BIG DATA – INTCDB22DW143

INTERNSHIP PROJECT

RETAIL INDUSTRY

SUBMITTED BY:

NAME	EMPLOYEE ID	GROUP
Anjani Sharma	2152442	DWH07
Anubhab Biswas	2151655	DWH07
Avinaba Karmakar	2153071	DWH07
Bratati Rout	2153035	DWH07
Debmoy Dutta	2151647	DWH07

- **Introduction**

We define the scope and objectives, and relate them to the requirements of Retail Industry. This is a good test case to see how we can manage huge amount of data using bigdata tools and techniques. The following are the tables in this proposed system - Sales, Customers and Branch.

- **Scope of the system**

The scope of the system is explained through its modules as follows

· Sales – This table is to display and sale of new or used goods to consumers for personal or household consumption. The retail trade division includes motor vehicle retail, fuel retailing, food retailing, and other forms of store-based retail. This table also includes sales unit and sales amount.

· Customer –This table is to find the loyal customers of the retail industry. Based on the sales which are given by the customers from the specific locations. Dimension Customer Master which has the product history is to find the customer loyalty.

· Branch– This table is to find the branch which is giving high profitable revenue in the retail industry. This table includes branch details as branch address, branch location, branch manager.

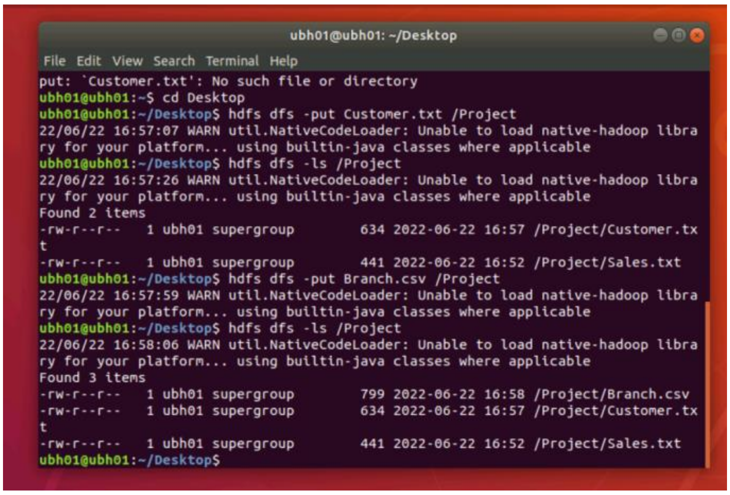
- **Objective**

This system is developed to manage the activities like finding the yearly sales revenue, sales customer, sales region, customer loyalty, how many female and male customers are there and how many units have been sold per branch and many more using hive.

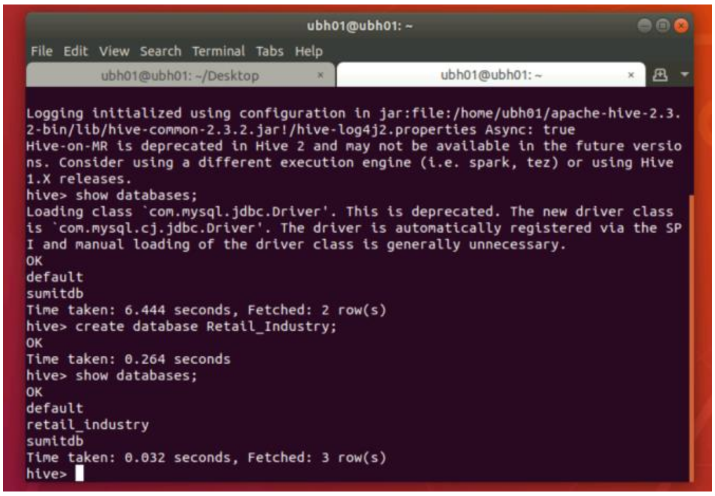
- **Procedure**

Step 1: Start your Hadoop Daemon

Step 2: Launch hive from terminal

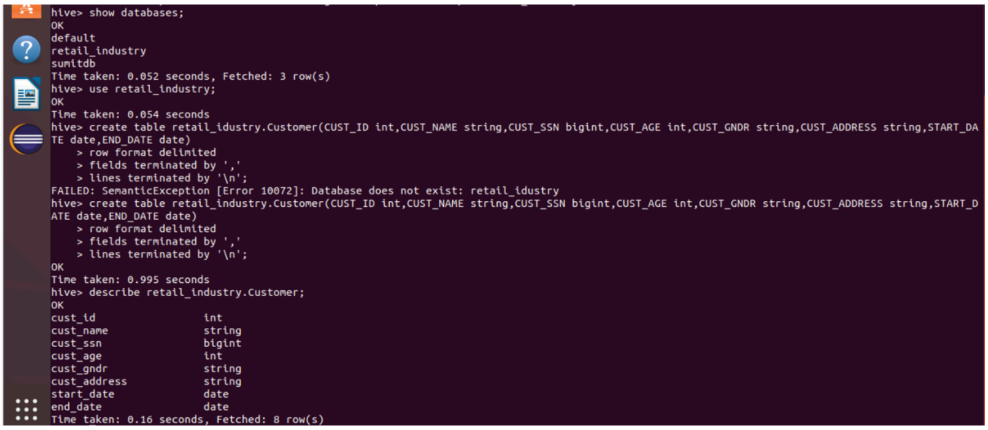


```
put: `Customer.txt': No such file or directory
ubh01@ubh01:~$ cd Desktop
ubh01@ubh01:~/Desktop$ hdfs dfs -put Customer.txt /Project
22/06/22 16:57:07 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
ubh01@ubh01:~/Desktop$ hdfs dfs -ls /Project
22/06/22 16:57:26 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 ubh01 supergroup        634 2022-06-22 16:57 /Project/Customer.tx
t
-rw-r--r--   1 ubh01 supergroup        441 2022-06-22 16:52 /Project/Sales.txt
ubh01@ubh01:~/Desktop$ hdfs dfs -put Branch.csv /Project
22/06/22 16:57:59 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
ubh01@ubh01:~/Desktop$ hdfs dfs -ls /Project
22/06/22 16:58:06 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r--   1 ubh01 supergroup        799 2022-06-22 16:58 /Project/Branch.csv
-rw-r--r--   1 ubh01 supergroup        634 2022-06-22 16:57 /Project/Customer.tx
t
-rw-r--r--   1 ubh01 supergroup        441 2022-06-22 16:52 /Project/Sales.txt
ubh01@ubh01:~/Desktop$
```



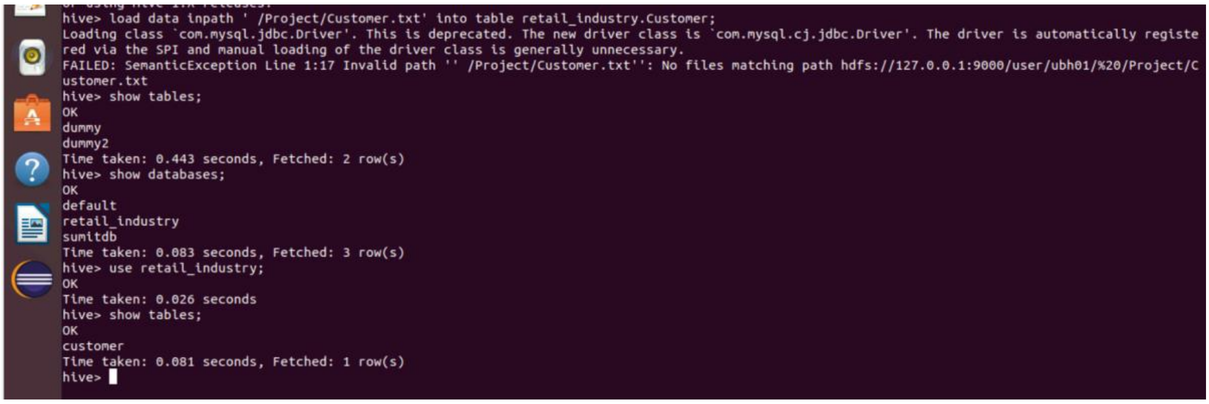
```
Logging initialized using configuration in jar:file:/home/ubh01/apache-hive-2.3.
2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
hive> show databases;
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class
is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SP
I and manual loading of the driver class is generally unnecessary.
OK
default
sumitdb
Time taken: 6.444 seconds, Fetched: 2 row(s)
hive> create database Retail_Industry;
OK
Time taken: 0.264 seconds
hive> show databases;
OK
default
retail_industry
sumitdb
Time taken: 0.032 seconds, Fetched: 3 row(s)
hive> 
```

Step 3: To insert data into the table let’s create a table

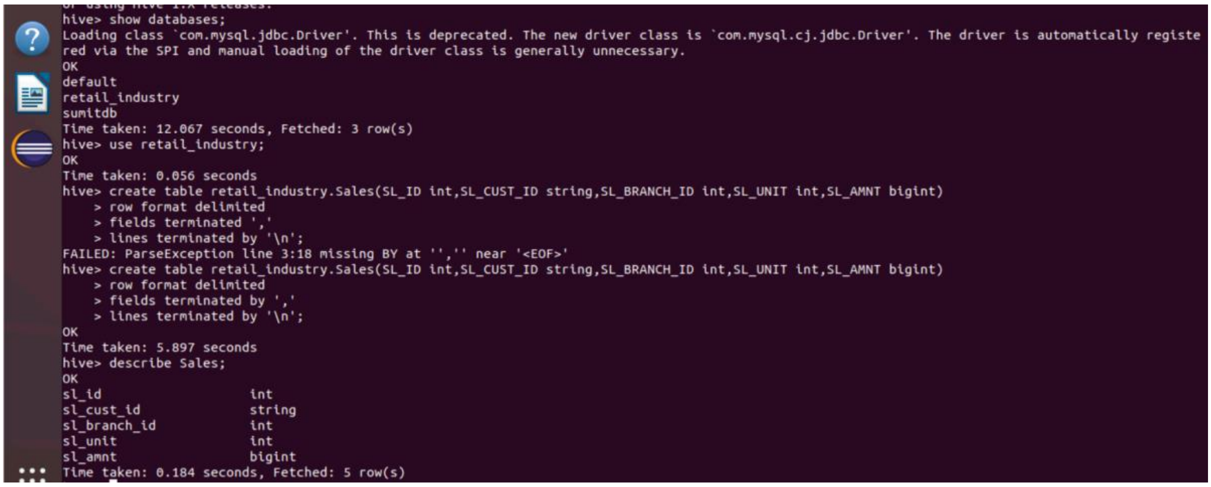


```
hive> show databases;
OK
default
retail_industry
sumitdb
Time taken: 0.052 seconds, Fetched: 3 row(s)
hive> use retail_industry;
OK
Time taken: 0.054 seconds
hive> create table retail_idustry.Customer(CUST_ID int,CUST_NAME string,CUST_SSN bigint,CUST_AGE int,CUST_GNDR string,CUST_ADDRESS string,START_DA
TE date,END_DATE date)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n';
FAILED: SemanticException [Error 10072]: Database does not exist: retail_idustry
hive> create table retail_industry.Customer(CUST_ID int,CUST_NAME string,CUST_SSN bigint,CUST_AGE int,CUST_GNDR string,CUST_ADDRESS string,START_D
ATE date,END_DATE date)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n';
OK
Time taken: 0.995 seconds
hive> describe retail_industry.Customer;
OK
cust_id                 int
cust_name               string
cust_ssn                bigint
cust_age                int
cust_gndr               string
cust_address            string
start_date              date
end_date                date
Time taken: 0.16 seconds, Fetched: 8 row(s)
```

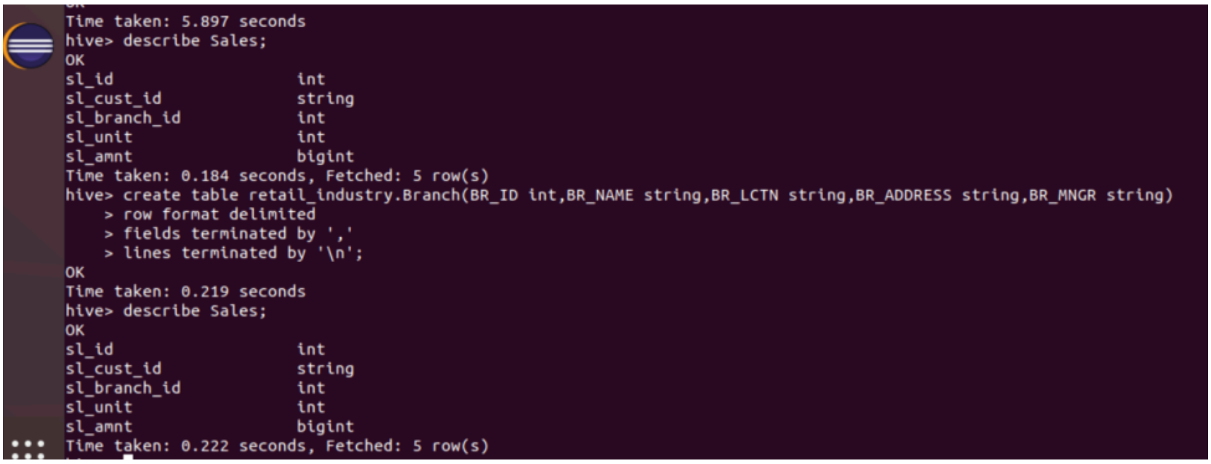
Step 4: Hive provides us the functionality to load pre-created table entities either from our local file system or from HDFS. The LOAD DATA statement is used to load data into the hive table.



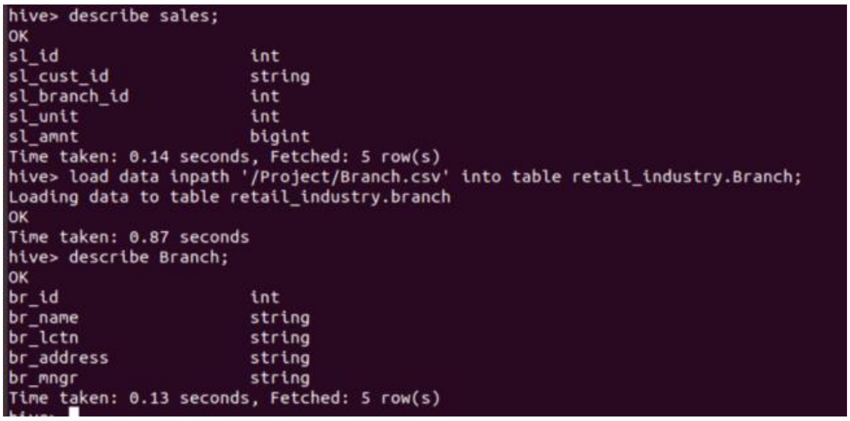
```
hive> load data inpath ' /Project/Customer.txt' into table retail_industry.Customer;
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registe
red via the SPI and manual loading of the driver class is generally unnecessary.
FAILED: SemanticException Line 1:17 Invalid path '' /Project/Customer.txt'': No files matching path hdfs://127.0.0.1:9000/user/ubh01/%20/Project/C
ustomer.txt
hive> show tables;
OK
dummy
dummy2
Time taken: 0.443 seconds, Fetched: 2 row(s)
hive> show databases;
OK
default
retail_industry
sumitdb
Time taken: 0.083 seconds, Fetched: 3 row(s)
hive> use retail_industry;
OK
Time taken: 0.026 seconds
hive> show tables;
OK
customer
Time taken: 0.081 seconds, Fetched: 1 row(s)
hive> 
```



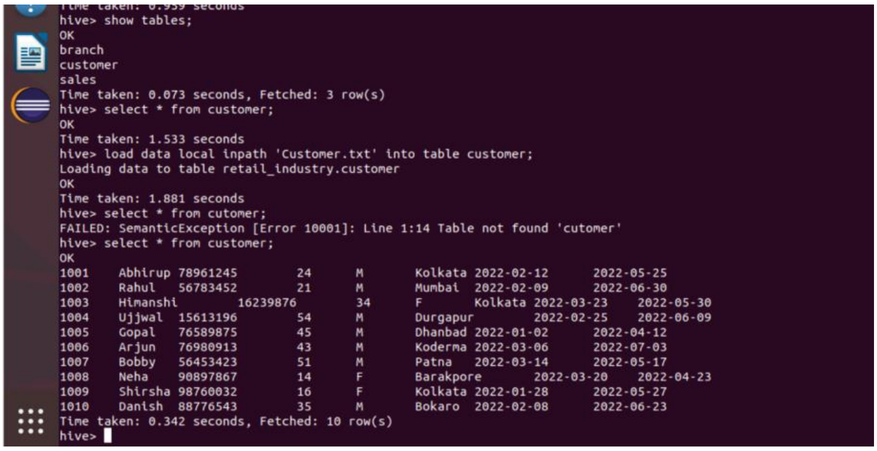
```
hive> show databases;
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registe
red via the SPI and manual loading of the driver class is generally unnecessary.
OK
default
retail_industry
sumitdb
Time taken: 12.067 seconds, Fetched: 3 row(s)
hive> use retail_industry;
OK
Time taken: 0.056 seconds
hive> create table retail_industry.Sales(SL_ID int,SL_CUST_ID string,SL_BRANCH_ID int,SL_UNIT int,SL_AMNT bigint)
    > row format delimited
    > fields terminated ','
    > lines terminated by '\n';
FAILED: ParseException line 3:18 missing BY at '',''' near '<EOF>'
hive> create table retail_industry.Sales(SL_ID int,SL_CUST_ID string,SL_BRANCH_ID int,SL_UNIT int,SL_AMNT bigint)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n';
OK
Time taken: 5.897 seconds
hive> describe Sales;
OK
sl_id                   int
sl_cust_id              string
sl_branch_id            int
sl_unit                 int
sl_amnt                 bigint
Time taken: 0.184 seconds, Fetched: 5 row(s)
```



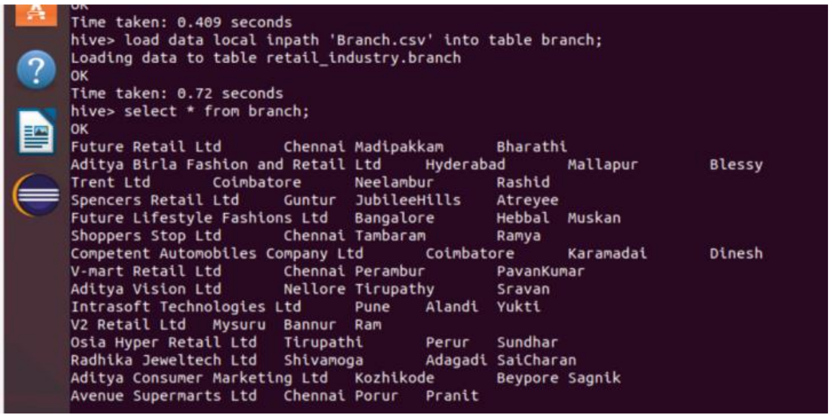
```
Time taken: 5.897 seconds
hive> describe Sales;
OK
sl_id                   int
sl_cust_id              string
sl_branch_id            int
sl_unit                 int
sl_amnt                 bigint
Time taken: 0.184 seconds, Fetched: 5 row(s)
hive> create table retail_industry.Branch(BR_ID int,BR_NAME string,BR_LCTN string,BR_ADDRESS string,BR_MNGR string)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n';
OK
Time taken: 0.219 seconds
hive> describe Sales;
OK
sl_id                   int
sl_cust_id              string
sl_branch_id            int
sl_unit                 int
sl_amnt                 bigint
Time taken: 0.222 seconds, Fetched: 5 row(s)
```



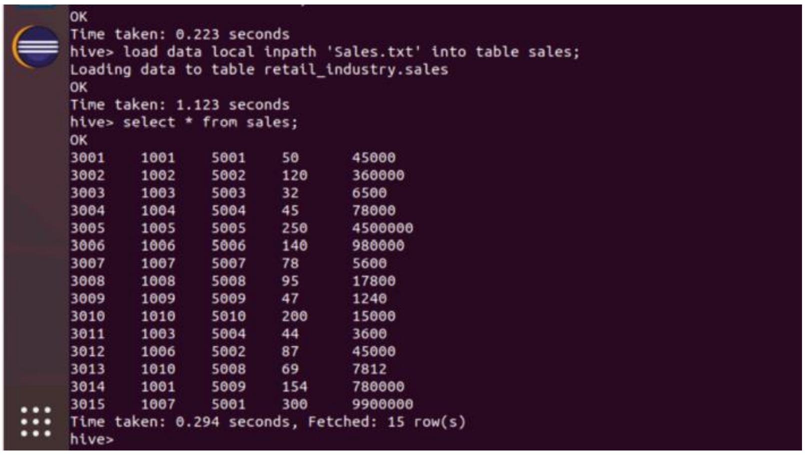
```
hive> describe sales;
OK
sl_id                   int
sl_cust_id              string
sl_branch_id            int
sl_unit                 int
sl_amnt                 bigint
Time taken: 0.14 seconds, Fetched: 5 row(s)
hive> load data inpath '/Project/Branch.csv' into table retail_industry.Branch;
Loading data to table retail_industry.branch
OK
Time taken: 0.87 seconds
hive> describe Branch;
OK
br_id                   int
br_name                 string
br_lctn                 string
br_address              string
br_mngr                 string
Time taken: 0.13 seconds, Fetched: 5 row(s)
hive>
```



```
Time taken: 0.959 seconds
hive> show tables;
OK
branch
customer
sales
Time taken: 0.073 seconds, Fetched: 3 row(s)
hive> select * from customer;
OK
Time taken: 1.533 seconds
hive> load data local inpath 'Customer.txt' into table customer;
Loading data to table retail_industry.customer
OK
Time taken: 1.881 seconds
hive> select * from cutomer;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'cutomer'
hive> select * from customer;
OK
1001    Abhirup 78961245        24      M       Kolkata 2022-02-12      2022-05-25
1002    Rahul   56783452        21      M       Mumbai  2022-02-09      2022-06-30
1003    Himanshi        16239876        34      F       Kolkata 2022-03-23      2022-05-30
1004    Ujjwal  15613196        54      M       Durgapur        2022-02-25      2022-06-09
1005    Gopal   76589875        45      M       Dhanbad 2022-01-02      2022-04-12
1006    Arjun   76980913        43      M       Koderma 2022-03-06      2022-07-03
1007    Bobby   56453423        51      M       Patna   2022-03-14      2022-05-17
1008    Neha    90897867        14      F       Barakpore       2022-03-20      2022-04-23
1009    Shirsha 98760032        16      F       Kolkata 2022-01-28      2022-05-27
1010    Danish  88776543        35      M       Bokaro  2022-02-08      2022-06-23
Time taken: 0.342 seconds, Fetched: 10 row(s)
hive>
```



```
OK
Time taken: 0.409 seconds
hive> load data local inpath 'Branch.csv' into table branch;
Loading data to table retail_industry.branch
OK
Time taken: 0.72 seconds
hive> select * from branch;
OK
Future Retail Ltd       Chennai Madipakkam      Bharathi
Aditya Birla Fashion and Retail Ltd     Hyderabad       Mallapur        Blessy
Trent Ltd       Coimbatore      Neelambur       Rashid
Spencers Retail Ltd     Guntur  JubileeHills    Atreyee
Future Lifestyle Fashions Ltd   Bangalore       Hebbal  Muskan
Shoppers Stop Ltd       Chennai Tambaram        Ramya
Competent Automobiles Company Ltd       Coimbatore      Karamadai       Dinesh
V-mart Retail Ltd       Chennai Perambur        PavanKumar
Aditya Vision Ltd       Nellore Tirupathy       Sravan
Intrasoft Technologies Ltd      Pune    Alandi  Yukti
V2 Retail Ltd   Mysuru  Bannur  Ram
Osia Hyper Retail Ltd   Tirupathi       Perur   Sundhar
Radhika Jeweltech Ltd   Shivamoga       Adagadi SaiCharan
Aditya Consumer Marketing Ltd   Kozhikode       Beypore Sagnik
Avenue Supermarts Ltd   Chennai Porur   Pranit
```



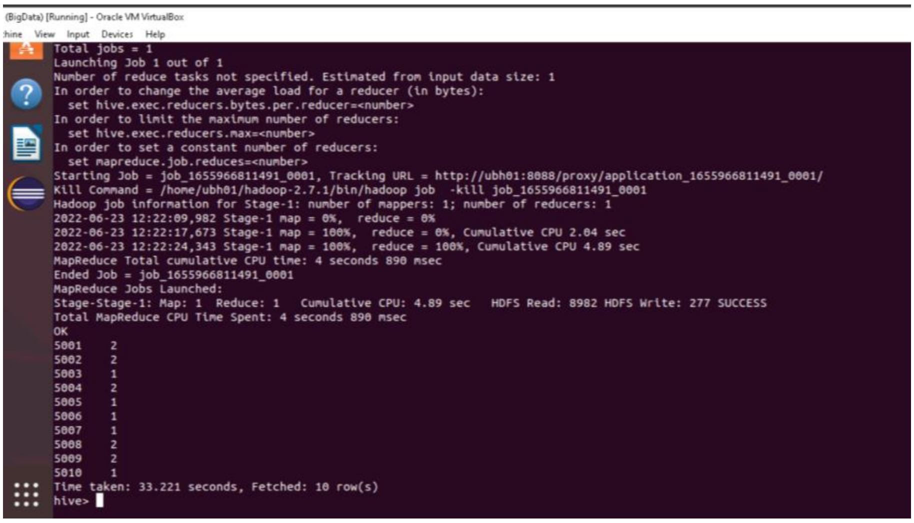
```
OK
Time taken: 0.223 seconds
hive> load data local inpath 'Sales.txt' into table sales;
Loading data to table retail_industry.sales
OK
Time taken: 1.123 seconds
hive> select * from sales;
OK
3001    1001    5001    50      45000
3002    1002    5002    120     360000
3003    1003    5003    32      6500
3004    1004    5004    45      78000
3005    1005    5005    250     4500000
3006    1006    5006    140     980000
3007    1007    5007    78      5600
3008    1008    5008    95      17800
3009    1009    5009    47      1240
3010    1010    5010    200     15000
3011    1003    5004    44      3600
3012    1006    5002    87      45000
3013    1010    5008    69      7812
3014    1001    5009    154     780000
3015    1007    5001    300     9900000
Time taken: 0.294 seconds, Fetched: 15 row(s)
hive>
```

Step 6: Writing queries to analyse the data present in sales, customer and branch files.

SALES.TXT

- **How many units have been sold per branch?**

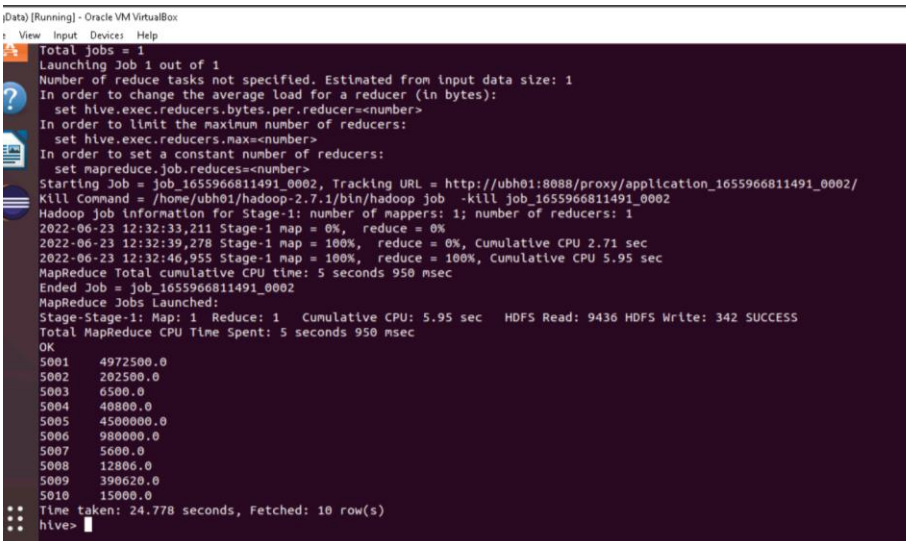
Select SL_BRANCH_ID , count(SL_UNIT) as no_of_units from sales group by SL_BRANCH_ID;


(BigData) [Running] - Oracle VM VirtualBox
hine View Input Devices Help

```
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_1655966811491_0001, Tracking URL = http://ubh01:8088/proxy/application_1655966811491_0001/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655966811491_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 12:22:09,982 Stage-1 map = 0%,  reduce = 0%
2022-06-23 12:22:17,673 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.04 sec
2022-06-23 12:22:24,343 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.89 sec
MapReduce Total cumulative CPU time: 4 seconds 890 msec
Ended Job = job_1655966811491_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.89 sec   HDFS Read: 8982 HDFS Write: 277 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 890 msec
OK
5001    2
5002    2
5003    1
5004    2
5005    1
5006    1
5007    1
5008    2
5009    2
5010    1
Time taken: 33.221 seconds, Fetched: 10 row(s)
hive>
```

- **Average sales amount per branch**

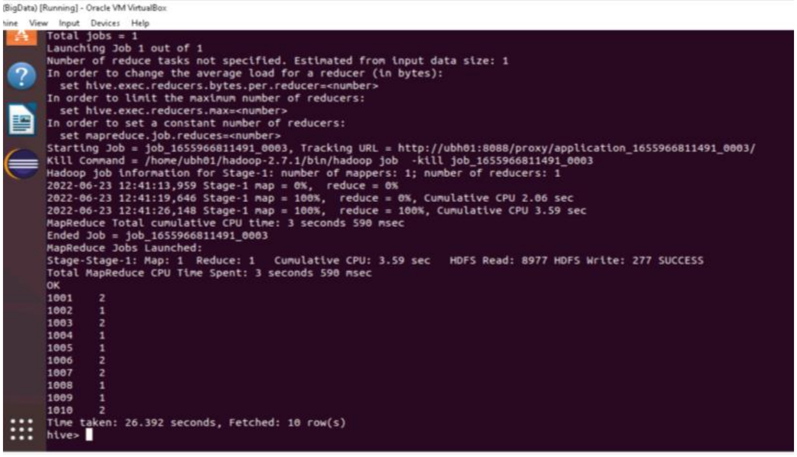
Select SL_BRANCH_ID,avg(SL_AMNT) from sales group by SL_BRANCH_ID;



```
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_1655966811491_0002, Tracking URL = http://ubh01:8088/proxy/application_1655966811491_0002/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655966811491_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 12:32:33,211 Stage-1 map = 0%,  reduce = 0%
2022-06-23 12:32:39,278 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.71 sec
2022-06-23 12:32:46,955 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.95 sec
MapReduce Total cumulative CPU time: 5 seconds 950 msec
Ended Job = job_1655966811491_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.95 sec   HDFS Read: 9436 HDFS Write: 342 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 950 msec
OK
5001    4972500.0
5002    202500.0
5003    6500.0
5004    40800.0
5005    4500000.0
5006    980000.0
5007    5600.0
5008    12806.0
5009    390620.0
5010    15000.0
Time taken: 24.778 seconds, Fetched: 10 row(s)
hive>
```

- **Total units bought by each customer.**

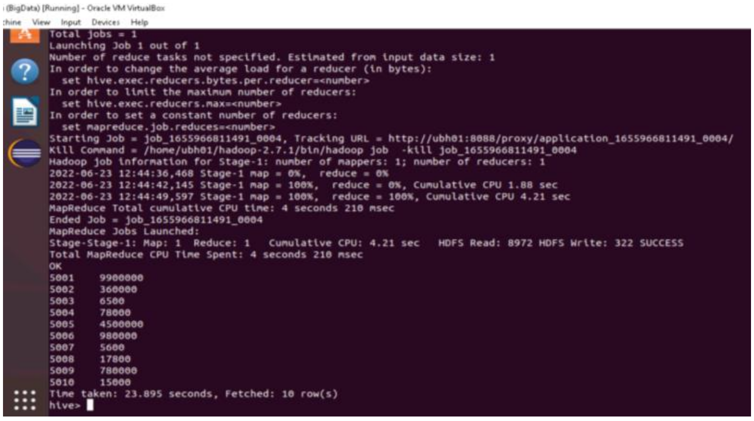
Select SL_CUST_ID , count(SL_UNIT) as no_of_units from sales group by SL_CUST_ID;


(BigData) [Running] - Oracle VM VirtualBox
hine View Input Devices Help

```
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_1655966811491_0003, Tracking URL = http://ubh01:8088/proxy/application_1655966811491_0003/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655966811491_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 12:41:13,959 Stage-1 map = 0%,  reduce = 0%
2022-06-23 12:41:19,646 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.06 sec
2022-06-23 12:41:26,148 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.59 sec
MapReduce Total cumulative CPU time: 3 seconds 590 msec
Ended Job = job_1655966811491_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.59 sec   HDFS Read: 8977 HDFS Write: 277 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 590 msec
OK
1001    2
1002    1
1003    2
1004    1
1005    1
1006    2
1007    2
1008    1
1009    1
1010    2
Time taken: 26.392 seconds, Fetched: 10 row(s)
hive>
```

- **Maximum sales in a particular branch.**

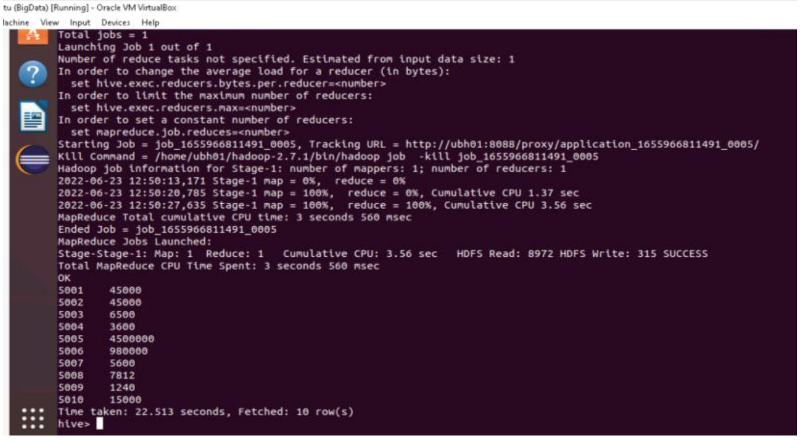
Select SL_BRANCH_ID, max(SL_AMNT) from sales group by SL_BRANCH_ID;


(BigData) [Running] - Oracle VM VirtualBox
chine View Input Devices Help

```
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_1655966811491_0004, Tracking URL = http://ubh01:8088/proxy/application_1655966811491_0004/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655966811491_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 12:44:36,468 Stage-1 map = 0%,  reduce = 0%
2022-06-23 12:44:42,145 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.88 sec
2022-06-23 12:44:49,597 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.21 sec
MapReduce Total cumulative CPU time: 4 seconds 210 msec
Ended Job = job_1655966811491_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.21 sec   HDFS Read: 8972 HDFS Write: 322 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 210 msec
OK
5001    9900000
5002    360000
5003    6500
5004    78000
5005    4500000
5006    980000
5007    5600
5008    17800
5009    780000
5010    15000
Time taken: 23.895 seconds, Fetched: 10 row(s)
hive>
```

- **Minimum sales in a particular branch**.

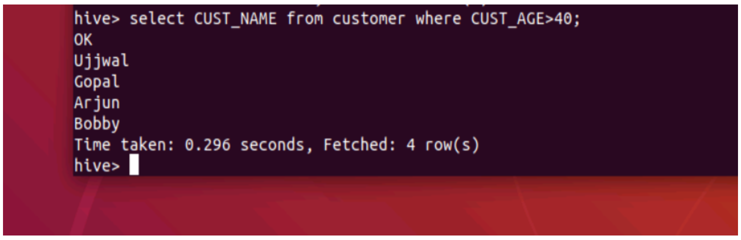
  Select SL_BRANCH_ID, min(SL_AMNT) from sales group by SL_BRANCH_ID;


tu (BigData) [Running] - Oracle VM VirtualBox
lachine View Input Devices Help

```
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_1655966811491_0005, Tracking URL = http://ubh01:8088/proxy/application_1655966811491_0005/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655966811491_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 12:50:13,171 Stage-1 map = 0%,  reduce = 0%
2022-06-23 12:50:20,785 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.37 sec
2022-06-23 12:50:27,635 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.56 sec
MapReduce Total cumulative CPU time: 3 seconds 560 msec
Ended Job = job_1655966811491_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.56 sec   HDFS Read: 8972 HDFS Write: 315 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 560 msec
OK
5001    45000
5002    45000
5003    6500
5004    3600
5005    4500000
5006    980000
5007    5600
5008    7812
5009    1240
5010    15000
Time taken: 22.513 seconds, Fetched: 10 row(s)
hive>
```

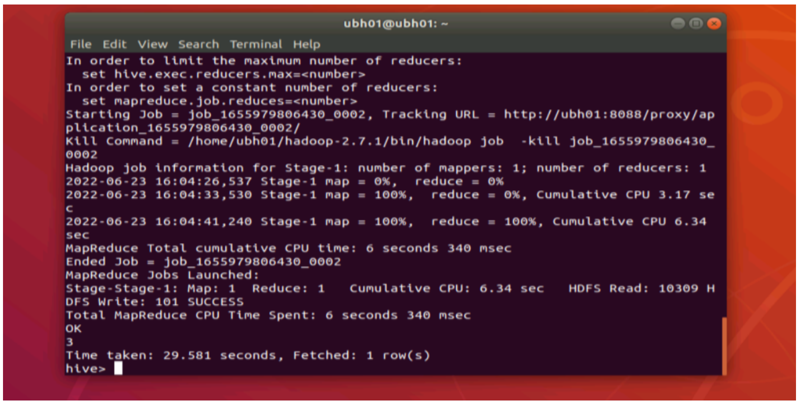
CUSTOMER.TXT

- **Display the customer names who are above 40?**
  Select CUST_NAME from customer where CUST_AGE>40;



```
hive> select CUST_NAME from customer where CUST_AGE>40;
OK
Ujjwal
Gopal
Arjun
Bobby
Time taken: 0.296 seconds, Fetched: 4 row(s)
hive> 
```

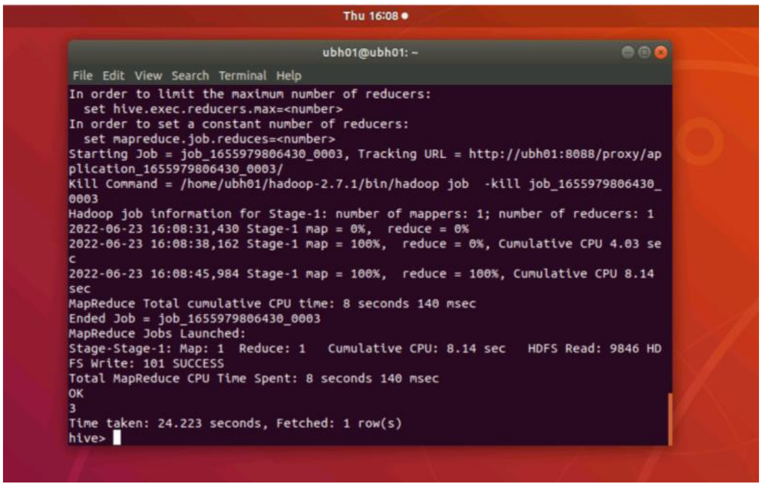
- **How many customers are there from a particular city(Kolkata).**
  Select count(CUST_NAME) from customer where CUST_ADDRESS = ‘Kolkata’;



```
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_1655979806430_0002, Tracking URL = http://ubh01:8088/proxy/ap
plication_1655979806430_0002/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655979806430_
0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 16:04:26,537 Stage-1 map = 0%,  reduce = 0%
2022-06-23 16:04:33,530 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.17 se
c
2022-06-23 16:04:41,240 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.34
sec
MapReduce Total cumulative CPU time: 6 seconds 340 msec
Ended Job = job_1655979806430_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.34 sec   HDFS Read: 10309 H
DFS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 340 msec
OK
3
Time taken: 29.581 seconds, Fetched: 1 row(s)
hive> 
```

- **How many female customers are there**?

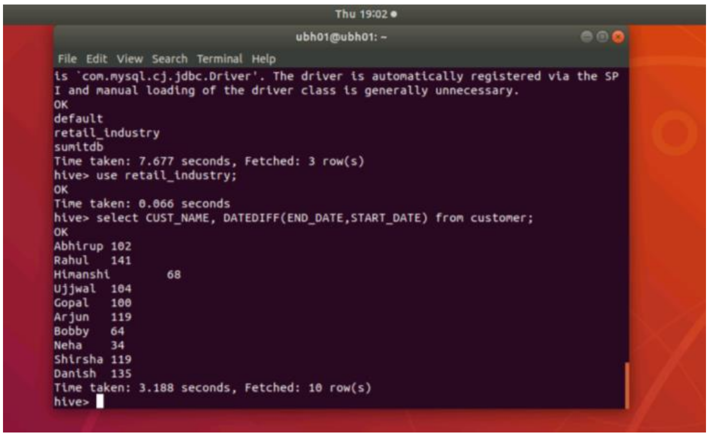
Select count(cust_id) as female_customers from customer where cust_gndr='F';



```
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_1655979806430_0003, Tracking URL = http://ubh01:8088/proxy/ap
plication_1655979806430_0003/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655979806430_
0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 16:08:31,430 Stage-1 map = 0%,  reduce = 0%
2022-06-23 16:08:38,162 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.03 se
c
2022-06-23 16:08:45,984 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.14
sec
MapReduce Total cumulative CPU time: 8 seconds 140 msec
Ended Job = job_1655979806430_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.14 sec   HDFS Read: 9846 HD
FS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 140 msec
OK
3
Time taken: 24.223 seconds, Fetched: 1 row(s)
hive> 
```

- **Display customer names and loyalty of customers in number of days from the customer table.**

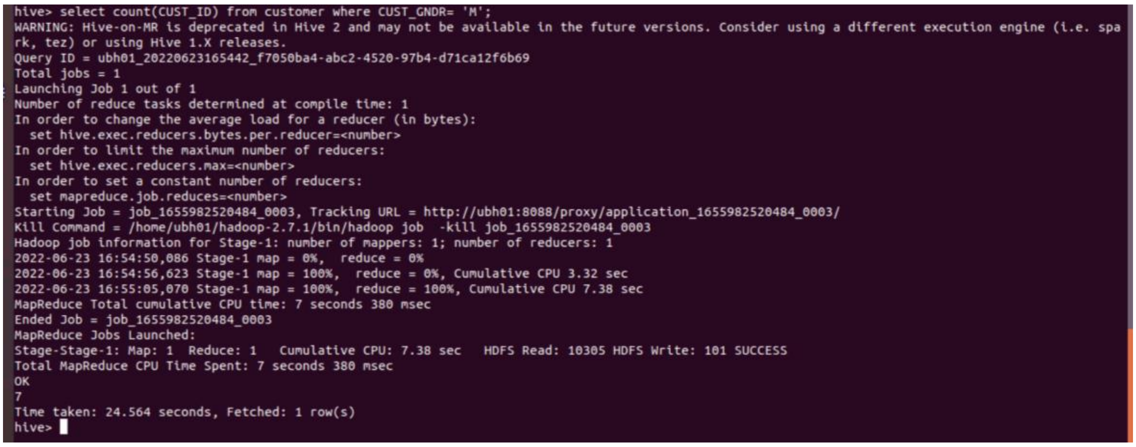
  Select CUST_NAME, DATEDIFF(END_DATE,START_DATE) from customer;



```
is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SP
I and manual loading of the driver class is generally unnecessary.
OK
default
retail_industry
sumitdb
Time taken: 7.677 seconds, Fetched: 3 row(s)
hive> use retail_industry;
OK
Time taken: 0.066 seconds
hive> select CUST_NAME, DATEDIFF(END_DATE,START_DATE) from customer;
OK
Abhirup 102
Rahul   141
Himanshi        68
Ujjwal  104
Gopal   100
Arjun   119
Bobby   64
Neha    34
Shirsha 119
Danish  135
Time taken: 3.188 seconds, Fetched: 10 row(s)
hive> 
```

- **How many male customers are there?**

  Select count(cust_id) as male_customers from customer where cust_gndr='M';

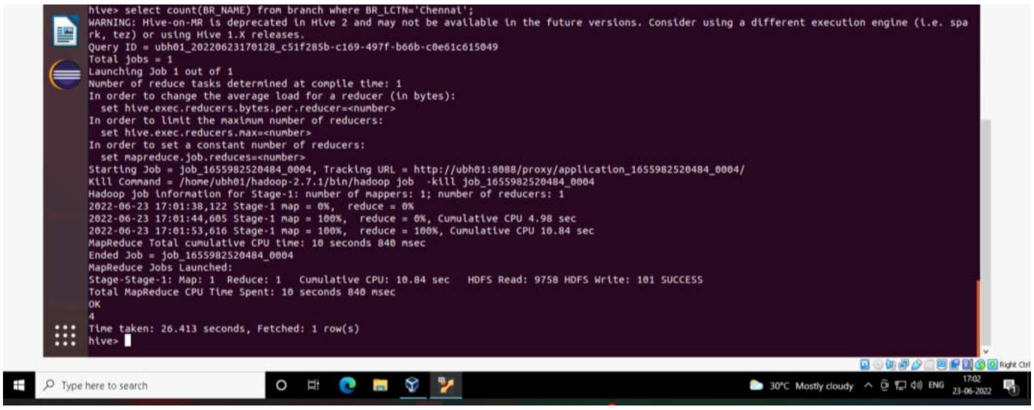


```
hive> select count(CUST_ID) from customer where CUST_GNDR= 'M';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spa
rk, tez) or using Hive 1.X releases.
Query ID = ubh01_20220623165442_f7050ba4-abc2-4520-97b4-d71ca12f6b69
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1655982520484_0003, Tracking URL = http://ubh01:8088/proxy/application_1655982520484_0003/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655982520484_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 16:54:50,086 Stage-1 map = 0%,  reduce = 0%
2022-06-23 16:54:56,623 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.32 sec
2022-06-23 16:55:05,070 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.38 sec
MapReduce Total cumulative CPU time: 7 seconds 380 msec
Ended Job = job_1655982520484_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.38 sec   HDFS Read: 10305 HDFS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 380 msec
OK
7
Time taken: 24.564 seconds, Fetched: 1 row(s)
hive> 
```

BRANCH.CSV

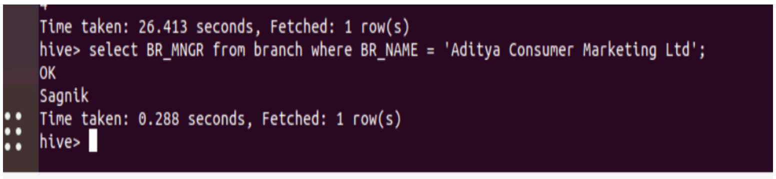
- **How many branches are there in a particular city?**

  Select count(BR_NAME) from branch where BR_LCTN = ‘Chennai’;



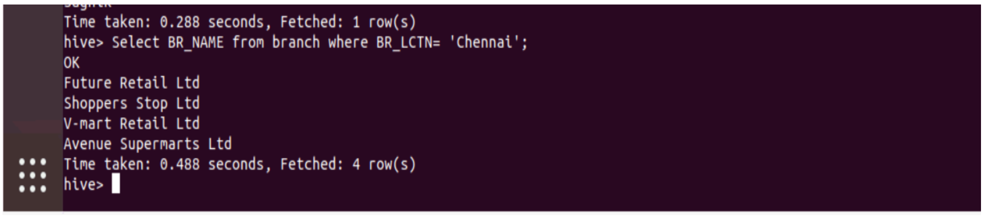
```
hive> select count(BR_NAME) from branch where BR_LCTN='Chennai';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spa
rk, tez) or using Hive 1.X releases.
Query ID = ubh01_20220623170128_c51f285b-c169-497f-b66b-c0e61c615049
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1655982520484_0004, Tracking URL = http://ubh01:8088/proxy/application_1655982520484_0004/
Kill Command = /home/ubh01/hadoop-2.7.1/bin/hadoop job  -kill job_1655982520484_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-06-23 17:01:38,122 Stage-1 map = 0%,  reduce = 0%
2022-06-23 17:01:44,605 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.98 sec
2022-06-23 17:01:53,616 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.84 sec
MapReduce Total cumulative CPU time: 10 seconds 840 msec
Ended Job = job_1655982520484_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.84 sec   HDFS Read: 9758 HDFS Write: 101 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 840 msec
OK
4
Time taken: 26.413 seconds, Fetched: 1 row(s)
hive> 
```

- **Who is the manager of Aditya Consumer Marketing Ltd.**
  Select br_mgnr from branch where br_name='Aditya Consumer Marketing Ltd';



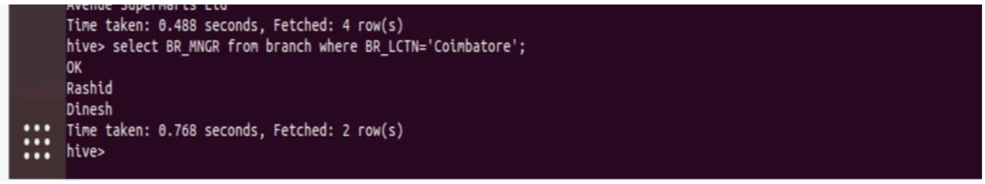
```
Time taken: 26.413 seconds, Fetched: 1 row(s)
hive> select BR_MNGR from branch where BR_NAME = 'Aditya Consumer Marketing Ltd';
OK
Sagnik
Time taken: 0.288 seconds, Fetched: 1 row(s)
hive> 
```

- **Display all the branches in particular city**
  Select br_name from branch where br_lctn='Chennai';



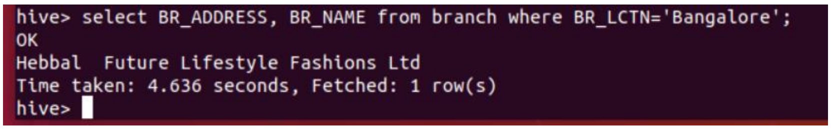
```
Sagnik
Time taken: 0.288 seconds, Fetched: 1 row(s)
hive> Select BR_NAME from branch where BR_LCTN= 'Chennai';
OK
Future Retail Ltd
Shoppers Stop Ltd
V-mart Retail Ltd
Avenue Supermarts Ltd
Time taken: 0.488 seconds, Fetched: 4 row(s)
hive> 
```

- **Display the names of all the managers in Coimbatore**.
  Select br_mgnr from branch where br_lctn='Coimbatore';



```
Avenue Supermarts Ltd
Time taken: 0.488 seconds, Fetched: 4 row(s)
hive> select BR_MNGR from branch where BR_LCTN='Coimbatore';
OK
Rashid
Dinesh
Time taken: 0.768 seconds, Fetched: 2 row(s)
hive>
```

- **Display the branch location and branch name of a branch in that particular city**.
  Select br_address, br_name from branch where br_lctn='Bangalore';



```
hive> select BR_ADDRESS, BR_NAME from branch where BR_LCTN='Bangalore';
OK
Hebbal  Future Lifestyle Fashions Ltd
Time taken: 4.636 seconds, Fetched: 1 row(s)
hive> 
```

Conclusion

Finally we will make our own review and conclusion on Hive, based on our project.

Hadoop is a flexible and open source implementation for analyzing large datasets using map-reduce, but relatively difficult to implement and programming.

As a result, Hive provides easy to use platform for the users who are comfortable in SQL language for map-reduce programming. The performance discrepancies between Hive and conventional SQL rely on the difference between single node operation and distributed framework. In real word experient, how difficult reduce tasks performs on a query determines the performance of the distributed framework (Hive), which can also been seen as large context switch overhead.