**Data Testing-Assignment-Part2 solution-Srimathi Rajendran**

**Create a directory in HDFS called input**

[cloudera@quickstart ~]$ ls

lcloudera-manager Downloads kerberos Pictures workspace

cm\_api.py eclipse lib Public

Desktop enterprise-deployment.json Music Templates

Documents express-deployment.json parcels Videos

[cloudera@quickstart ~]$ hadoop version

Hadoop 2.6.0-cdh5.13.0

Subversion http://github.com/cloudera/hadoop -r 42e8860b182e55321bd5f5605264da4adc8882be

Compiled by jenkins on 2017-10-04T18:08Z

Compiled with protoc 2.5.0

From source with checksum 5e84c185f8a22158e2b0e4b8f85311

This command was run using /usr/lib/hadoop/hadoop-common-2.6.0-cdh5.13.0.jar

[cloudera@quickstart ~]$ hadoop fs -ls /

Found 6 items

drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks

drwxr-xr-x - hbase supergroup 0 2019-01-03 20:08 /hbase

drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr

drwxrwxrwt - hdfs supergroup 0 2019-01-02 02:33 /tmp

drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /user

drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var

[cloudera@quickstart ~]$ hadoop fs -ls

[cloudera@quickstart ~]$ hadoop fs -mkdir SDET

mkdir: Cannot create directory /user/cloudera/SDET. Name node is in safe mode.

[cloudera@quickstart ~]$ hadoop dfsadmin -safemode leave

DEPRECATED: Use of this script to execute hdfs command is deprecated.

Instead use the hdfs command for it.

Safe mode is OFF

**cloudera@quickstart ~]$ hadoop fs -mkdir input**

[cloudera@quickstart ~]$

[cloudera@quickstart ~]$ ls

cloudera-manager Downloads parcels

cm\_api.py eclipse Pictures

data\_quality\_input enterprise-deployment.json Public

data\_quality\_input~ express-deployment.json Templates

data\_quality\_input - Copy.csv kerberos Videos

Desktop lib workspace

Documents Music

cloudera@quickstart ~]$ ls

cloudera-manager Downloads parcels

cm\_api.py eclipse Pictures

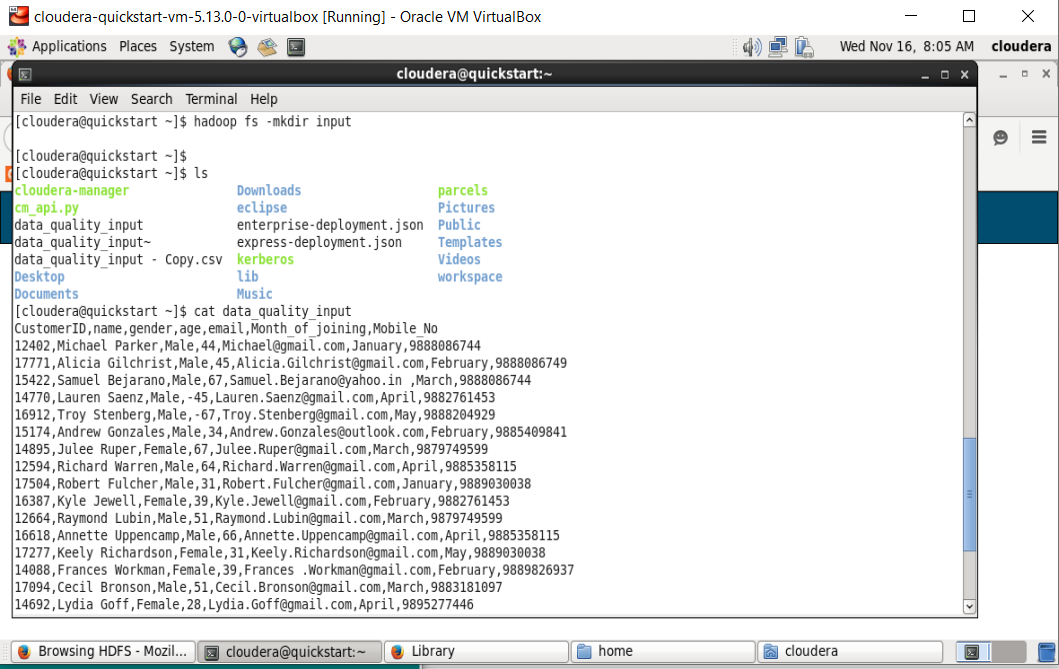
data\_quality\_input enterprise-deployment.json Public

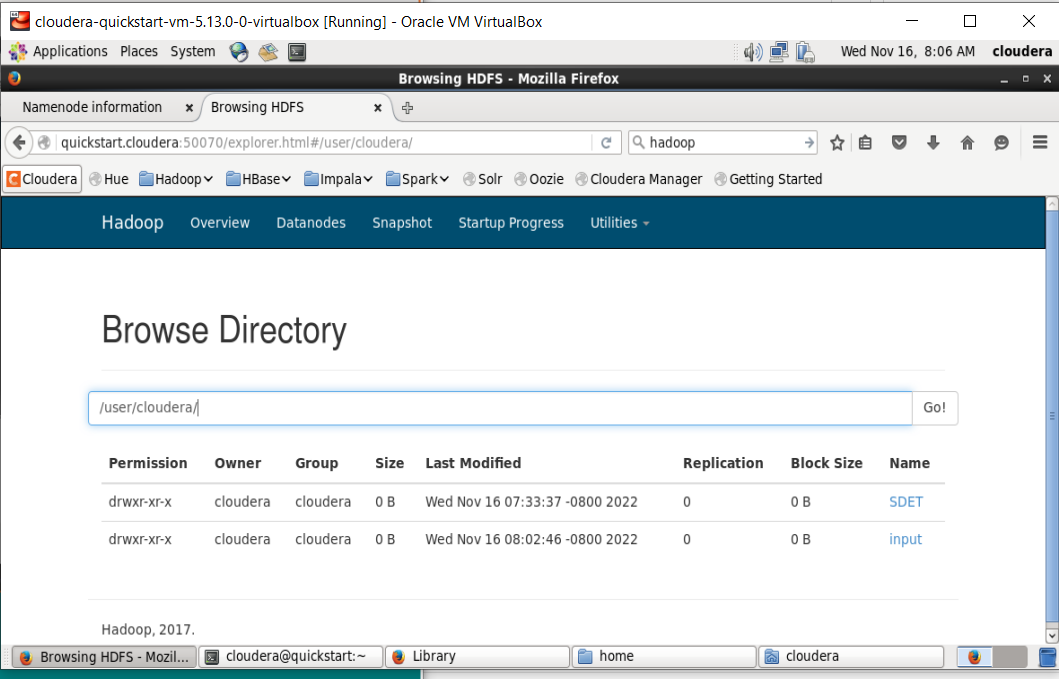
data\_quality\_input~ express-deployment.json Templates

data\_quality\_input - Copy.csv kerberos Videos

Desktop lib workspace

Documents Music





**Load the file data\_quality\_input.csv into this directory**

[cloudera@quickstart ~]$ cat data\_quality\_input

CustomerID,name,gender,age,email,Month\_of\_joining,Mobile\_No

12402,Michael Parker,Male,44,Michael@gmail.com,January,9888086744

17771,Alicia Gilchrist,Male,45,Alicia.Gilchrist@gmail.com,February,9888086749

15422,Samuel Bejarano,Male,67,Samuel.Bejarano@yahoo.in ,March,9888086744

14770,Lauren Saenz,Male,-45,Lauren.Saenz@gmail.com,April,9882761453

16912,Troy Stenberg,Male,-67,Troy.Stenberg@gmail.com,May,9888204929

15174,Andrew Gonzales,Male,34,Andrew.Gonzales@outlook.com,February,9885409841

14895,Julee Ruper,Female,67,Julee.Ruper@gmail.com,March,9879749599

12594,Richard Warren,Male,64,Richard.Warren@gmail.com,April,9885358115

17504,Robert Fulcher,Male,31,Robert.Fulcher@gmail.com,January,9889030038

16387,Kyle Jewell,Female,39,Kyle.Jewell@gmail.com,February,9882761453

12664,Raymond Lubin,Male,51,Raymond.Lubin@gmail.com,March,9879749599

16618,Annette Uppencamp,Male,66,Annette.Uppencamp@gmail.com,April,9885358115

17277,Keely Richardson,Female,31,Keely.Richardson@gmail.com,May,9889030038

14088,Frances Workman,Female,39,Frances .Workman@gmail.com,February,9889826937

17094,Cecil Bronson,Male,51,Cecil.Bronson@gmail.com,March,9883181097

14692,Lydia Goff,Female,28,Lydia.Goff@gmail.com,April,9895277446

12970,Zoe Ruiz,Female,24,Zoe.Ruiz@gmail.com,January,9880266361

17889,Lois Robins,Male,42,Lois.Robins@gmail.com,February,9890913257

17644,Paul Canter,Male,24,Paul.Canter@gmail.com,March,9887737046

17714,Casey Bring,Male,70,Casey.Bring@gmail.com,April,9896115615

[cloudera@quickstart ~]$ hadoop fs -copyFromLocal data\_quality\_input input

^C[cloudera@quickstart ~]$ hadoop fs -ls input

^C[cloudera@quickstart ~]$

[cloudera@quickstart ~]$ **hadoop fs -copyFromLocal data\_quality\_input input**

[cloudera@quickstart ~]$ **hadoop fs -ls input**

**Found 1 items**

**-rw-r--r-- 1 cloudera cloudera 1488 2022-11-16 08:02 input/data\_quality\_input**

[cloudera@quickstart ~]$ **hadoop dfs -cat input/data\_quality\_input**

**DEPRECATED: Use of this script to execute hdfs command is deprecated.**

**Instead use the hdfs command for it.**

**CustomerID,name,gender,age,email,Month\_of\_joining,Mobile\_No**

12402,Michael Parker,Male,44,Michael@gmail.com,January,9888086744

17771,Alicia Gilchrist,Male,45,Alicia.Gilchrist@gmail.com,February,9888086749

15422,Samuel Bejarano,Male,67,Samuel.Bejarano@yahoo.in ,March,9888086744

14770,Lauren Saenz,Male,-45,Lauren.Saenz@gmail.com,April,9882761453

16912,Troy Stenberg,Male,-67,Troy.Stenberg@gmail.com,May,9888204929

15174,Andrew Gonzales,Male,34,Andrew.Gonzales@outlook.com,February,9885409841

14895,Julee Ruper,Female,67,Julee.Ruper@gmail.com,March,9879749599

12594,Richard Warren,Male,64,Richard.Warren@gmail.com,April,9885358115

17504,Robert Fulcher,Male,31,Robert.Fulcher@gmail.com,January,9889030038

16387,Kyle Jewell,Female,39,Kyle.Jewell@gmail.com,February,9882761453

12664,Raymond Lubin,Male,51,Raymond.Lubin@gmail.com,March,9879749599

16618,Annette Uppencamp,Male,66,Annette.Uppencamp@gmail.com,April,9885358115

17277,Keely Richardson,Female,31,Keely.Richardson@gmail.com,May,9889030038

14088,Frances Workman,Female,39,Frances .Workman@gmail.com,February,9889826937

17094,Cecil Bronson,Male,51,Cecil.Bronson@gmail.com,March,9883181097

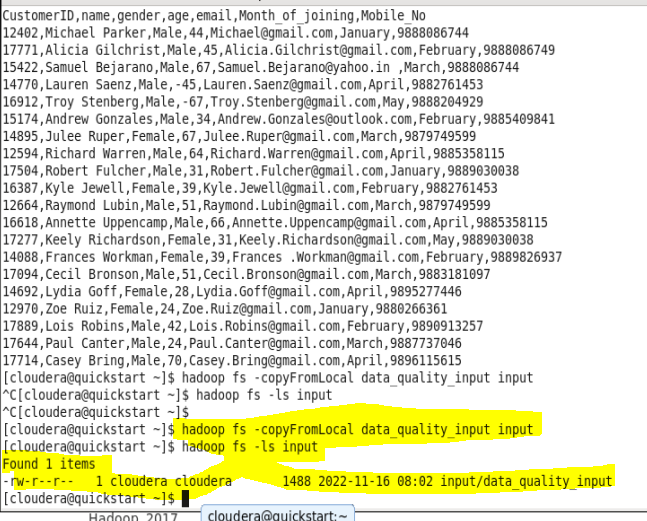
14692,Lydia Goff,Female,28,Lydia.Goff@gmail.com,April,9895277446

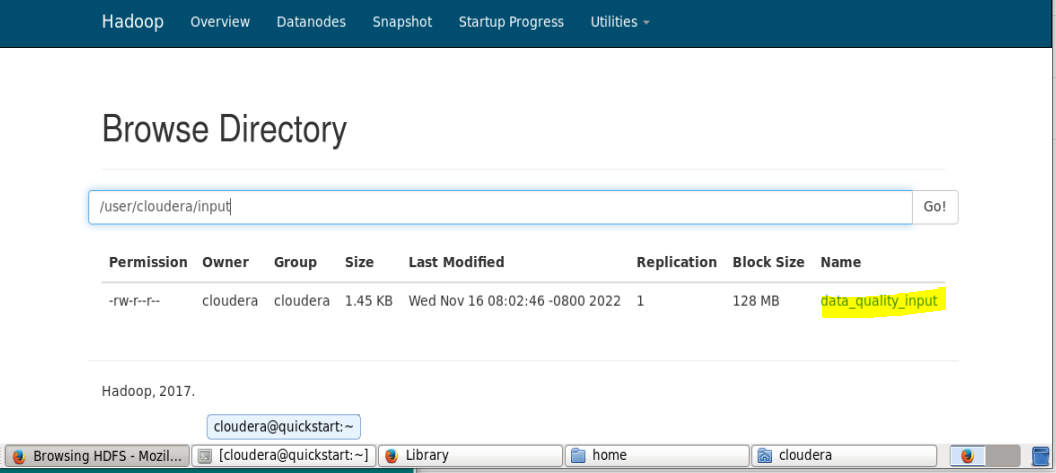
12970,Zoe Ruiz,Female,24,Zoe.Ruiz@gmail.com,January,9880266361

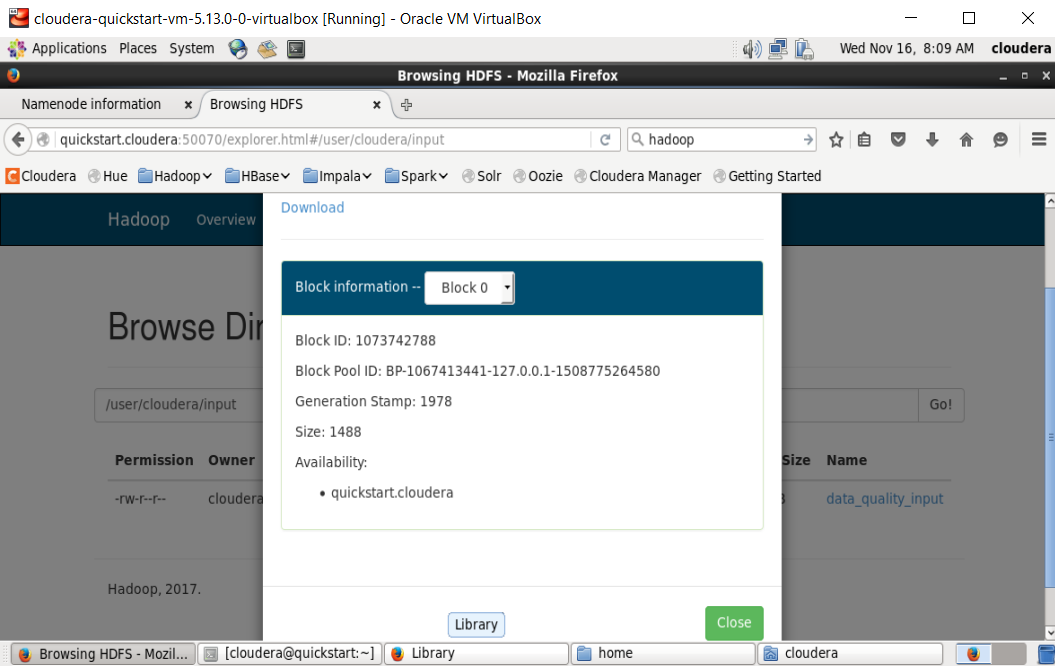
17889,Lois Robins,Male,42,Lois.Robins@gmail.com,February,9890913257

17644,Paul Canter,Male,24,Paul.Canter@gmail.com,March,9887737046

17714,Casey Bring,Male,70,Casey.Bring@gmail.com,April,9896115615

****





**Understand the given input data and create hive external table for** **this**

[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.properties

WARNING: Hive CLI is deprecated and migration to Beeline is recommended.

hive>

OK

Time taken: 19.059 seconds

hive> use srimathi;

OK

Time taken: 0.796 seconds

hive> **create table customer(CustomerID int,name string,gender string,age int,email string,Month\_of\_joining date,Mobile\_No int)**

**> row format delimited**

**> fields terminated by ','**

**> location '/user/cloudera/input'**

**> TBLPROPERTIES("skip.header.line.count"="1");**

OK

Time taken: 11.266 seconds

hive> select \* from customer

> select \* from customer;

FAILED: ParseException line 2:0 missing EOF at 'select' near 'customer'

hive> **select \* from customer;**

**OK**

12402 Michael Parker Male 44 Michael@gmail.com NULL NULL

17771 Alicia Gilchrist Male 45 Alicia.Gilchrist@gmail.com NULL NULL

15422 Samuel Bejarano Male 67 Samuel.Bejarano@yahoo.in NULL NULL

14770 Lauren Saenz Male -45 Lauren.Saenz@gmail.com NULL NULL

16912 Troy Stenberg Male -67 Troy.Stenberg@gmail.com NULL NULL

15174 Andrew Gonzales Male 34 Andrew.Gonzales@outlook.com NULL NULL

14895 Julee Ruper Female 67 Julee.Ruper@gmail.com NULL NULL

12594 Richard Warren Male 64 Richard.Warren@gmail.com NULL NULL

17504 Robert Fulcher Male 31 Robert.Fulcher@gmail.com NULL NULL

16387 Kyle Jewell Female 39 Kyle.Jewell@gmail.com NULL NULL

12664 Raymond Lubin Male 51 Raymond.Lubin@gmail.com NULL NULL

16618 Annette Uppencamp Male 66 Annette.Uppencamp@gmail.com NULL NULL

17277 Keely Richardson Female 31 Keely.Richardson@gmail.com NULL NULL

14088 Frances Workman Female 39 Frances .Workman@gmail.com NULL NULL

17094 Cecil Bronson Male 51 Cecil.Bronson@gmail.com NULL NULL

14692 Lydia Goff Female 28 Lydia.Goff@gmail.com NULL NULL

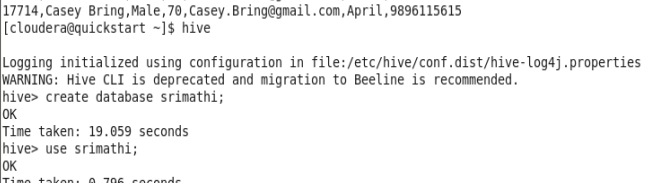
12970 Zoe Ruiz Female 24 Zoe.Ruiz@gmail.com NULL NULL

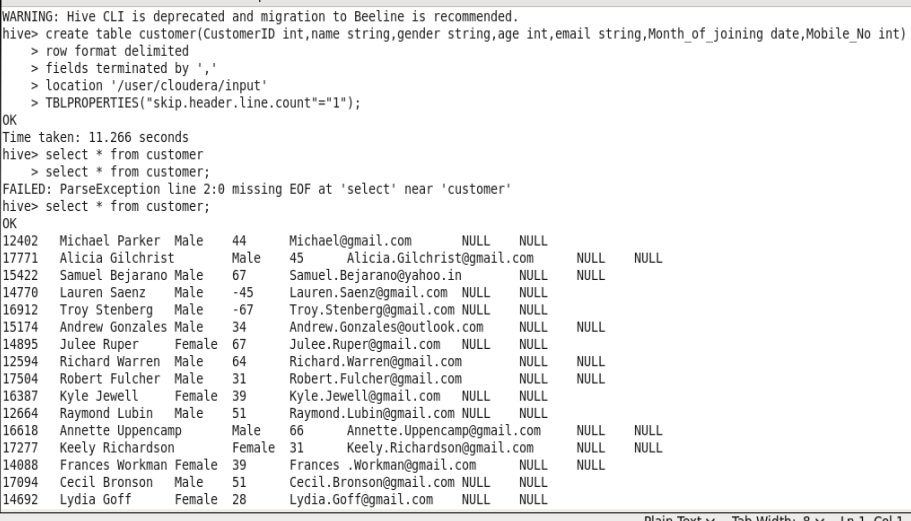
17889 Lois Robins Male 42 Lois.Robins@gmail.com NULL NULL

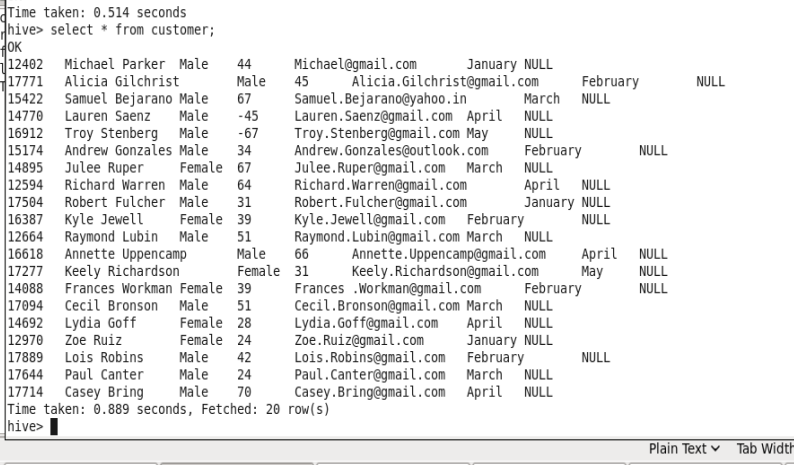
17644 Paul Canter Male 24 Paul.Canter@gmail.com NULL NULL

17714 Casey Bring Male 70 Casey.Bring@gmail.com NULL NULL

Time taken: 4.565 seconds, Fetched: 20 row(s)





****

**List down the distinct month names in the given input**

hive> **select DISTINCT Month\_of\_joining from customer;**

Query ID = cloudera\_20221116095454\_90eed8ab-6681-4116-927b-ac3ad5a3bc9b

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1668610156287\_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application\_1668610156287\_0004/

Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job\_1668610156287\_0004

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-11-16 09:55:05,453 Stage-1 map = 0%, reduce = 0%

2022-11-16 09:55:23,698 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.03 sec

2022-11-16 09:56:03,267 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.56 sec

MapReduce Total cumulative CPU time: 4 seconds 560 msec

Ended Job = job\_1668610156287\_0004

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.56 sec HDFS Read: 9275 HDFS Write: 33 SUCCESS

**Total MapReduce CPU Time Spent: 4 seconds 560 msec**

**OK**

**April**

**February**

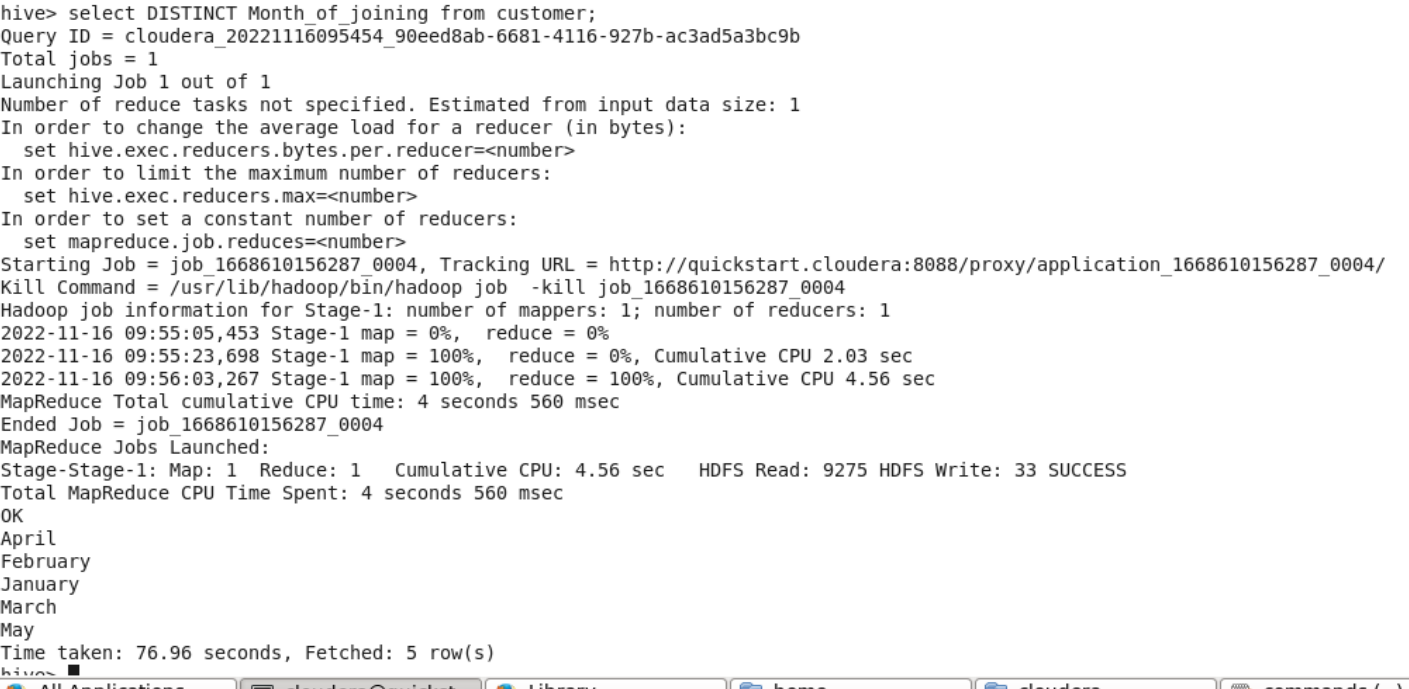
**January**

**March**

**May**

Time taken: 76.96 seconds, Fetched: 5 row(s)

hive>



**Calculate the total number of Male and Female in this data**

hive> **select count(\*), gender from customer group by gender;**

Query ID = cloudera\_20221116095050\_7823996f-2b88-441d-ab82-28b45b3c26c8

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1668610156287\_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application\_1668610156287\_0003/

Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job\_1668610156287\_0003

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2022-11-16 09:50:36,306 Stage-1 map = 0%, reduce = 0%

2022-11-16 09:51:22,697 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.02 sec

2022-11-16 09:52:23,468 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.02 sec

2022-11-16 09:53:24,388 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.02 sec

2022-11-16 09:54:08,171 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.95 sec

MapReduce Total cumulative CPU time: 4 seconds 950 msec

Ended Job = job\_1668610156287\_0003

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.95 sec HDFS Read: 10000 HDFS Write: 17 SUCCESS

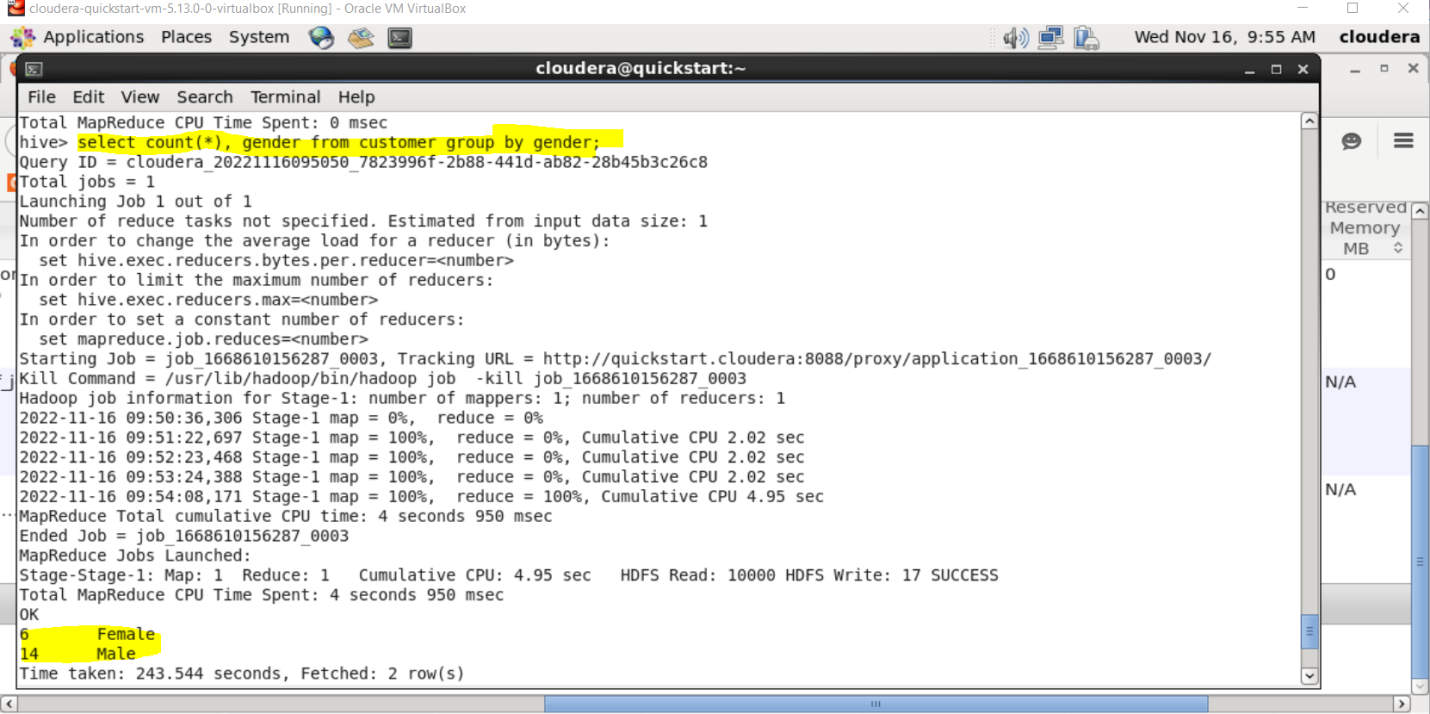
**Total MapReduce CPU Time Spent: 4 seconds 950 msec**

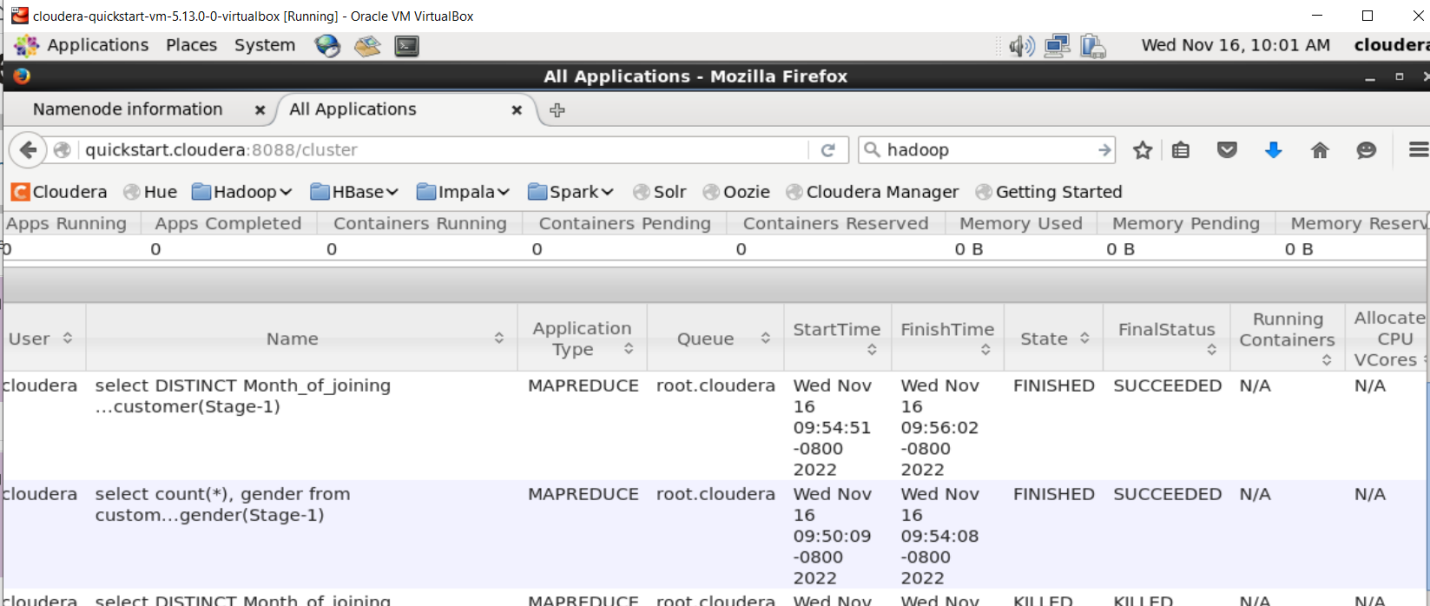
**OK**

**6 Female**

**14 Male**

Time taken: 243.544 seconds, Fetched: 2 row(s)

****

****

Storing- Name node, secondary name node, data node

Processing- Resource manager, Node manager

Name node is used to store the files and directories. It helps in mapping of the files as blocks to data nodes. It perform operation like opening, closing and renaming the files

Data nodes- Data node has the actual test date and it performs operations like create, delete, replicate based on the instruction from name node

Secondary name node- It is kind of back up of name node. It keeps the copy of merged namespace image which will help in the event of name node failing

Resource manager- It allocates resources among all applications and manage all the nodes

Node manager- Node manager perform health check of all the nodes and report it to resource manager. It monitors the containers and resource usage and reports it to resource manager

**Take the records only the age between 30 and 50 and store it in HDFS**

hive> **select \* from customer where age BETWEEN 30 AND 50;**

OK

12402 Michael Parker Male 44 Michael@gmail.com January NULL

17771 Alicia Gilchrist Male 45 Alicia.Gilchrist@gmail.com February NULL

15174 Andrew Gonzales Male 34 Andrew.Gonzales@outlook.com February NULL

17504 Robert Fulcher Male 31 Robert.Fulcher@gmail.com January NULL

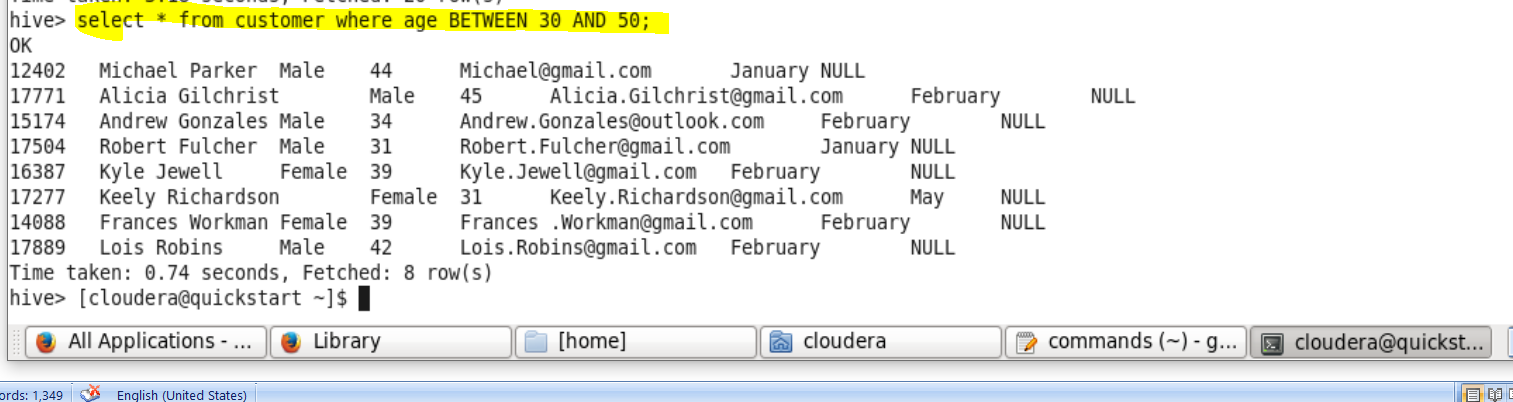
16387 Kyle Jewell Female 39 Kyle.Jewell@gmail.com February NULL

17277 Keely Richardson Female 31 Keely.Richardson@gmail.com May NULL

14088 Frances Workman Female 39 Frances .Workman@gmail.com February NULL

17889 Lois Robins Male 42 Lois.Robins@gmail.com February NULL

Time taken: 0.74 seconds, Fetched: 8 row(s)



**Create the Database and table for the above results**

mysql> use result;

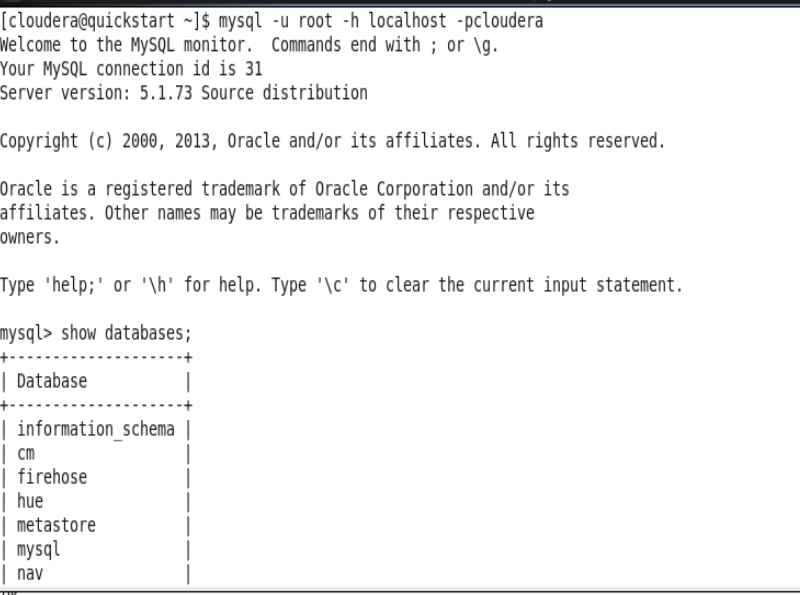
Database changed

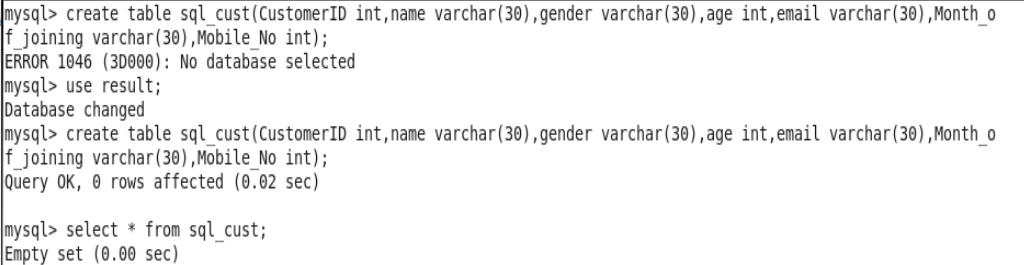
mysql> create table sql\_cust(CustomerID int,name varchar(30),gender varchar(30),age int,email varchar(30),Month\_of\_joining varchar(30),Mobile\_No int);

Query OK, 0 rows affected (0.02 sec)

mysql> select \* from sql\_cust;

Empty set (0.00 sec)

****

****

**Do the data export to transfer the data from HDFS to MySQL table**

[cloudera@quickstart ~]$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/result --table sql\_cust --username root --password cloudera --export-dir /user/hive/warehouse/srimathi.db/customer --m 1 --driver com.mysql.jdbc.driver --inpu-field-terminated-by ','

sqoop import --connect jdbc:mysql://127.0.0.1:3306/result --username root --password cloudera --table sql\_cust --hive-import --hive-table srimathi.customer --m 1

Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.

Please set $ACCUMULO\_HOME to the root of your Accumulo installation.

22/11/17 04:30:15 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0

22/11/17 04:30:15 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.

22/11/17 04:30:15 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override

22/11/17 04:30:15 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.

22/11/17 04:30:15 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.

22/11/17 04:30:15 INFO tool.CodeGenTool: Beginning code generation

22/11/17 04:30:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM `sql\_cust` AS t LIMIT 1

22/11/17 04:30:16 INFO manager.SqlManager: Executing SQL statement: SELECT t.\* FROM `sql\_cust` AS t LIMIT 1

22/11/17 04:30:16 INFO orm.CompilationManager: HADOOP\_MAPRED\_HOME is /usr/lib/hadoop-mapreduce

Note: /tmp/sqoop-cloudera/compile/041ba0896703f0c33cc09b3a329afef5/sql\_cust.java uses or overrides a deprecated API.

Note: Recompile with -Xlint:deprecation for details.

22/11/17 04:30:21 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-cloudera/compile/041ba0896703f0c33cc09b3a329afef5/sql\_cust.jar

22/11/17 04:30:21 WARN manager.MySQLManager: It looks like you are importing from mysql.

22/11/17 04:30:21 WARN manager.MySQLManager: This transfer can be faster! Use the --direct

22/11/17 04:30:21 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.

22/11/17 04:30:21 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)

22/11/17 04:30:21 INFO mapreduce.ImportJobBase: Beginning import of sql\_cust

22/11/17 04:30:21 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address

22/11/17 04:30:21 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar

22/11/17 04:30:23 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps

22/11/17 04:30:23 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032

22/11/17 04:30:26 WARN hdfs.DFSClient: Caught exception

java.lang.InterruptedException

at java.lang.Object.wait(Native Method)

at java.lang.Thread.join(Thread.java:1281)

at java.lang.Thread.join(Thread.java:1355)

at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)

at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)

at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)

22/11/17 04:30:28 INFO db.DBInputFormat: Using read commited transaction isolation

22/11/17 04:30:28 INFO mapreduce.JobSubmitter: number of splits:1

22/11/17 04:30:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1668610156287\_0006

22/11/17 04:30:30 INFO impl.YarnClientImpl: Submitted application application\_1668610156287\_0006

22/11/17 04:30:30 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application\_1668610156287\_0006/

22/11/17 04:30:30 INFO mapreduce.Job: Running job: job\_1668610156287\_0006

22/11/17 04:30:46 INFO mapreduce.Job: Job job\_1668610156287\_0006 running in uber mode : false

22/11/17 04:30:46 INFO mapreduce.Job: map 0% reduce 0%

22/11/17 04:31:23 INFO mapreduce.Job: map 100% reduce 0%

22/11/17 04:31:25 INFO mapreduce.Job: Job job\_1668610156287\_0006 completed successfully

22/11/17 04:31:27 INFO mapreduce.Job: Counters: 30

File System Counters

FILE: Number of bytes read=0

FILE: Number of bytes written=171186

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=87

HDFS: Number of bytes written=587

HDFS: Number of read operations=4

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

Job Counters

Launched map tasks=1

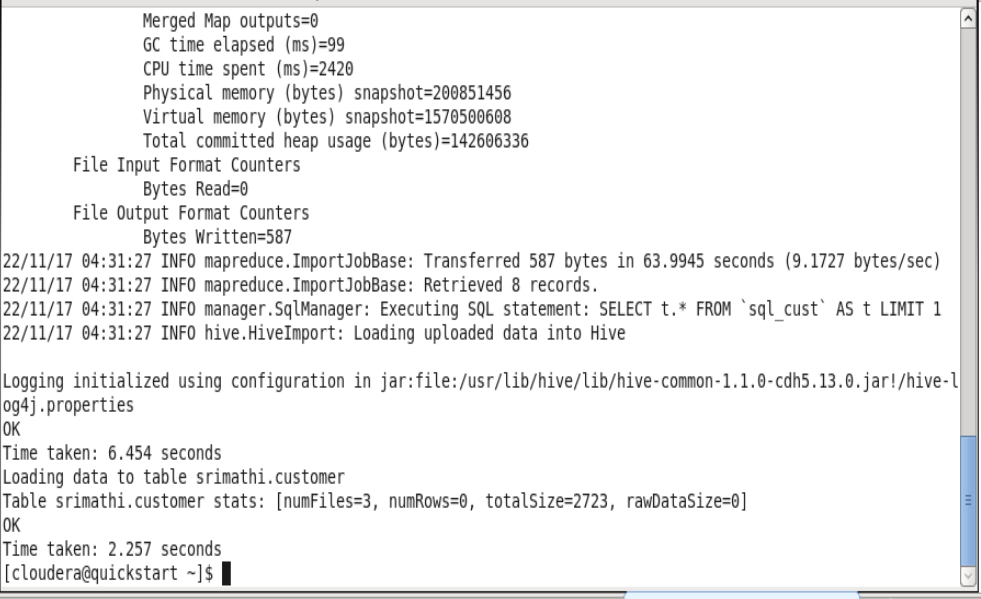
Other local map tasks=1

Total time spent by all maps in occupied slots (ms)=29822

Total time spent by all reduces in occupied slots (ms)=0

Total time spent by all map tasks (ms)=29822

****

** **

**Verify all the results are stored in MySQL table**

mysql> select \* from sql\_cust;

+------------+------------------+--------+------+-----------------------------+------------------+------------+

| CustomerID | name | gender | age | email | Month\_of\_joining | Mobile\_No |

+------------+------------------+--------+------+-----------------------------+------------------+------------+

| 12402 | Michael Parker | Male | 44 | Michael@gmail.com | January | 2147483647 |

| 17771 | Alicia Gilchrist | Male | 45 | Alicia.Gilchrist@gmail.com | February | 2147483647 |

| 15174 | Andrew Gonzales | Male | 34 | Andrew.Gonzales@outlook.com | February | 2147483647 |

| 17504 | Robert Fulcher | Male | 31 | Robert.Fulcher@gmail.com | January | 2147483647 |

| 16387 | Kyle Jewell | Female | 39 | Kyle.Jewell@gmail.com | February | 2147483647 |

| 17277 | Keely Richardson | Female | 31 | Keely.Richardson@gmail.com | May | 2147483647 |

| 14088 | Frances Workman | Female | 39 | Frances .Workman@gmail.com | February | 2147483647 |

| 17889 | Lois Robins | Male | 42 | Lois.Robins@gmail.com | February | 2147483647 |

+------------+------------------+--------+------+-----------------------------+------------------+------------+

8 rows in set (0.01 sec)

