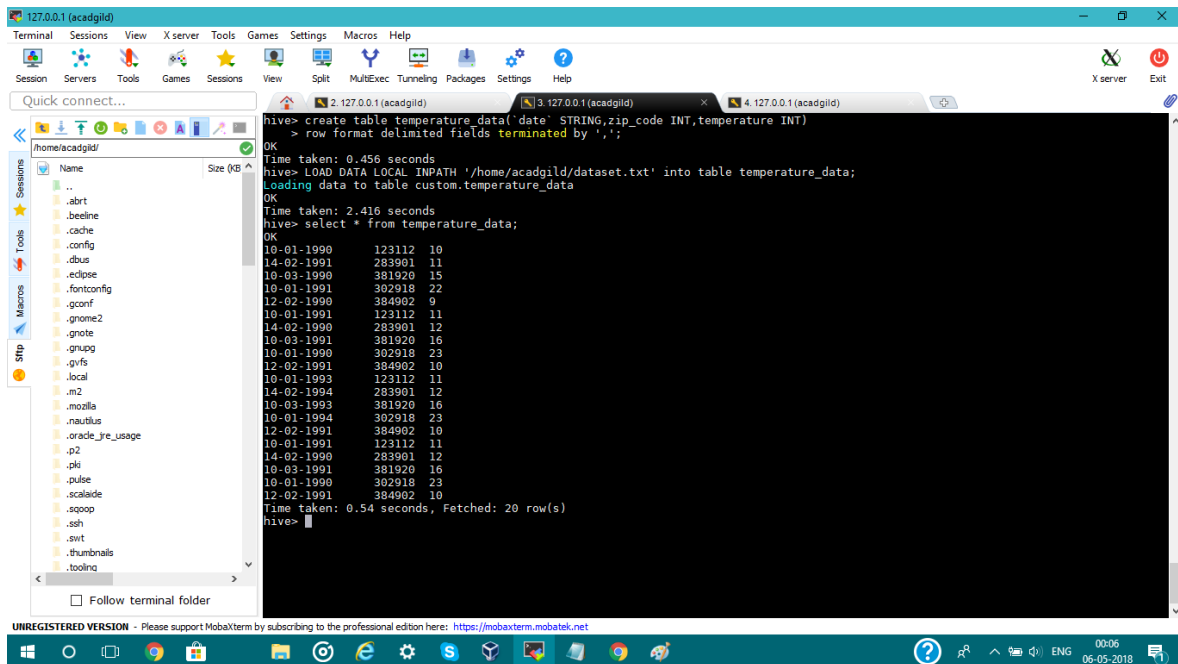


Assignment 8

Task1:

Create a database named 'custom' and temperature data within it. Load data from dataset.txt into it.



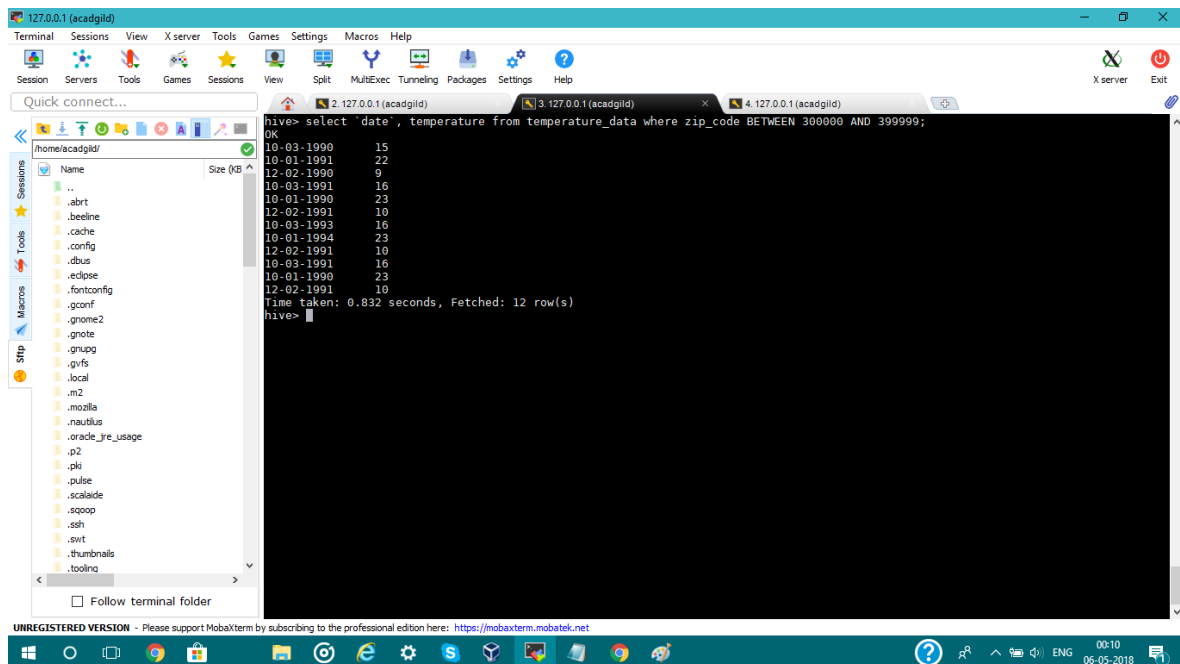
The screenshot shows a MobaXterm terminal window with three tabs. The active tab is titled '127.0.0.1 (acadgild)'. The terminal displays the following commands and output:

```
hive> create database custom;
OK
Time taken: 0.456 seconds
hive> use custom;
OK
Time taken: 0.001 seconds
hive> create table temperature_data('date' STRING,zip_code INT,temperature INT)
> row format delimited fields terminated by ',';
OK
Time taken: 0.456 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/dataset.txt' into table temperature_data;
Loading data to table custom.temperature_data
Time taken: 2.416 seconds
hive> select * from temperature_data;
OK
10-01-1990      123112      10
14-02-1991      283901      11
10-03-1990      381920      15
10-01-1991      302918      22
12-02-1990      384902      9
10-01-1991      123112      11
14-02-1990      283901      12
10-03-1991      381920      16
10-01-1990      302918      23
12-02-1991      384902      10
10-01-1993      123112      11
14-02-1994      283901      12
10-03-1993      381920      16
10-01-1994      302918      23
12-02-1991      384902      10
10-01-1991      123112      11
14-02-1990      283901      12
10-03-1991      381920      16
10-01-1990      302918      23
12-02-1991      384902      10
Time taken: 0.54 seconds, Fetched: 20 row(s)
hive>
```

- Database is created using the command: create database custom;
- Then 'use custom' command is given use this database.
- Then after that the required table is created and data is loaded into it from dataset.txt.

Task2

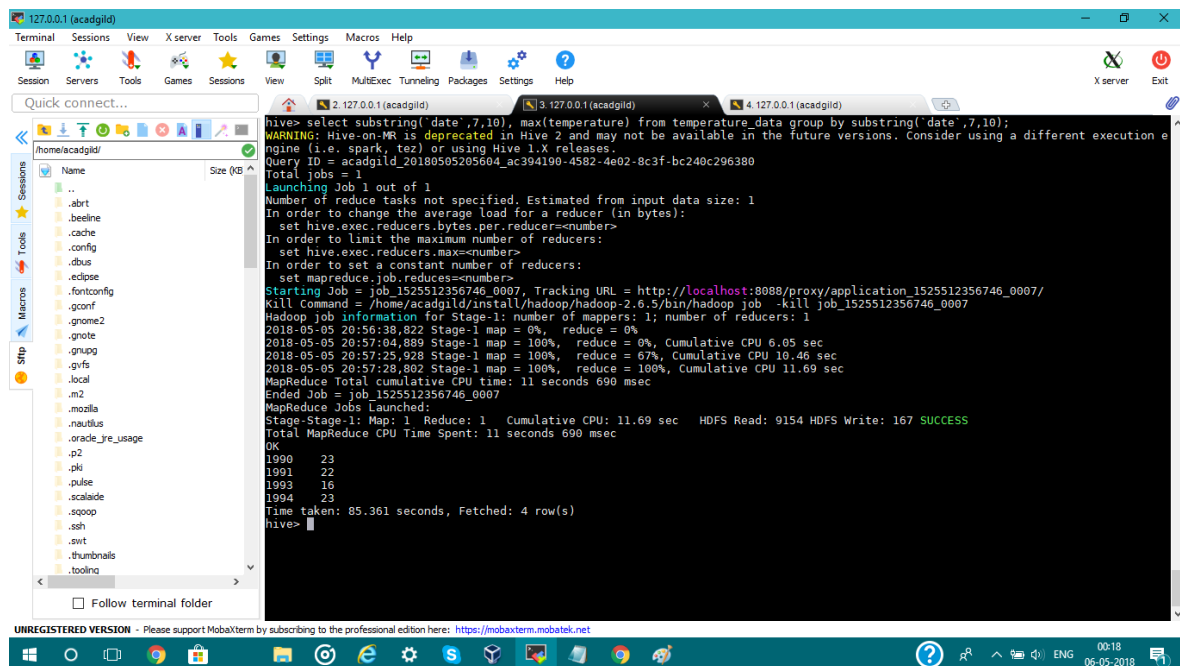
- a. Fetch data and temperature from the table where the zip code is between 300000 and 399999



The screenshot shows a MobaXterm window with a terminal session. The terminal displays a Hive query execution for a table named 'temperature_data'. The query filters for zip codes between 300000 and 399999. The output shows 12 rows of data, including date and temperature. The time taken for the query is 0.832 seconds.

```
hive> select date, temperature from temperature_data where zip_code BETWEEN 300000 AND 399999;
OK
10-03-1990      15
10-01-1991      22
12-02-1990       9
10-03-1991      16
10-01-1990      23
12-02-1991      10
10-03-1993      16
10-01-1994      23
12-02-1991      10
10-03-1991      16
10-01-1990      23
12-02-1991      10
Time taken: 0.832 seconds, Fetched: 12 row(s)
hive>
```

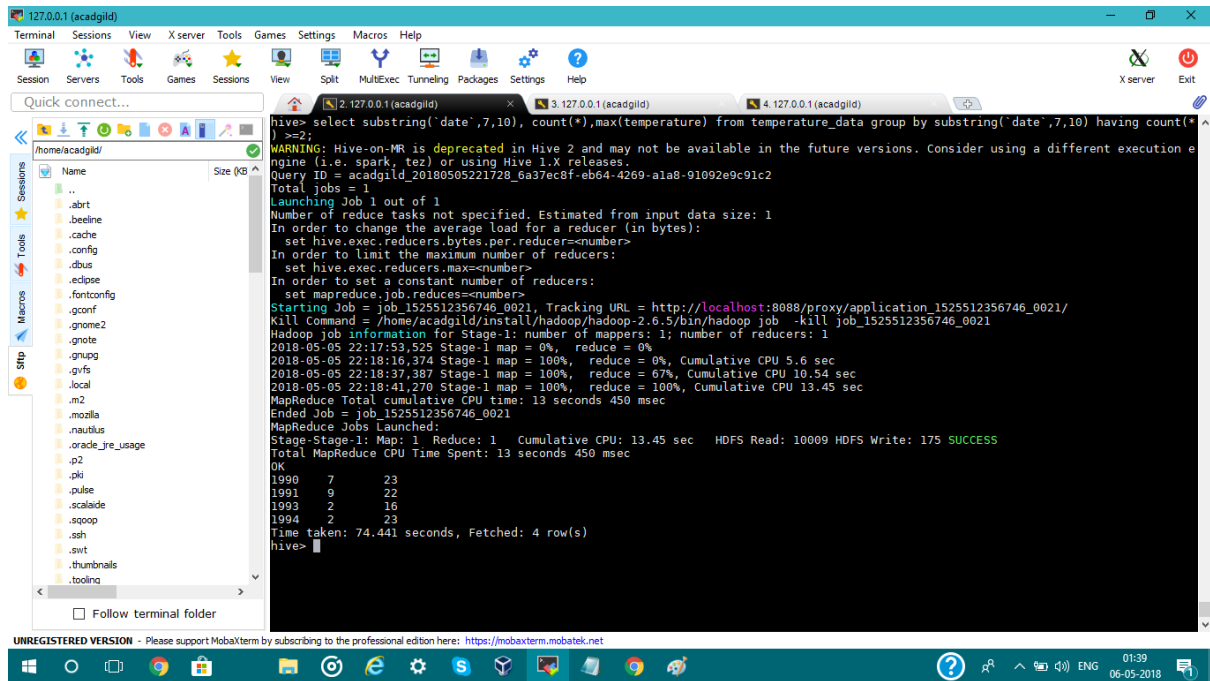
- b. Calculate maximum temperature corresponding to every year from temperature_data table.



The screenshot shows a MobaXterm window with a terminal session. The terminal displays a Hive query execution for a table named 'temperature_data'. The query calculates the maximum temperature for each year. The output shows 4 rows of data, including year and maximum temperature. The time taken for the query is 85.361 seconds.

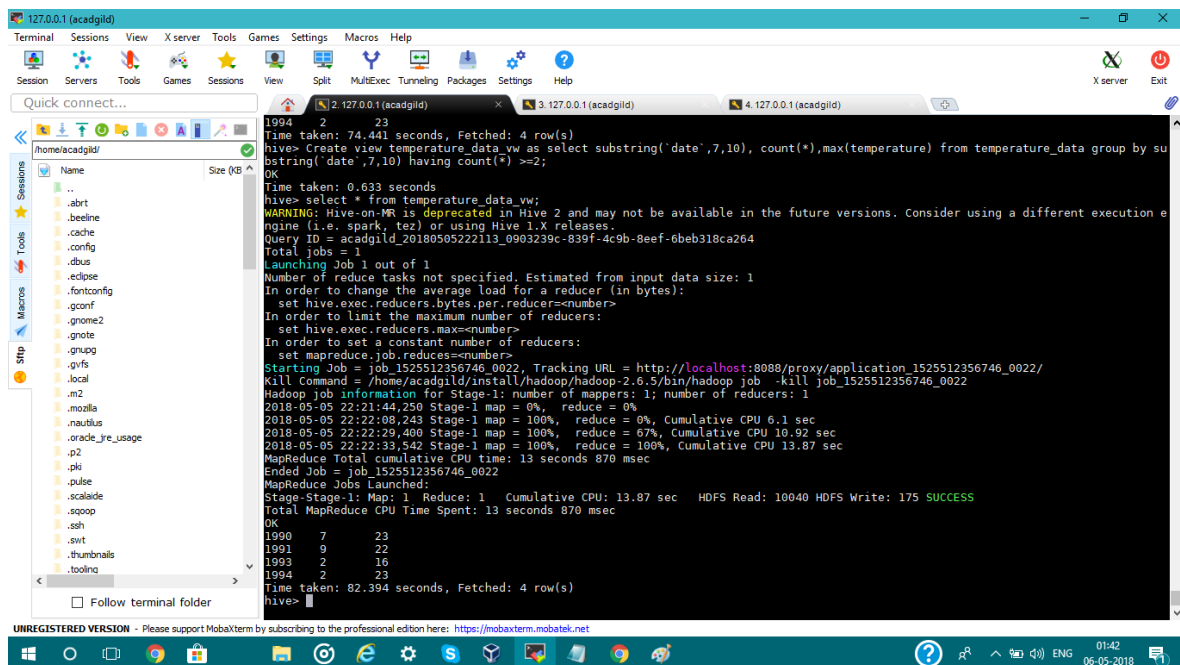
```
hive> select substring('date',7,10), max(temperature) from temperature_data group by substring('date',7,10);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180505205604_ac394190-4582-4e02-8c3f-bc240c296380
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.reducers=<number>
Starting Job = job_1525512356746_0007, Tracking URL = http://localhost:8088/proxy/application_1525512356746_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1525512356746_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-05 20:56:38,822 Stage-1 map = 0%, reduce = 0%
2018-05-05 20:57:04,889 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.05 sec
2018-05-05 20:57:25,928 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 10.46 sec
2018-05-05 20:57:28,802 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.69 sec
MapReduce Total cumulative CPU time: 11 seconds 690 msec
Ended Job = job_1525512356746_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 11.69 sec HDFS Read: 9154 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 690 msec
OK
1990      23
1991      22
1993      16
1994      23
Time taken: 85.361 seconds, Fetched: 4 row(s)
hive>
```

- c. Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.



```
hive> select substring('date',7,10), count(*), max(temperature) from temperature_data group by substring('date',7,10) having count(*) >=2;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180505221728_6a37ec8f-eb64-4269-ala8-91092e9c91c2
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_1525512356746_0021, Tracking URL = http://localhost:8088/proxy/application_1525512356746_0021/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1525512356746_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-05 22:17:53,525 Stage-1 map = 0%, reduce = 0%
2018-05-05 22:18:16,374 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.6 sec
2018-05-05 22:18:37,387 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 10.54 sec
2018-05-05 22:18:41,270 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.45 sec
MapReduce Total cumulative CPU time: 13 seconds 450 msec
Ended Job = job_1525512356746_0021
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 13.45 sec HDFS Read: 10009 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 450 msec
OK
1990 7 23
1991 9 22
1993 2 16
1994 2 23
Time taken: 74.441 seconds, Fetched: 4 row(s)
hive>
```

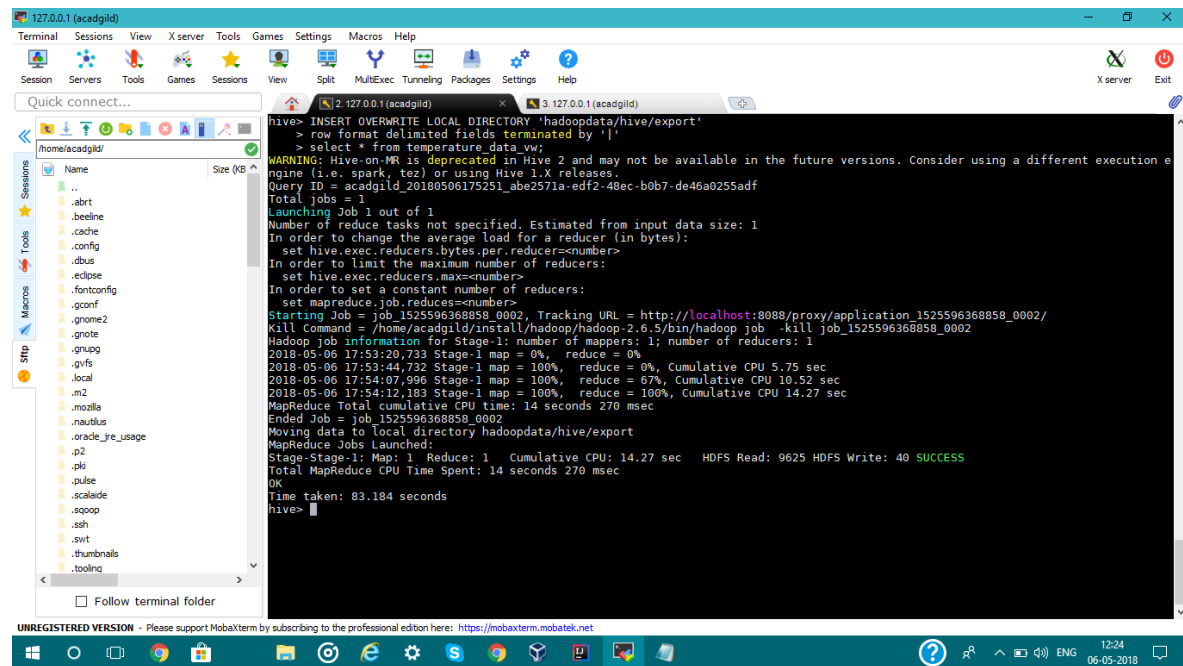
- d. Create a view on the top of last query, name it temperature_data_vw.



```
1994 2 23
Time taken: 74.441 seconds, Fetched: 4 row(s)
hive> create view temperature_data_vw as select substring('date',7,10), count(*), max(temperature) from temperature_data group by substring('date',7,10) having count(*) >=2;
Time taken: 0.633 seconds
hive> select * from temperature_data_vw;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180505222113_0903239c-839f-4c9b-8eef-6beb318ca264
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_1525512356746_0022, Tracking URL = http://localhost:8088/proxy/application_1525512356746_0022/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1525512356746_0022
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-05 22:21:44,250 Stage-1 map = 0%, reduce = 0%
2018-05-05 22:22:08,243 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.1 sec
2018-05-05 22:22:29,408 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 10.92 sec
2018-05-05 22:22:33,542 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.87 sec
MapReduce Total cumulative CPU time: 13 seconds 870 msec
Ended Job = job_1525512356746_0022
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 13.87 sec HDFS Read: 10040 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 13 seconds 870 msec
OK
1990 7 23
1991 9 22
1993 2 16
1994 2 23
Time taken: 82.394 seconds, Fetched: 4 row(s)
hive>
```

e. List of employees having no entries in employee_expenses file

Command

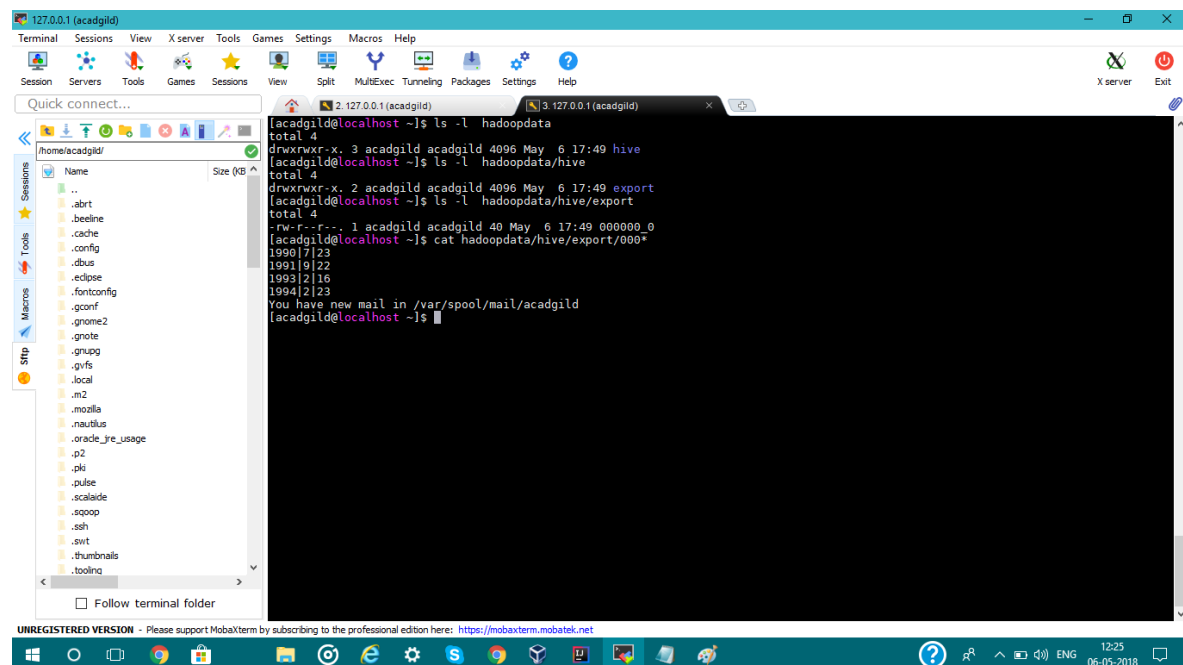


The screenshot shows the MobaXterm interface with a terminal window. The terminal displays a Hive command to insert data from a view into a new directory. The output shows the job execution details, including the number of mappers and reducers, and the final success status.

```
hive> INSERT OVERWRITE LOCAL DIRECTORY 'hadoopdata/hive/export'
> row format delimited fields terminated by '|'
> select * from temperature_data vw;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180506175251_abe2571a-edf2-48ec-b0b7-de46a0255adf
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_1525596368858_0002, Tracking URL = http://localhost:8088/proxy/application_1525596368858_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1525596368858_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-06 17:53:20,733 Stage-1 map = 0%, reduce = 0%
2018-05-06 17:53:44,732 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.75 sec
2018-05-06 17:54:07,998 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 10.52 sec
2018-05-06 17:54:12,183 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 14.27 sec
MapReduce Total cumulative CPU time: 14 seconds 270 msec
Ended Job = job_1525596368858_0002
Moving data to local directory hadoopdata/hive/export
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 14.27 sec HDFS Read: 9625 HDFS Write: 40 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 270 msec
OK
Time taken: 83.184 seconds
hive>
```

Here, we are inserting data from the view, created from above query, into a new directory (hadoopdata/hive/export) on local. The fields in this are | separated.

Output



The screenshot shows the MobaXterm interface with a terminal window. The terminal displays the output of the Hive command, showing the directory structure and the data inserted into the new directory. The output is displayed in a color-coded format.

```
[acadgild@localhost ~]$ ls -l hadoopdata
total 4
drwxrwxr-x. 3 acadgild acadgild 4096 May 6 17:49 hive
[acadgild@localhost ~]$ ls -l hadoopdata/hive
total 4
drwxrwxr-x. 2 acadgild acadgild 4096 May 6 17:49 export
[acadgild@localhost ~]$ ls -l hadoopdata/hive/export
total 4
-rw-r--r--. 1 acadgild acadgild 40 May 6 17:49 000000.0
[acadgild@localhost ~]$ cat hadoopdata/hive/export/000*
1990|7|23
1991|9|22
1993|2|16
1994|2|23
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```