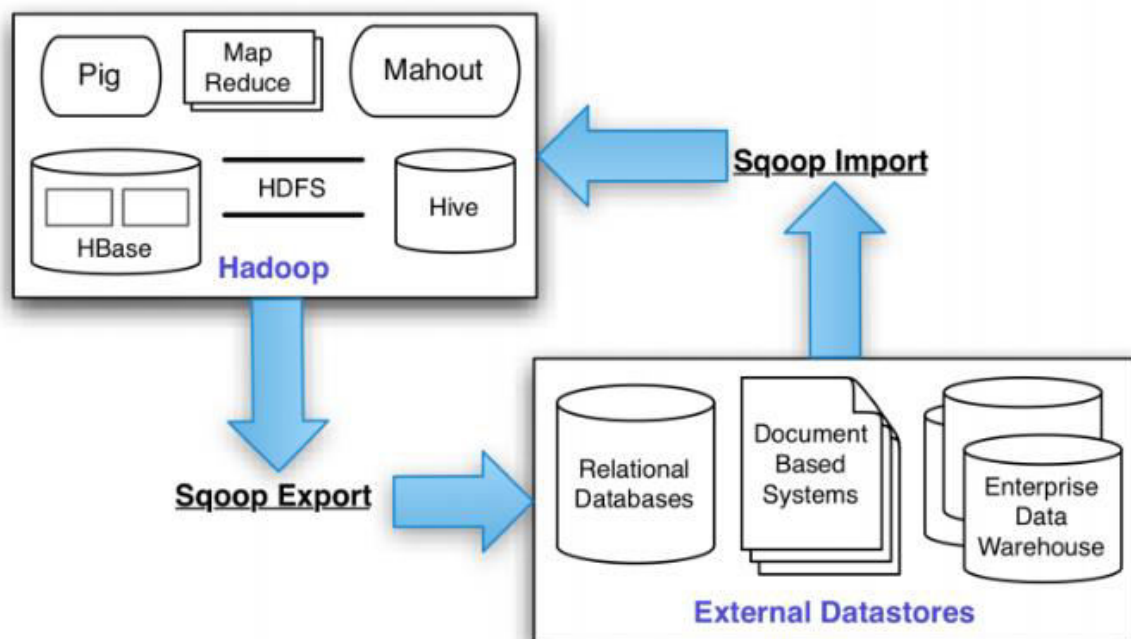


Documentation of Sqoop

-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.



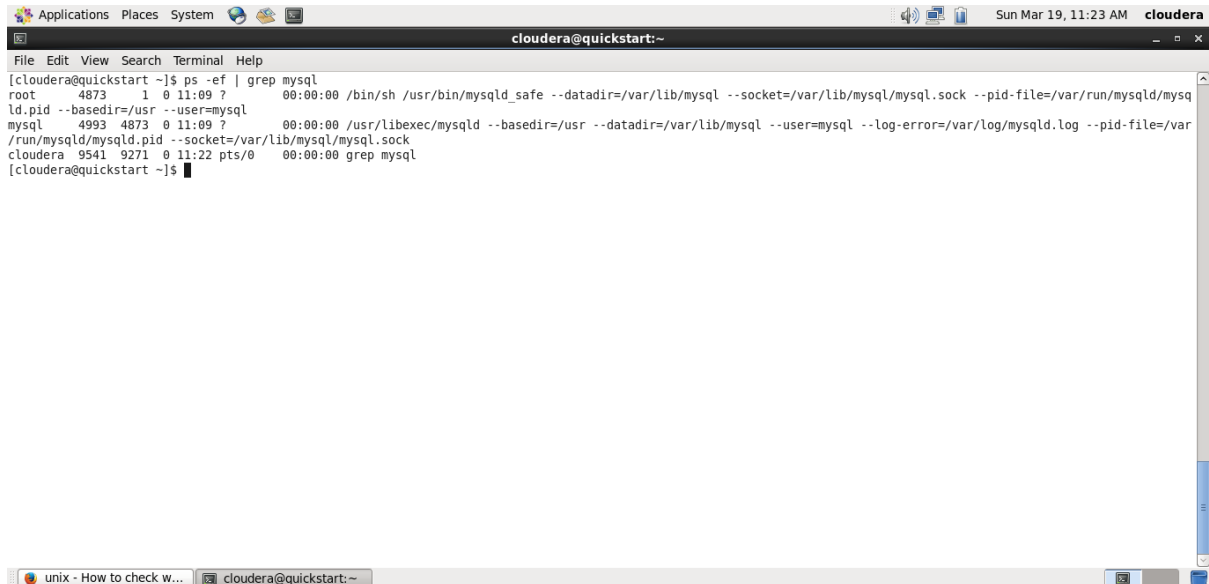
Sqoop is a tool designed to transfer data between Hadoop and relational databases. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the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

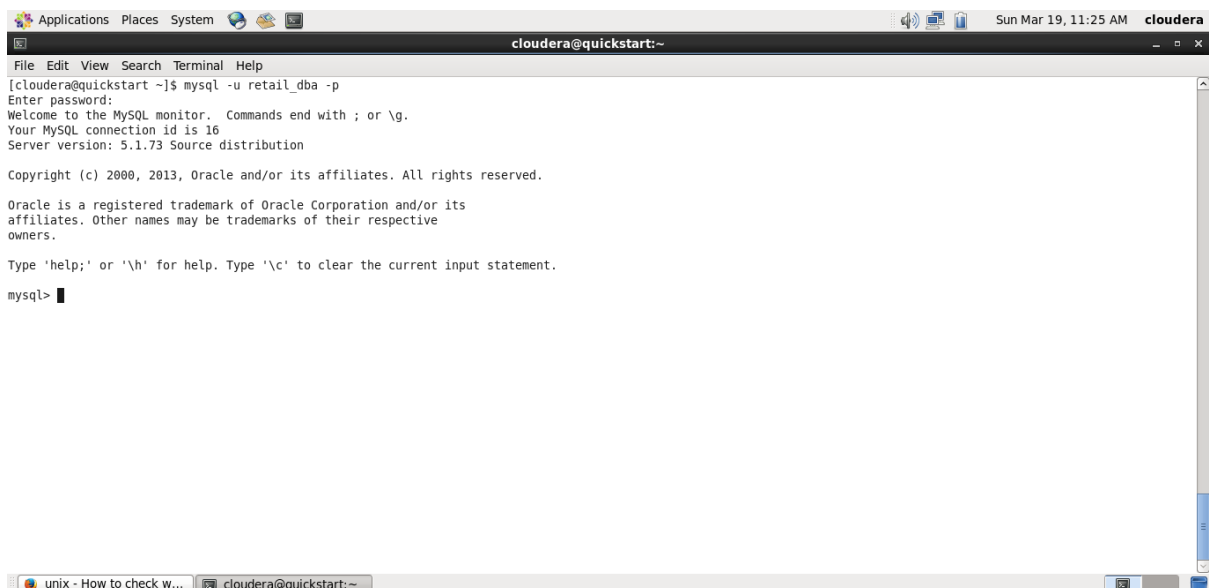


```
Applications Places System cloudera@quickstart:~
[cloudera@quickstart ~]$ ps -ef | grep mysql
root      4873      1  0 11:09 ?        00:00:00 /bin/sh /usr/bin/mysqld_safe --datadir=/var/lib/mysql --socket=/var/lib/mysql/mysql.sock --pid-file=/var/run/mysqld/mysq
ld.pid --basedir=/usr --user=mysql
mysql     4993  4873  0 11:09 ?        00:00:00 /usr/libexec/mysqld --basedir=/usr --datadir=/var/lib/mysql --user=mysql --log-error=/var/log/mysqld.log --pid-file=/var
/run/mysqld/mysqld.pid --socket=/var/lib/mysql/mysql.sock
cloudera  9541  9271  0 11:22 pts/0    00:00:00 grep mysql
[cloudera@quickstart ~]$
```

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`



```
Applications Places System cloudera@quickstart:~
[cloudera@quickstart ~]$ mysql -u retail_dba -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 16
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

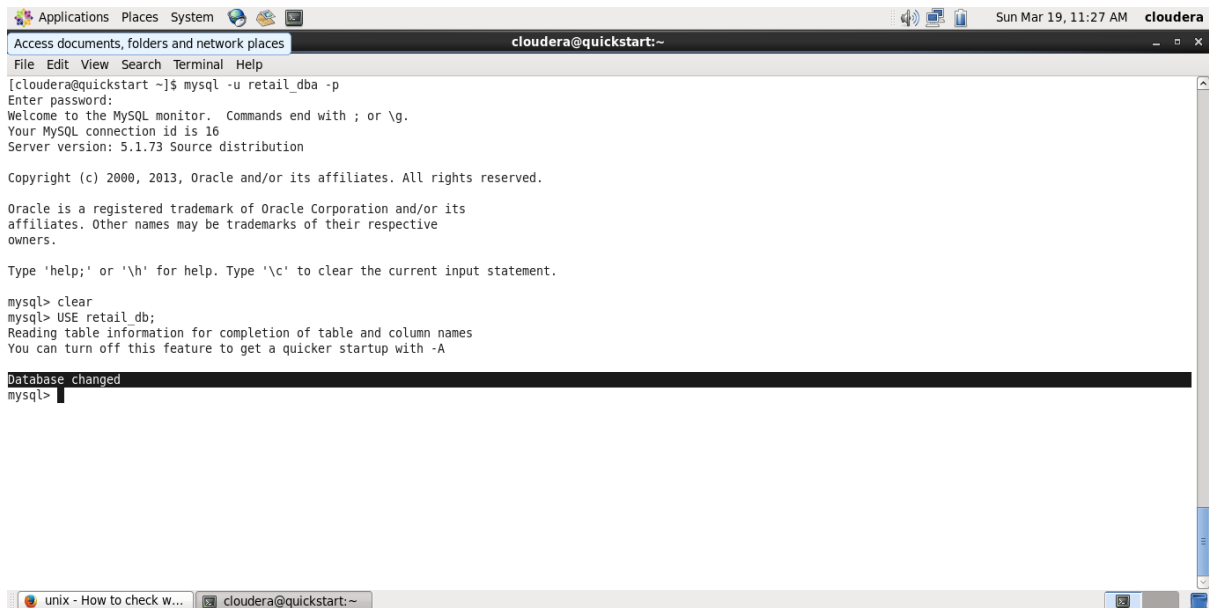
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql>
```

Step 3:

Then the database is changed using the following command.

Command: USE retail_db;



The screenshot shows a terminal window titled 'cloudera@quickstart:~'. The user has entered the command 'mysql -u retail_dba -p' and provided a password. The MySQL prompt 'mysql>' is shown. The user enters 'clear', then 'USE retail db;'. The output shows 'Database changed' and the prompt returns to 'mysql>'. The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The status bar at the bottom shows 'unix - How to check w...' and 'cloudera@quickstart:~'.

```
cloudera@quickstart:~$ mysql -u retail_dba -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 16
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

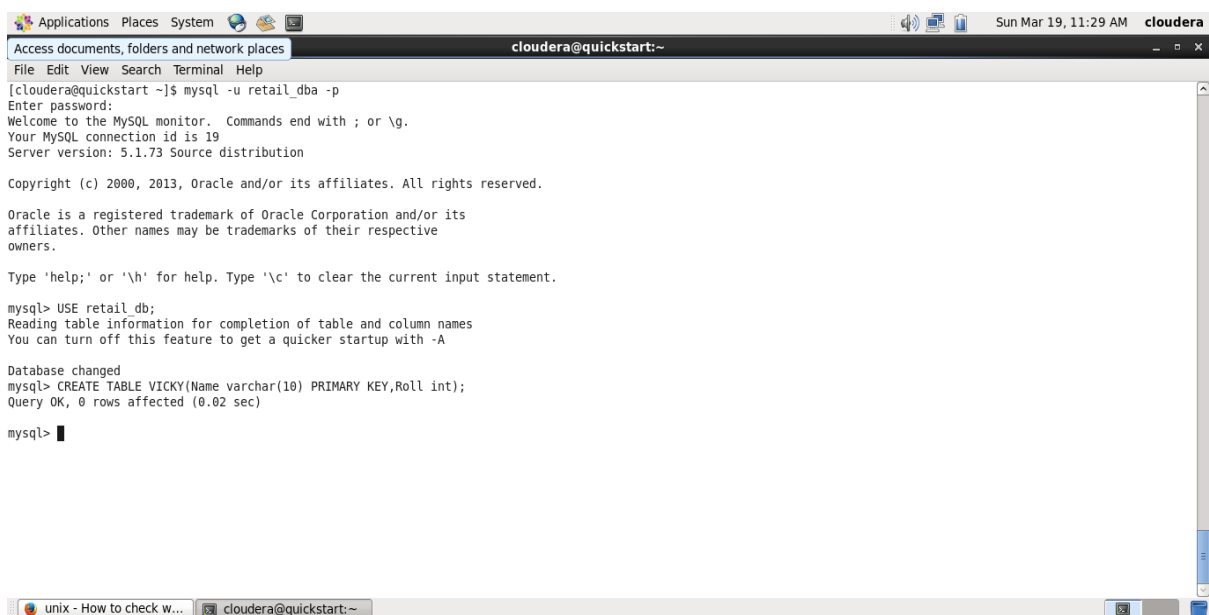
mysql> clear
mysql> USE retail db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql>
```

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,...);



The screenshot shows a terminal window titled 'cloudera@quickstart:~'. The user has entered the command 'mysql -u retail_dba -p' and provided a password. The MySQL prompt 'mysql>' is shown. The user enters 'USE retail db;'. The output shows 'Database changed' and the prompt returns to 'mysql>'. The user enters 'CREATE TABLE VICKY(Name varchar(10) PRIMARY KEY,Roll int);'. The output shows 'Query OK, 0 rows affected (0.02 sec)' and the prompt returns to 'mysql>'. The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The status bar at the bottom shows 'unix - How to check w...' and 'cloudera@quickstart:~'.

```
cloudera@quickstart:~$ mysql -u retail_dba -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 19
Server version: 5.1.73 Source distribution

Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> USE retail db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

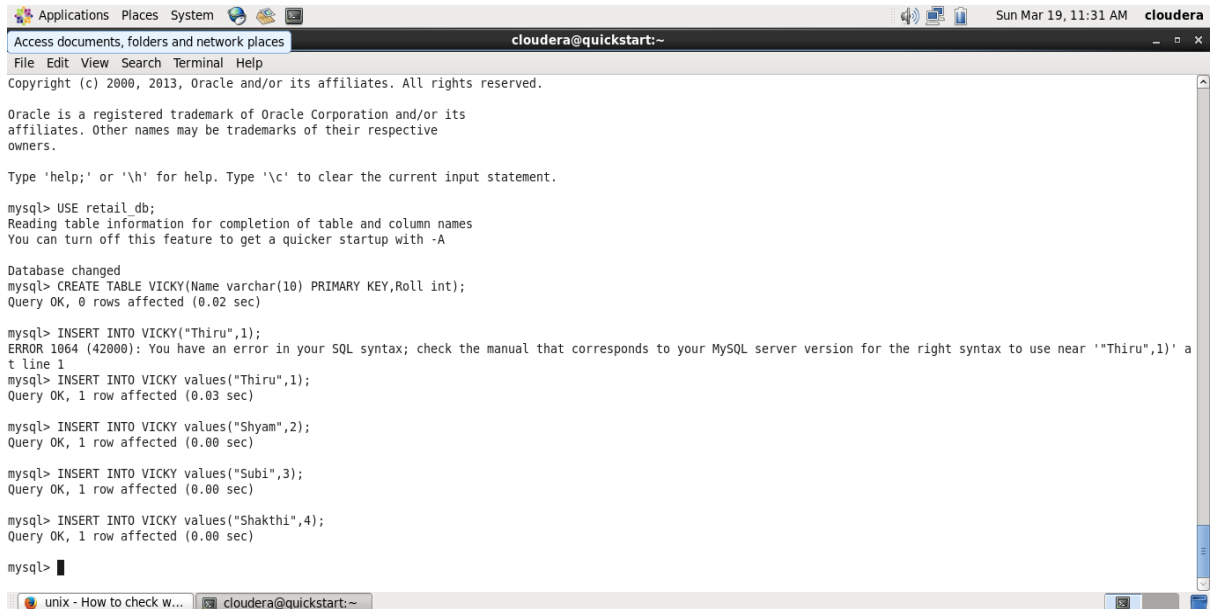
Database changed
mysql> CREATE TABLE VICKY(Name varchar(10) PRIMARY KEY,Roll int);
Query OK, 0 rows affected (0.02 sec)

mysql>
```

Step 5:

Insert some data in the created table.

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



```
Applications Places System cloudera@quickstart:~
Access documents, folders and network places
File Edit View Search Terminal Help
Copyright (c) 2000, 2013, Oracle and/or its affiliates. All rights reserved.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> USE retail db;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> CREATE TABLE VICKY(Name varchar(10) PRIMARY KEY,Roll int);
Query OK, 0 rows affected (0.02 sec)

mysql> INSERT INTO VICKY("Thiru",1);
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '"Thiru",1)' a
t line 1
mysql> INSERT INTO VICKY values("Thiru",1);
Query OK, 1 row affected (0.03 sec)

mysql> INSERT INTO VICKY values("Shyam",2);
Query OK, 1 row affected (0.00 sec)

mysql> INSERT INTO VICKY values("Subi",3);
Query OK, 1 row affected (0.00 sec)

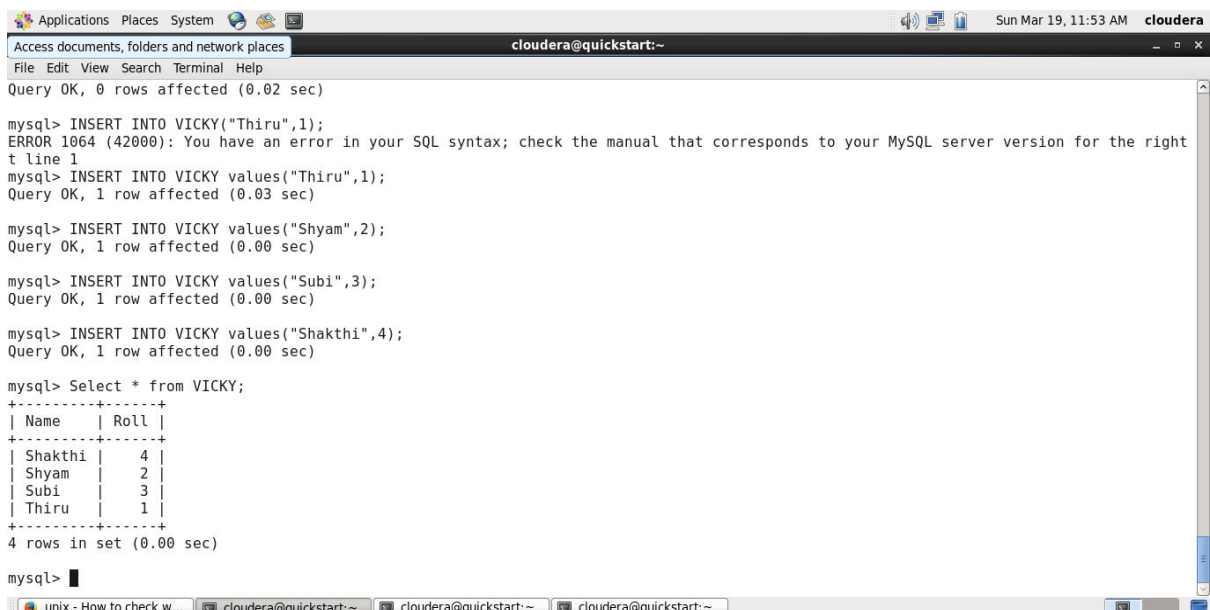
mysql> INSERT INTO VICKY values("Shakthi",4);
Query OK, 1 row affected (0.00 sec)

mysql>
```

Step 6:

To display the data in the table.

Command: Select * from <table_name>;



```
Applications Places System cloudera@quickstart:~
Access documents, folders and network places
File Edit View Search Terminal Help
Query OK, 0 rows affected (0.02 sec)

mysql> INSERT INTO VICKY("Thiru",1);
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right
t line 1
mysql> INSERT INTO VICKY values("Thiru",1);
Query OK, 1 row affected (0.03 sec)

mysql> INSERT INTO VICKY values("Shyam",2);
Query OK, 1 row affected (0.00 sec)

mysql> INSERT INTO VICKY values("Subi",3);
Query OK, 1 row affected (0.00 sec)

mysql> INSERT INTO VICKY values("Shakthi",4);
Query OK, 1 row affected (0.00 sec)

mysql> Select * from VICKY;
+-----+-----+
| Name | Roll |
+-----+-----+
| Shakthi | 4 |
| Shyam | 2 |
| Subi | 3 |
| Thiru | 1 |
+-----+-----+
4 rows in set (0.00 sec)

mysql>
```

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>`

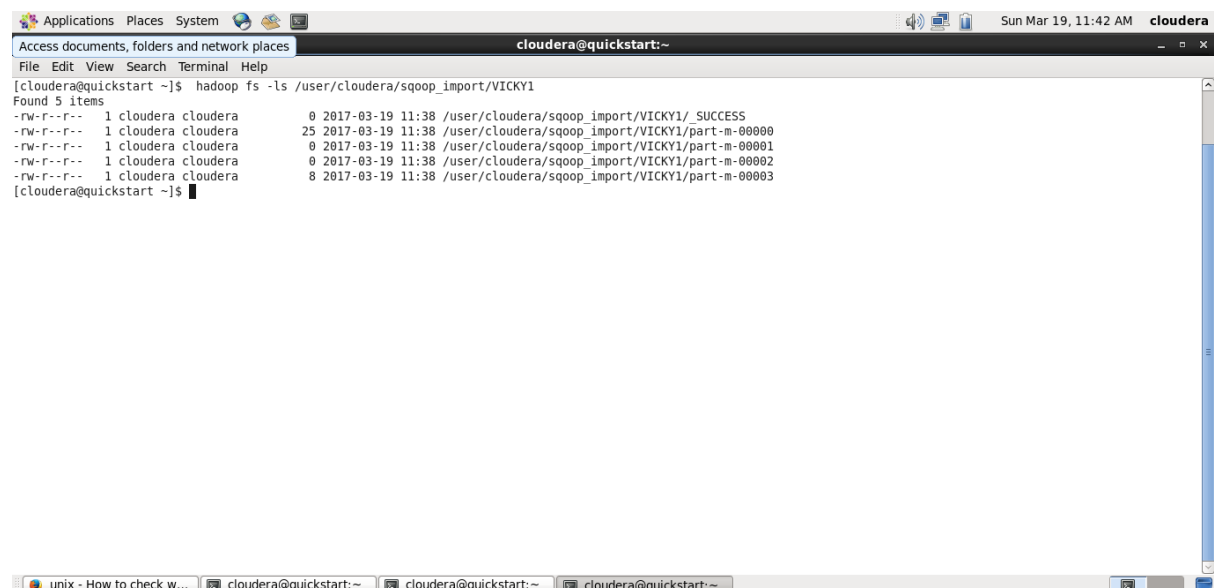


```
Applications Places System cloudera@quickstart:~
Access documents, folders and network places cloudera@quickstart:~
File Edit View Search Terminal Help
HDFS: Number of bytes read=525
HDFS: Number of bytes written=33
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
  Launched map tasks=4
  Other local map tasks=4
  Total time spent by all maps in occupied slots (ms)=185861
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=185861
  Total vcore-seconds taken by all map tasks=185861
  Total megabyte-seconds taken by all map tasks=190321664
Map-Reduce Framework
  Map input records=4
  Map output records=4
  Input split bytes=525
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=2591
  CPU time spent (ms)=5000
  Physical memory (bytes) snapshot=418635776
  Virtual memory (bytes) snapshot=6023929856
  Total committed heap usage (bytes)=243007488
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=33
17/03/19 11:38:53 INFO mapreduce.ImportJobBase: Transferred 33 bytes in 75.5491 seconds (0.4368 bytes/sec)
17/03/19 11:38:53 INFO mapreduce.ImportJobBase: Retrieved 4 records.
[cloudera@quickstart ~]$
```

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>`



```
Applications Places System cloudera@quickstart:~
Access documents, folders and network places cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/sqoop_import/VICKY1
Found 5 items
-rw-r--r-- 1 cloudera cloudera 0 2017-03-19 11:38 /user/cloudera/sqoop_import/VICKY1/_SUCCESS
-rw-r--r-- 1 cloudera cloudera 25 2017-03-19 11:38 /user/cloudera/sqoop_import/VICKY1/part-m-00000
-rw-r--r-- 1 cloudera cloudera 0 2017-03-19 11:38 /user/cloudera/sqoop_import/VICKY1/part-m-00001
-rw-r--r-- 1 cloudera cloudera 0 2017-03-19 11:38 /user/cloudera/sqoop_import/VICKY1/part-m-00002
-rw-r--r-- 1 cloudera cloudera 8 2017-03-19 11:38 /user/cloudera/sqoop_import/VICKY1/part-m-00003
[cloudera@quickstart ~]$
```

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

Command: `hadoop fs -cat <target_directory>/*`

A terminal window titled 'cloudera@quickstart:~' showing the command `hadoop fs -cat /user/cloudera/sqoop_import/VICKY1/*` being executed. The output lists five records: 'Princy,5', 'Shakthi,4', 'Shyam,2', 'Subi,3', and 'Thiru,1'. The terminal interface includes a menu bar (File, Edit, View, Search, Terminal, Help) and a taskbar at the bottom with several open windows.

```
cloudera@quickstart:~  
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/sqoop_import/VICKY1/*  
Princy,5  
Shakthi,4  
Shyam,2  
Subi,3  
Thiru,1  
[cloudera@quickstart ~]$
```

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

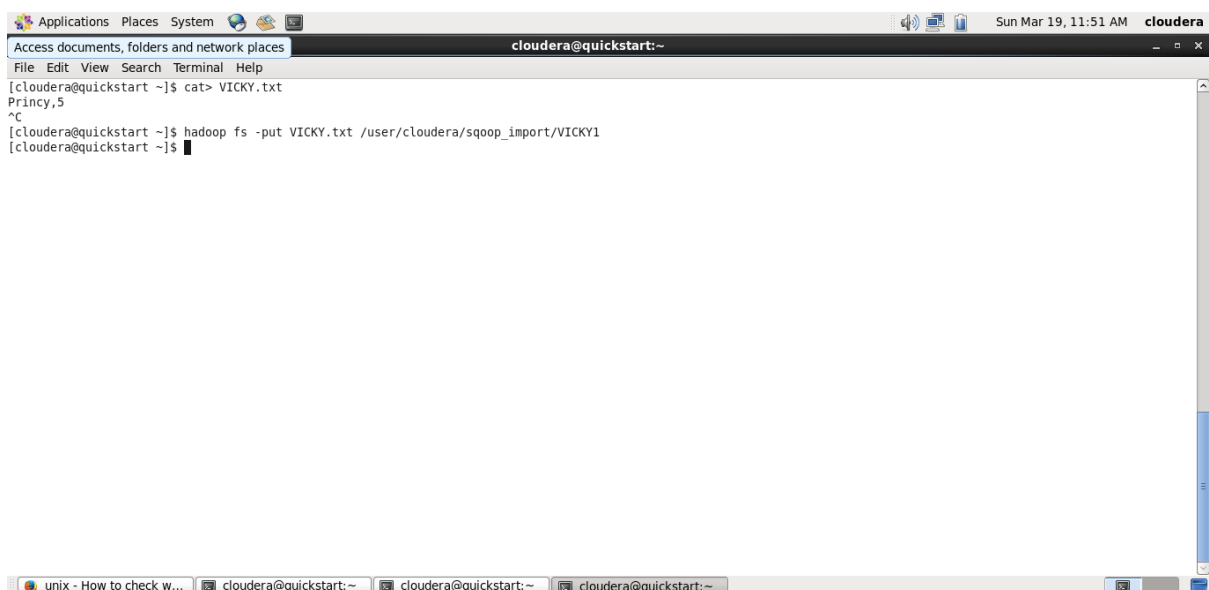
Exporting a table from Hadoop to Sqoop:-

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.

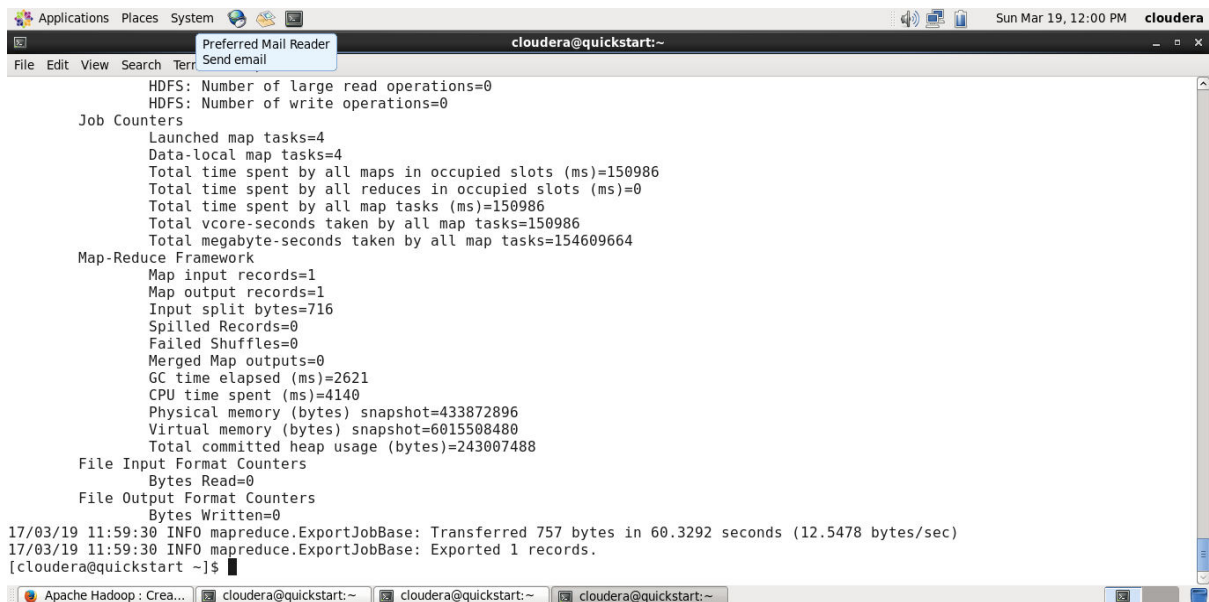
Command: `hadoop fs -put <filename>.txt <target_directory>`

A terminal window titled 'cloudera@quickstart:~' showing two commands being executed. First, `cat> VICKY.txt` is used to create a text file. Then, `hadoop fs -put VICKY.txt /user/cloudera/sqoop_import/VICKY1` is used to upload the file to HDFS. The terminal interface is consistent with the previous screenshot, showing a menu bar and a taskbar.

```
cloudera@quickstart:~  
[cloudera@quickstart ~]$ cat> VICKY.txt  
Princy,5  
^C  
[cloudera@quickstart ~]$ hadoop fs -put VICKY.txt /user/cloudera/sqoop_import/VICKY1  
[cloudera@quickstart ~]$
```

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`



```
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters
  Launched map tasks=4
  Data-local map tasks=4
  Total time spent by all maps in occupied slots (ms)=150986
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=150986
  Total vcore-seconds taken by all map tasks=150986
  Total megabyte-seconds taken by all map tasks=154609664
Map-Reduce Framework
  Map input records=1
  Map output records=1
  Input split bytes=716
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=2621
  CPU time spent (ms)=4140
  Physical memory (bytes) snapshot=433872896
  Virtual memory (bytes) snapshot=6015508480
  Total committed heap usage (bytes)=243007488
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=0
17/03/19 11:59:30 INFO mapreduce.ExportJobBase: Transferred 757 bytes in 60.3292 seconds (12.5478 bytes/sec)
17/03/19 11:59:30 INFO mapreduce.ExportJobBase: Exported 1 records.
[cloudera@quickstart ~]$
```

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>;`



```
Applications Places System cloudera@quickstart:~
Access documents, folders and network places
File Edit View Search Terminal Help
Query OK, 1 row affected (0.00 sec)

mysql> INSERT INTO VICKY values("Shakthi",4);
Query OK, 1 row affected (0.00 sec)

mysql> Select * from VICKY;
+-----+-----+
| Name  | Roll |
+-----+-----+
| Shakthi | 4    |
| Shyam  | 2    |
| Subi   | 3    |
| Thiru  | 1    |
+-----+-----+
4 rows in set (0.00 sec)

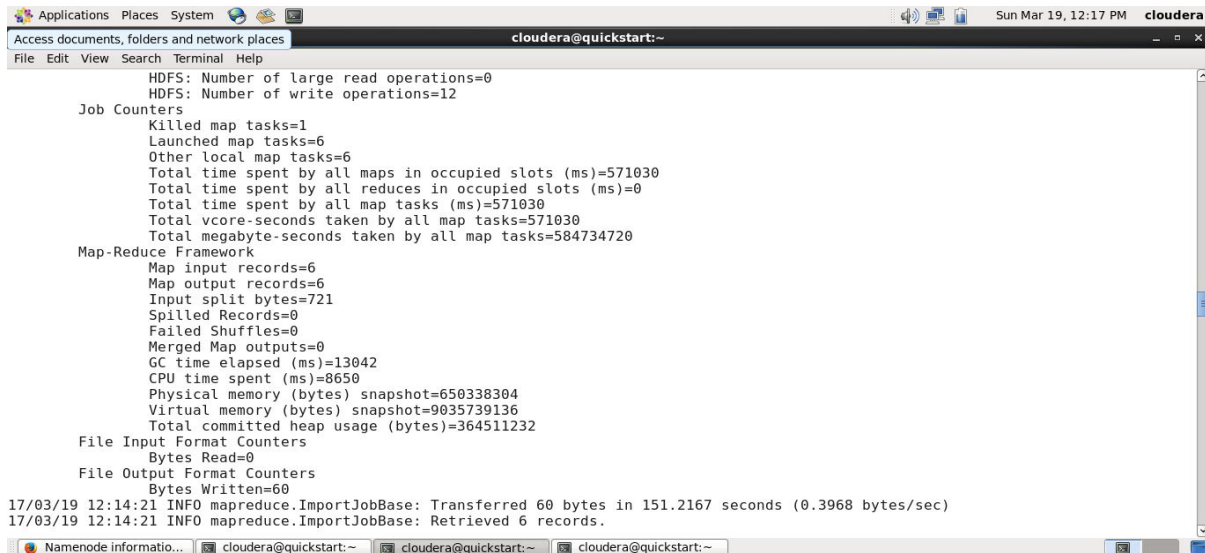
mysql> Select * from VICKY;
+-----+-----+
| Name  | Roll |
+-----+-----+
| Princy | 5    |
| Shakthi | 4    |
| Shyam  | 2    |
| Subi   | 3    |
| Thiru  | 1    |
+-----+-----+
5 rows in set (0.02 sec)

mysql>
```

Importing all tables from SQL to Hadoop:-

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

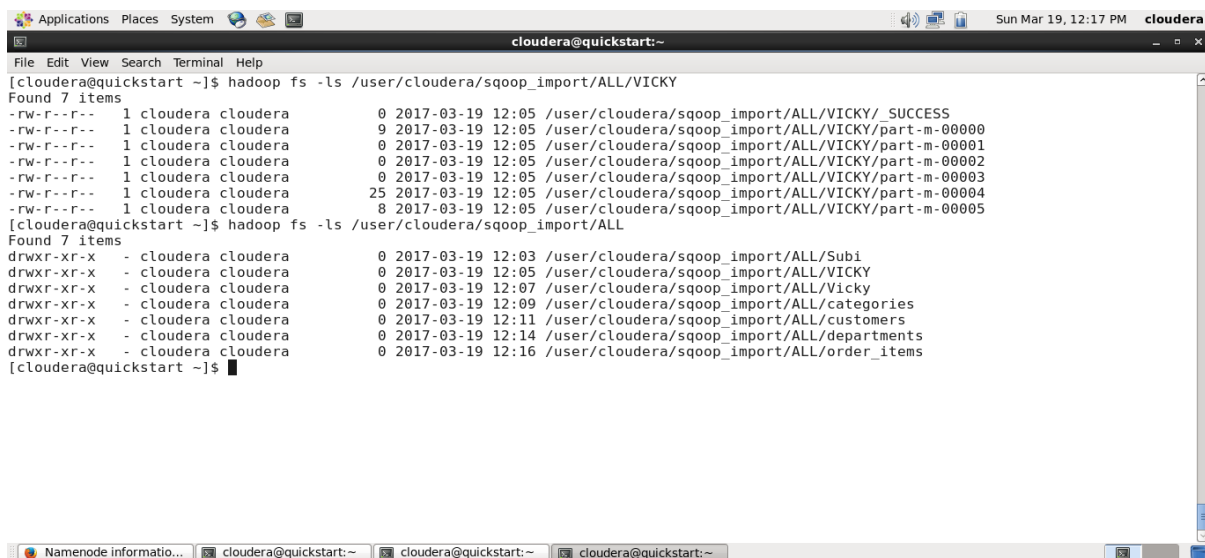
Command: `sqoop import-all-tables -m 4 --connect "jdbc:mysql://quickstart.cloudera:3306/retail_db" --username=retail_dba --password=cloudera --warehouse-dir=<target_directory>`



```
Applications Places System cloudera@quickstart:~
Access documents, folders and network places
File Edit View Search Terminal Help
HDFS: Number of large read operations=0
HDFS: Number of write operations=12
Job Counters
  Killed map tasks=1
  Launched map tasks=6
  Other local map tasks=6
  Total time spent by all maps in occupied slots (ms)=571030
  Total time spent by all reduces in occupied slots (ms)=0
  Total time spent by all map tasks (ms)=571030
  Total vcore-seconds taken by all map tasks=571030
  Total megabyte-seconds taken by all map tasks=584734720
Map-Reduce Framework
  Map input records=6
  Map output records=6
  Input split bytes=721
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=13042
  CPU time spent (ms)=8650
  Physical memory (bytes) snapshot=650338304
  Virtual memory (bytes) snapshot=9035739136
  Total committed heap usage (bytes)=364511232
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=60
17/03/19 12:14:21 INFO mapreduce.ImportJobBase: Transferred 60 bytes in 151.2167 seconds (0.3968 bytes/sec)
17/03/19 12:14:21 INFO mapreduce.ImportJobBase: Retrieved 6 records.
```

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

Command: `hadoop fs -ls <target_directory>/*`



```
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/sqoop_import/ALL/VICKY
Found 7 items
-rw-r--r-- 1 cloudera cloudera 0 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY/_SUCCESS
-rw-r--r-- 1 cloudera cloudera 9 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY/part-m-00000
-rw-r--r-- 1 cloudera cloudera 0 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY/part-m-00001
-rw-r--r-- 1 cloudera cloudera 0 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY/part-m-00002
-rw-r--r-- 1 cloudera cloudera 0 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY/part-m-00003
-rw-r--r-- 1 cloudera cloudera 25 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY/part-m-00004
-rw-r--r-- 1 cloudera cloudera 8 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY/part-m-00005
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/sqoop_import/ALL
Found 7 items
drwxr-xr-x - cloudera cloudera 0 2017-03-19 12:03 /user/cloudera/sqoop_import/ALL/Subi
drwxr-xr-x - cloudera cloudera 0 2017-03-19 12:05 /user/cloudera/sqoop_import/ALL/VICKY
drwxr-xr-x - cloudera cloudera 0 2017-03-19 12:07 /user/cloudera/sqoop_import/ALL/Vicky
drwxr-xr-x - cloudera cloudera 0 2017-03-19 12:09 /user/cloudera/sqoop_import/ALL/categories
drwxr-xr-x - cloudera cloudera 0 2017-03-19 12:11 /user/cloudera/sqoop_import/ALL/customers
drwxr-xr-x - cloudera cloudera 0 2017-03-19 12:14 /user/cloudera/sqoop_import/ALL/departments
drwxr-xr-x - cloudera cloudera 0 2017-03-19 12:16 /user/cloudera/sqoop_import/ALL/order_items
[cloudera@quickstart ~]$
```

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.