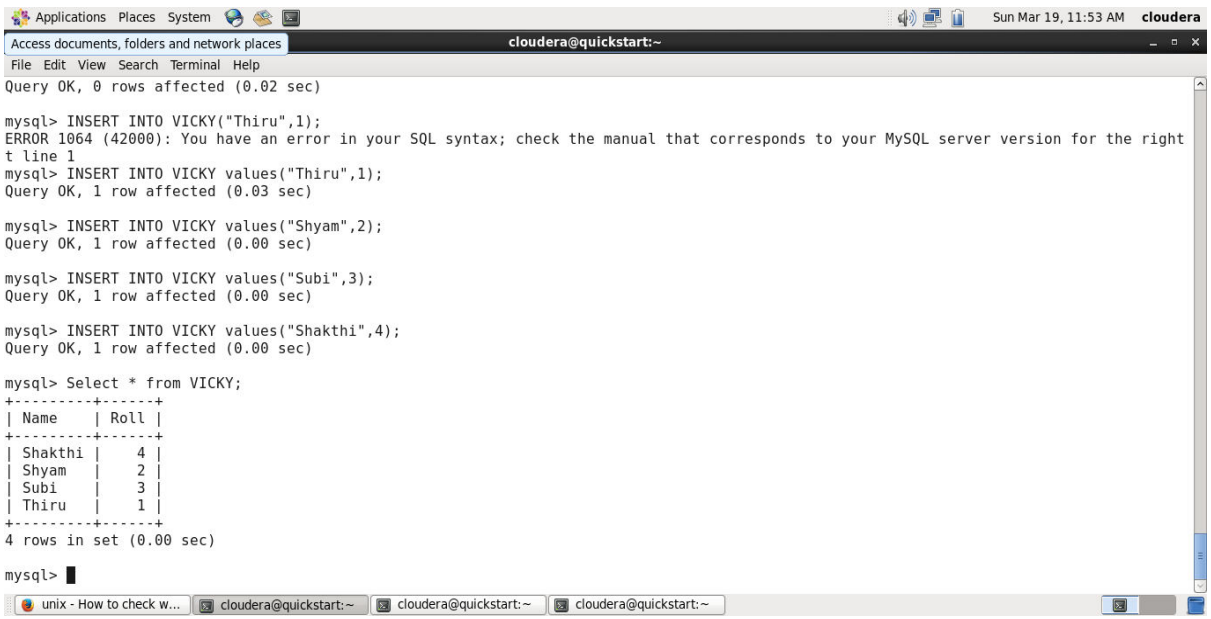
Command: INSERT INTO <*table_name*> VALUES (*value1*, *value2*, *value3*, ...);



Step 6:

To display the data in the table.

Command: Select * from <table_name>;



Importing a table from SQL to Hadoop:-

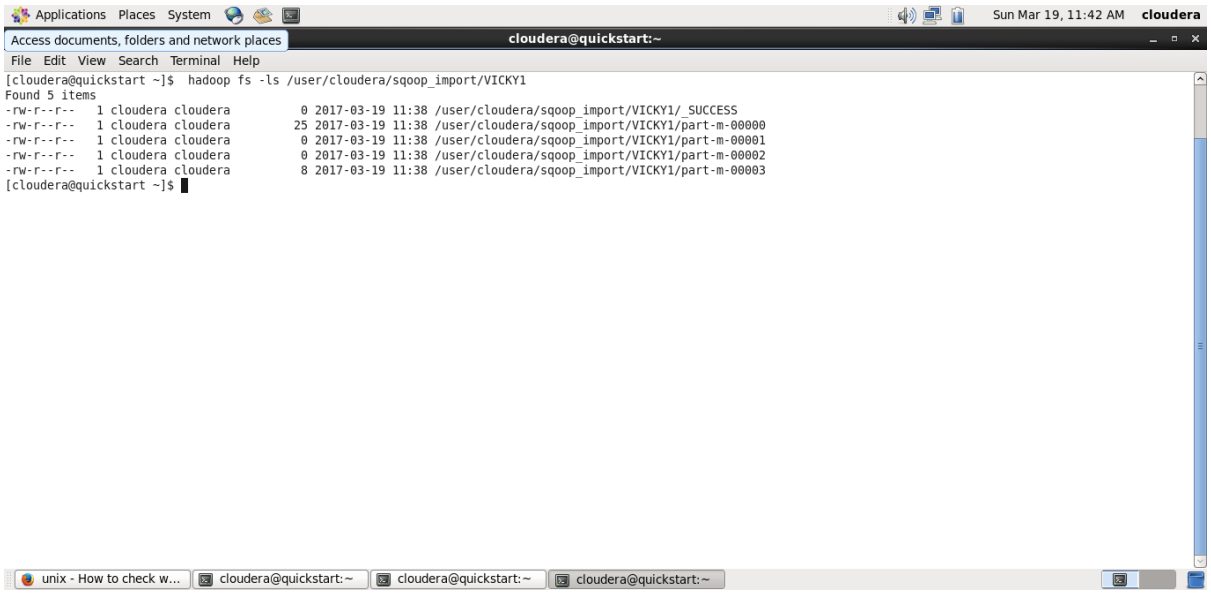
Step 7: Open a new terminal and connect to the database. Also give the target directory for the SQL table to be imported.

Command: sqoop import --connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" --username=retail_dba --password=cloudera --table <table_name> --as-textfile --target-dir=<target_directory>

Now the created table is imported to the target directory.

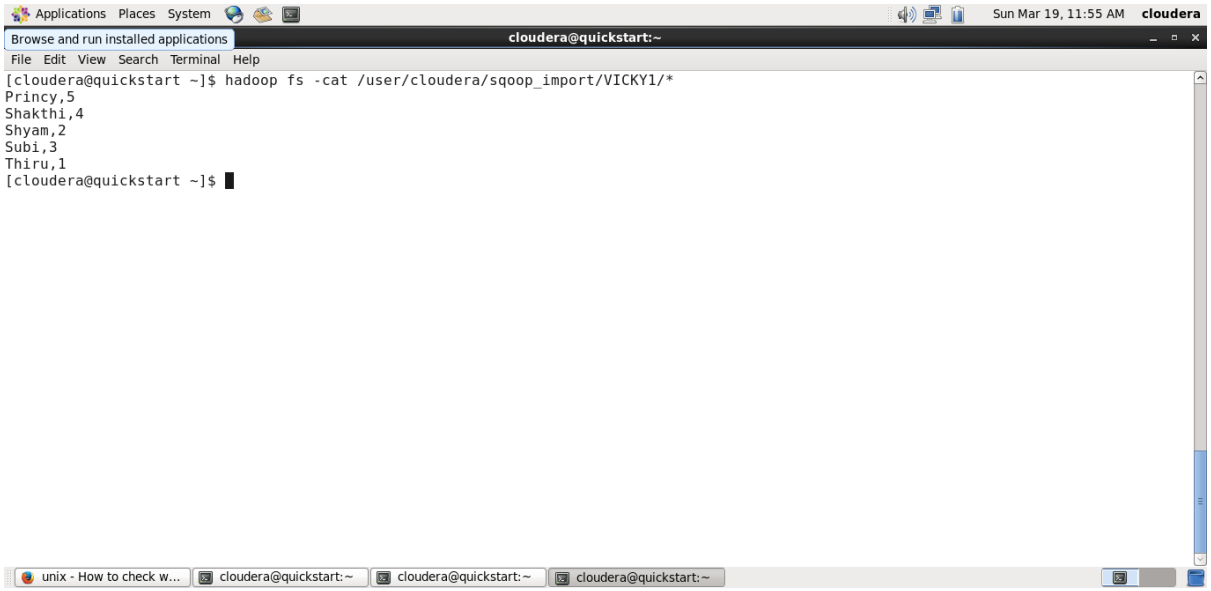
Step 8: Open a new terminal and check whether the table has been imported.

Command: hadoop fs -ls <target_directory>



Step 9: To display the records in the imported table.

Command: hadoop fs -cat <target_directory>/*



Thus the data table is successfully imported from the SQL database to HDFS.

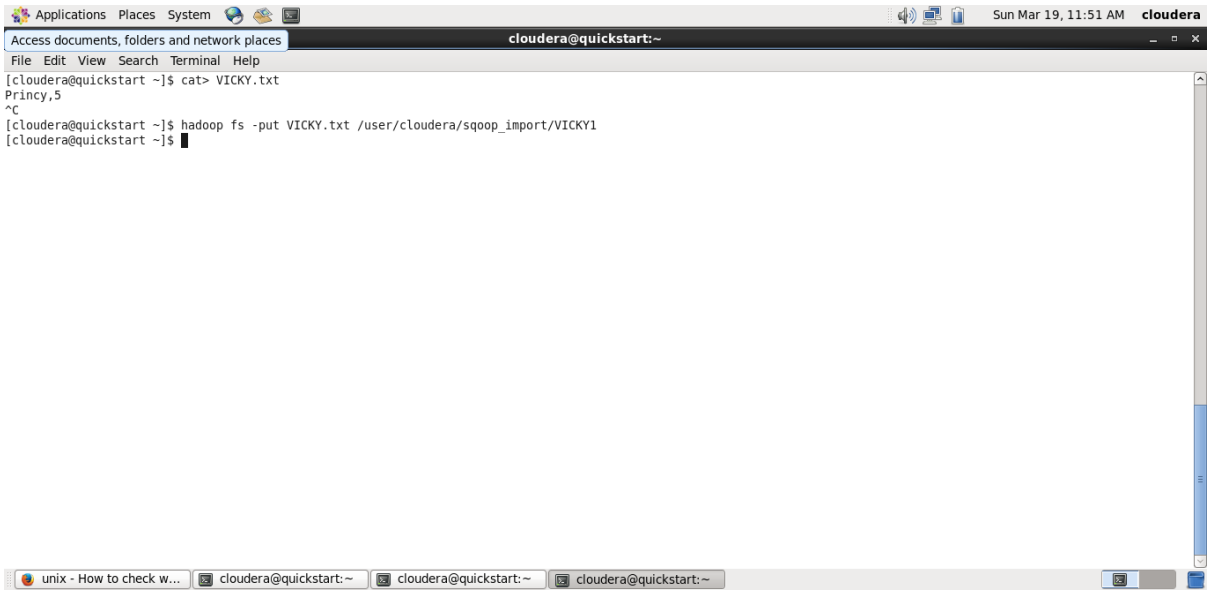
Exporting a table from Hadoop to Sqoop:-

Step 10: Create a normal text file with the same constraints that suits the database.

Command: cat> <filename>.txt

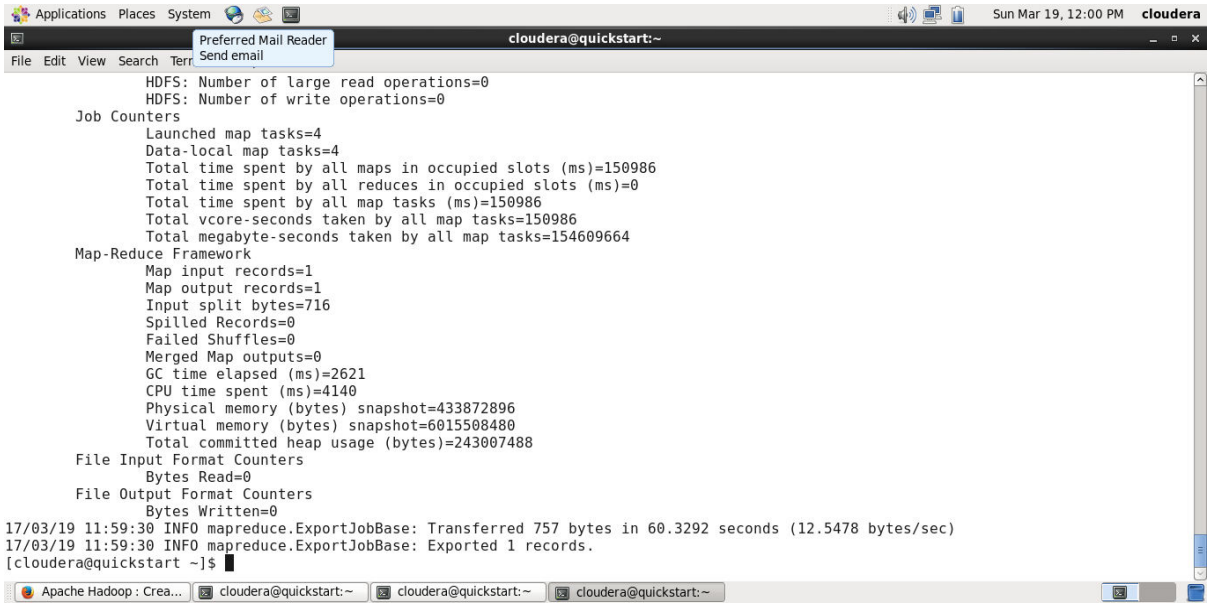
Step 11: Copy the created text file to the HDFS file system.

Command: hadoop fs -put <filename>.txt <target_directory>



Step 12: Now export the modified imported table to Database.

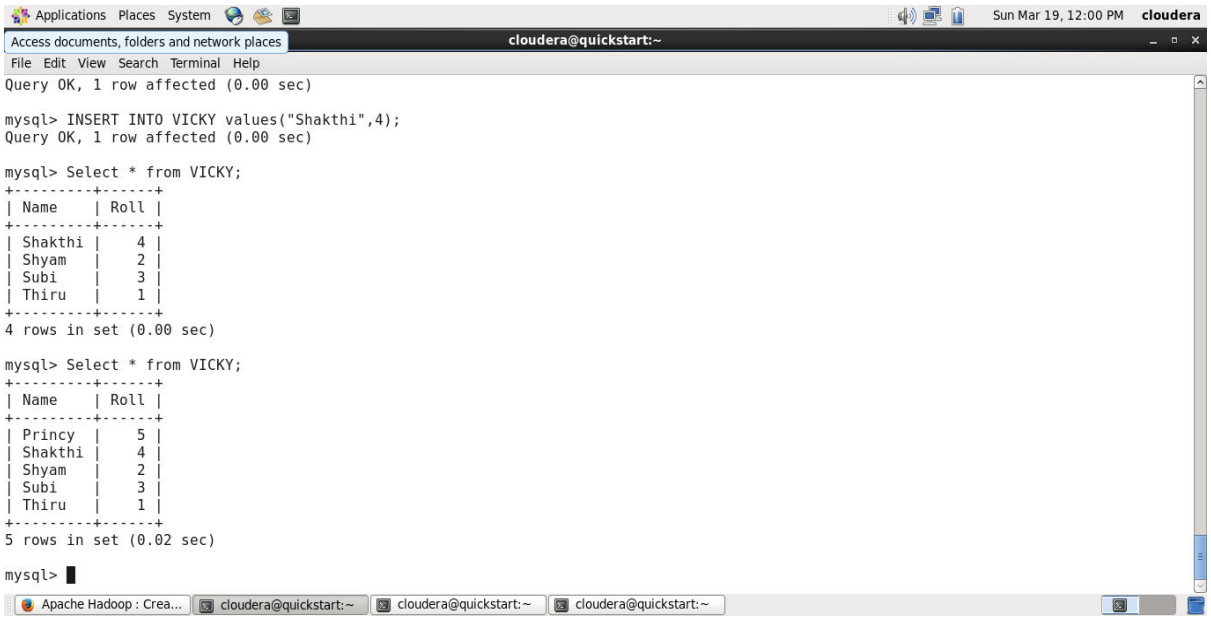
Command: sqoop export --connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" --username retail_dba --password cloudera --table <table_name> --export-dir <target_directory>/<filename>.txt



Thus the text file is successfully exported and appended to the existing table in the SQL Database.

Step 13: Switch to the SQL terminal to view the updated table.

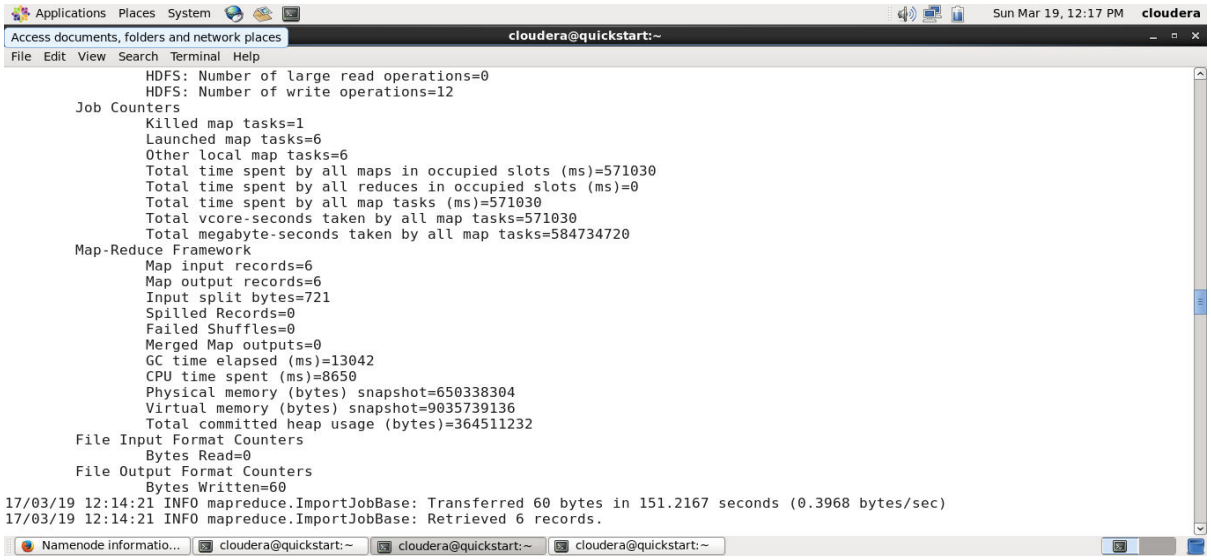
Command: Select * from <table_name>;



Importing all tables from SQL to Hadoop:-

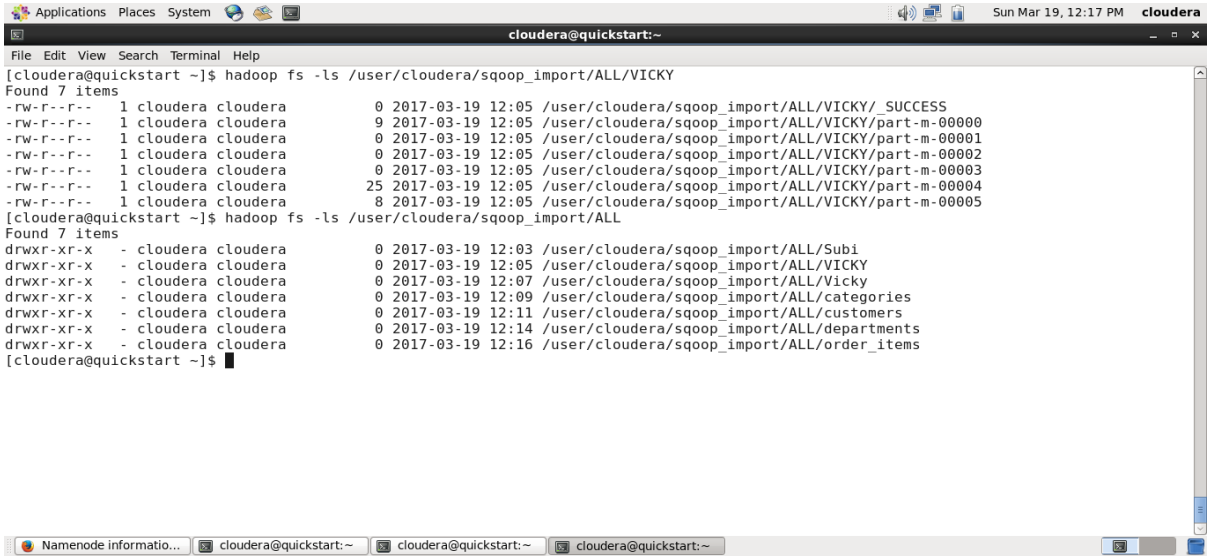
Step 14: To import all the tables from the SQL Database.

Command: sqoop import-all-tables -m 4 --connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" --username=retail_dba --password=cloudera --warehouse-dir=<target_directory>



Step 15: To display the list of the imported table.

Command: hadoop fs -ls <target_directory>/*



Some predefined table also gets imported to the HDFS system.

Conclusion:

Thus the tables were imported from SQL Database to HDFS and exported from HDFS to SQL Database using Sqoop.