APRIL 4, 2018

BIG DATA HADOOP & SPARK TRAINING

Assignment on Hive basics

RASHMI KRISHNA

Task 1

Created a database named 'custom' using the command “**Create database custom**” and use the created database using the command “**Use custom**”

Created a table named temperature_data inside custom having below fields:

1. date (mm-dd-yyyy) format

2. zip code

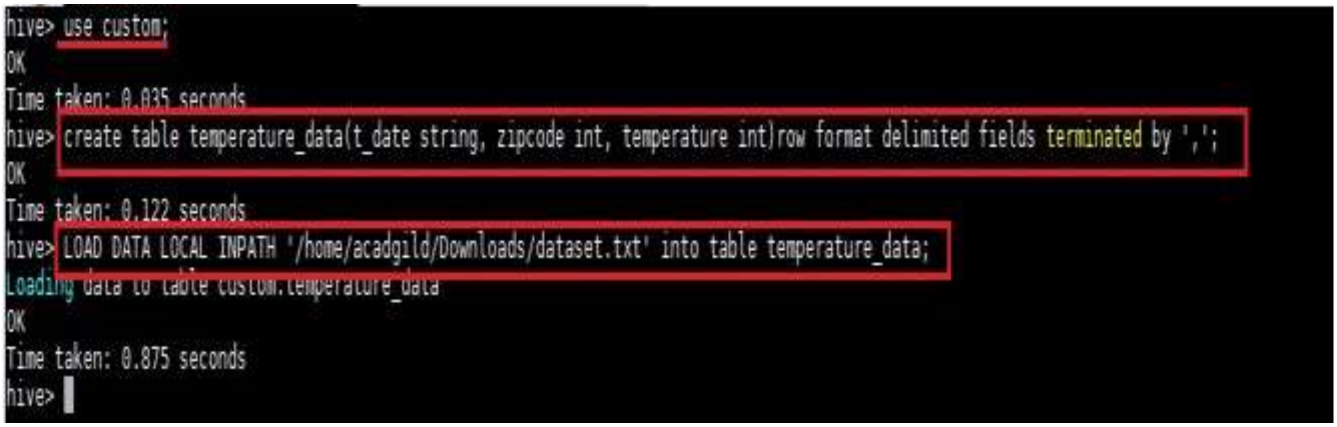
3. temperature

The table will be loaded from comma-delimited file.

create table temperature_data(t_date string, zipcode int, temperature int)row format delimited fields terminated by ',';

Load the dataset.txt (which is ',' delimited) in the table.

LOAD DATA LOCAL INPATH '/home/acadgild/Downloads/dataset.txt' into table temperature_data;



```
hive> use custom;
OK
Time taken: 0.035 seconds
hive> create table temperature_data(t_date string, zipcode int, temperature int)row format delimited fields terminated by ',';
OK
Time taken: 0.122 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/Downloads/dataset.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 0.875 seconds
hive>
```

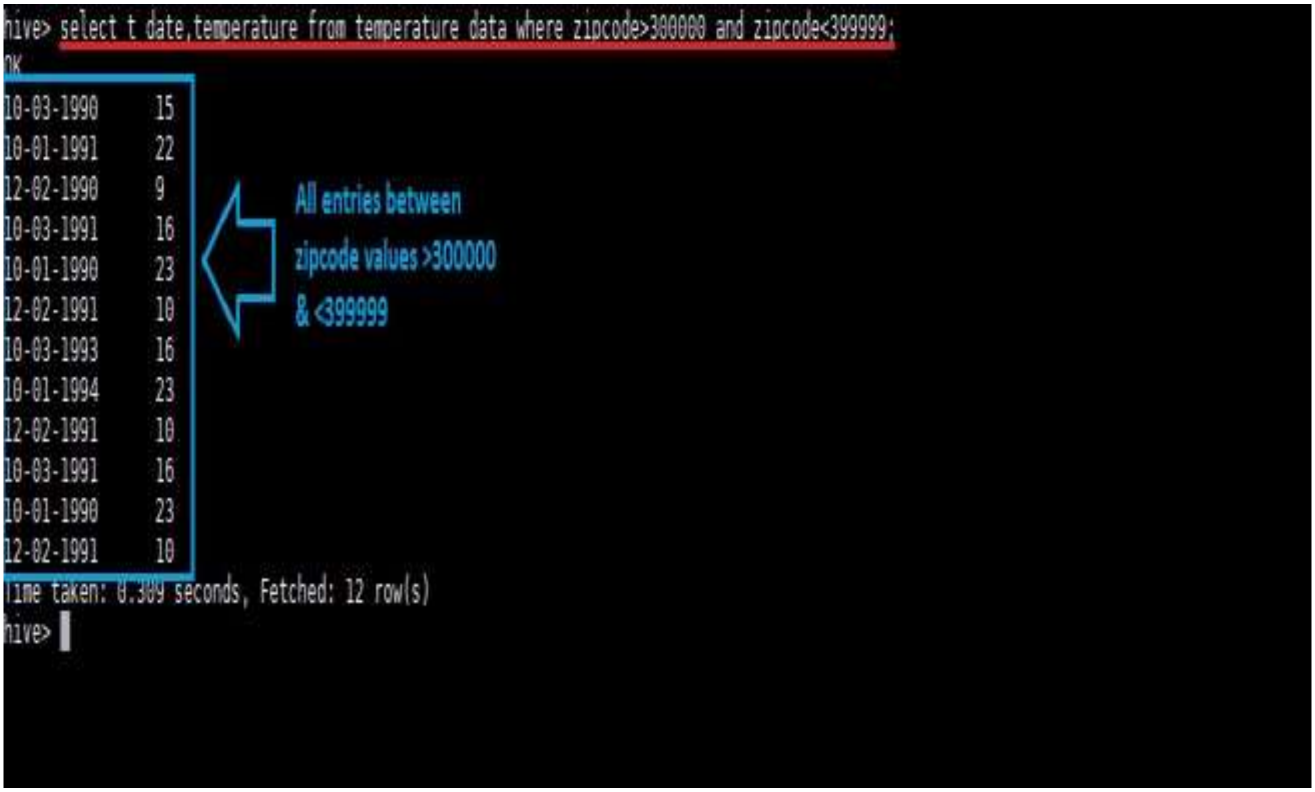
Task 2

● Fetch date and temperature from temperature_data where zip code is greater than

300000 and less than 399999.

To perform the above task use the query command :

select t_date,temperature from temperature_data where zipcode>300000 and zipcode<399999;



```
hive> select t_date,temperature from temperature_data where zipcode>300000 and zipcode<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: 0.309 seconds, Fetched: 12 row(s)
hive>
```

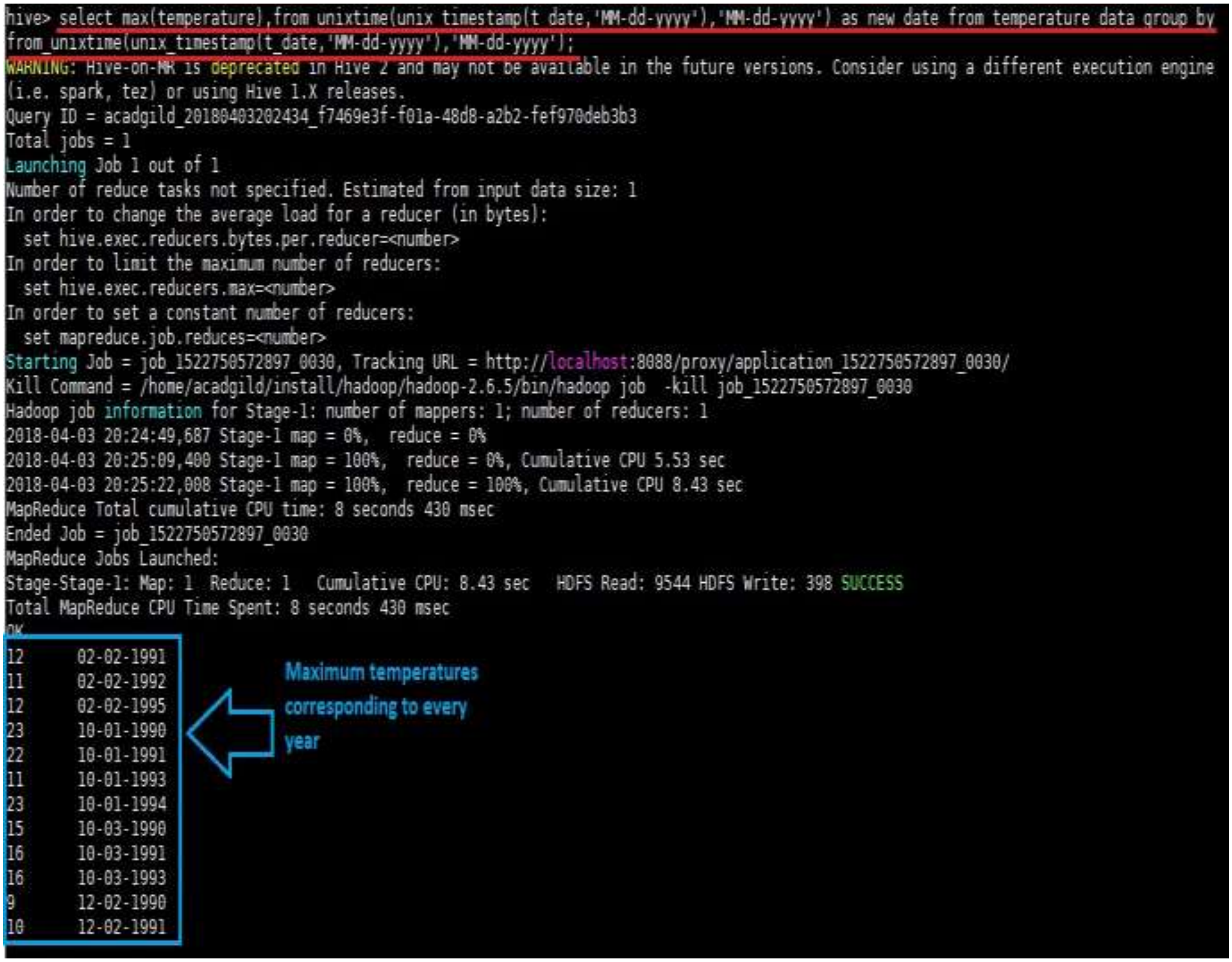
● Calculate maximum temperature corresponding to every year from temperature_data

table.

To perform the above task use the following command :

select max(temperature),from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'MM-dd-yyyy') as new_date from temperature_data group by from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'MM-dd-yyyy');

- **Max(temperature):** this will find the maximum temperature.
- **from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'MM-dd-yyyy'):** this to format date in the month-date-year format and display in the same format.
- This give the list of maximum temperature across all the years in the given data.



```
hive> select max(temperature),from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'MM-dd-yyyy') as new_date from temperature_data group by
from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'MM-dd-yyyy');
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_20180403202434_f7469e3f-f01a-48d8-a2b2-fef970deb3b3
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_1522750572897_0030, Tracking URL = http://localhost:8088/proxy/application_1522750572897_0030/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1522750572897_0030
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-04-03 20:24:49,687 Stage-1 map = 0%,  reduce = 0%
2018-04-03 20:25:09,400 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.53 sec
2018-04-03 20:25:22,008 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.43 sec
MapReduce Total cumulative CPU time: 8 seconds 430 msec
Ended Job = job_1522750572897_0030
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.43 sec   HDFS Read: 9544 HDFS Write: 398 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 430 msec
OK
12      02-02-1991
11      02-02-1992
12      02-02-1995
23      10-01-1990
22      10-01-1991
11      10-01-1993
23      10-01-1994
15      10-03-1990
16      10-03-1991
16      10-03-1993
9       12-02-1990
10      12-02-1991
```

Maximum temperatures corresponding to every year

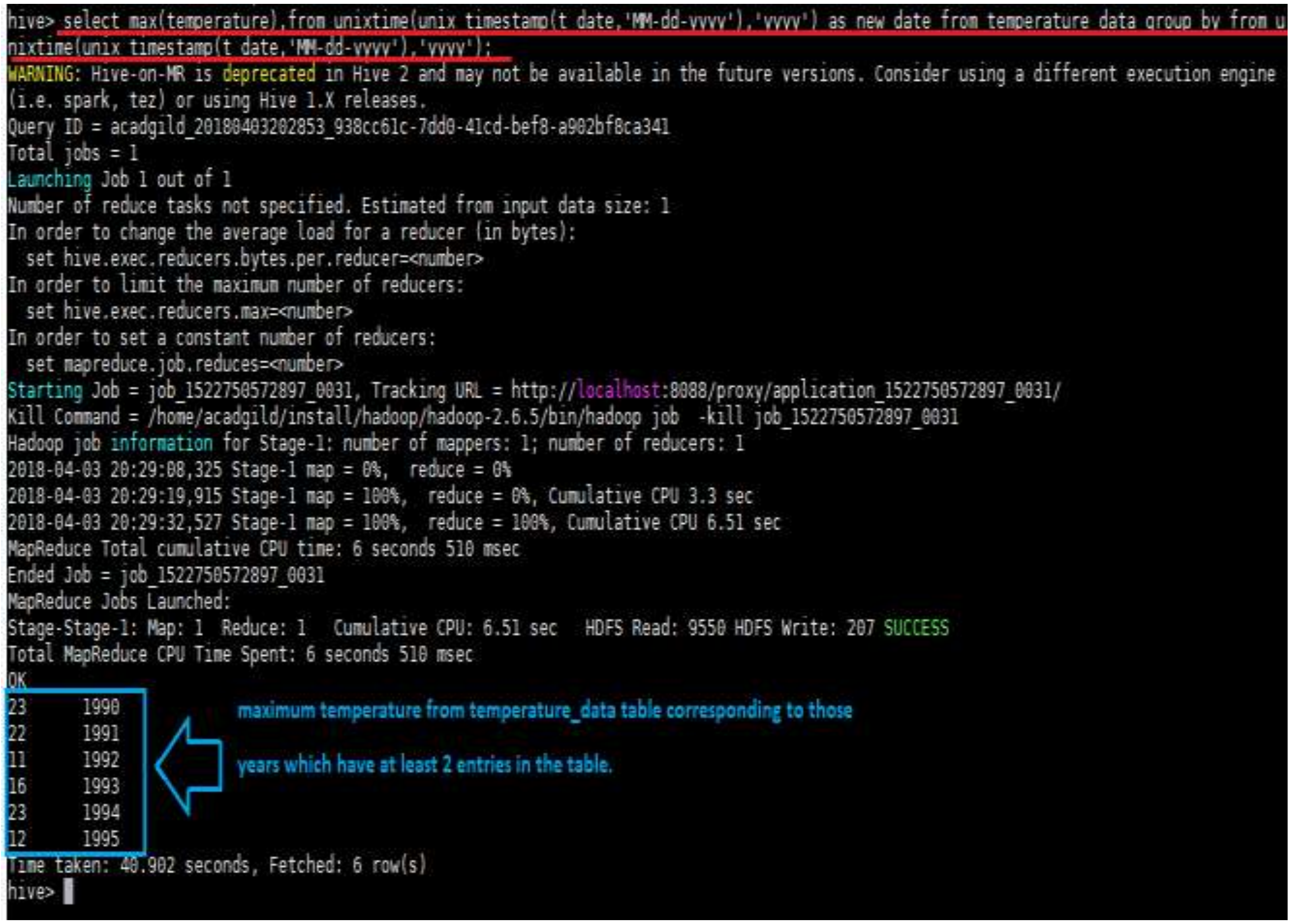
● Calculate maximum temperature from temperature_data table corresponding to those

years which have at least 2 entries in the table.

To perform the above task use the following command :

select max(temperature),from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy') as new_date from temperature_data group by from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy');

- **Max(temperature):** this will find the maximum temperature.
- **from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),’yyyy'):** this to format date in the month-date-year format and display only the year.
- This give the list of maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.



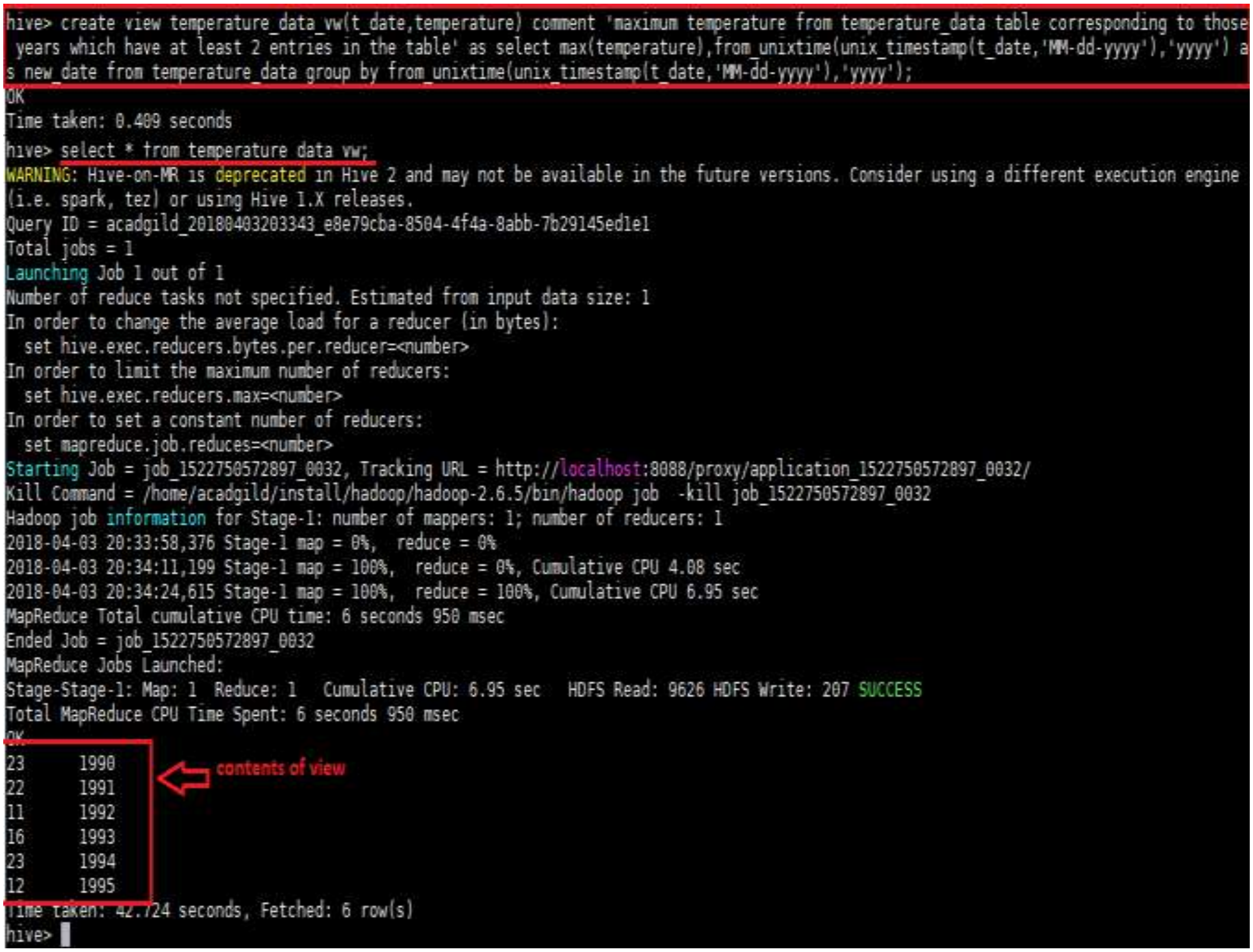
```
hive> select max(temperature),from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy') as new_date from temperature_data group by from_u
nixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy');
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine
(i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180403202853_938cc61c-7dd0-41cd-bef8-a902bf8ca341
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_1522750572897_0031, Tracking URL = http://localhost:8088/proxy/application_1522750572897_0031/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1522750572897_0031
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-04-03 20:29:08,325 Stage-1 map = 0%,  reduce = 0%
2018-04-03 20:29:19,915 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.3 sec
2018-04-03 20:29:32,527 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.51 sec
MapReduce Total cumulative CPU time: 6 seconds 510 msec
Ended Job = job_1522750572897_0031
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.51 sec   HDFS Read: 9550 HDFS Write: 207 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 510 msec
OK
23      1990
22      1991
11      1992
16      1993
23      1994
12      1995
Time taken: 40.902 seconds, Fetched: 6 row(s)
hive>
```

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

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

Created a view using the below command for the previously query:

create view temperature_data_vw(t_date,temperature) comment 'maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table' as select max(temperature),from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy') as new_date from temperature_data group by from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy');

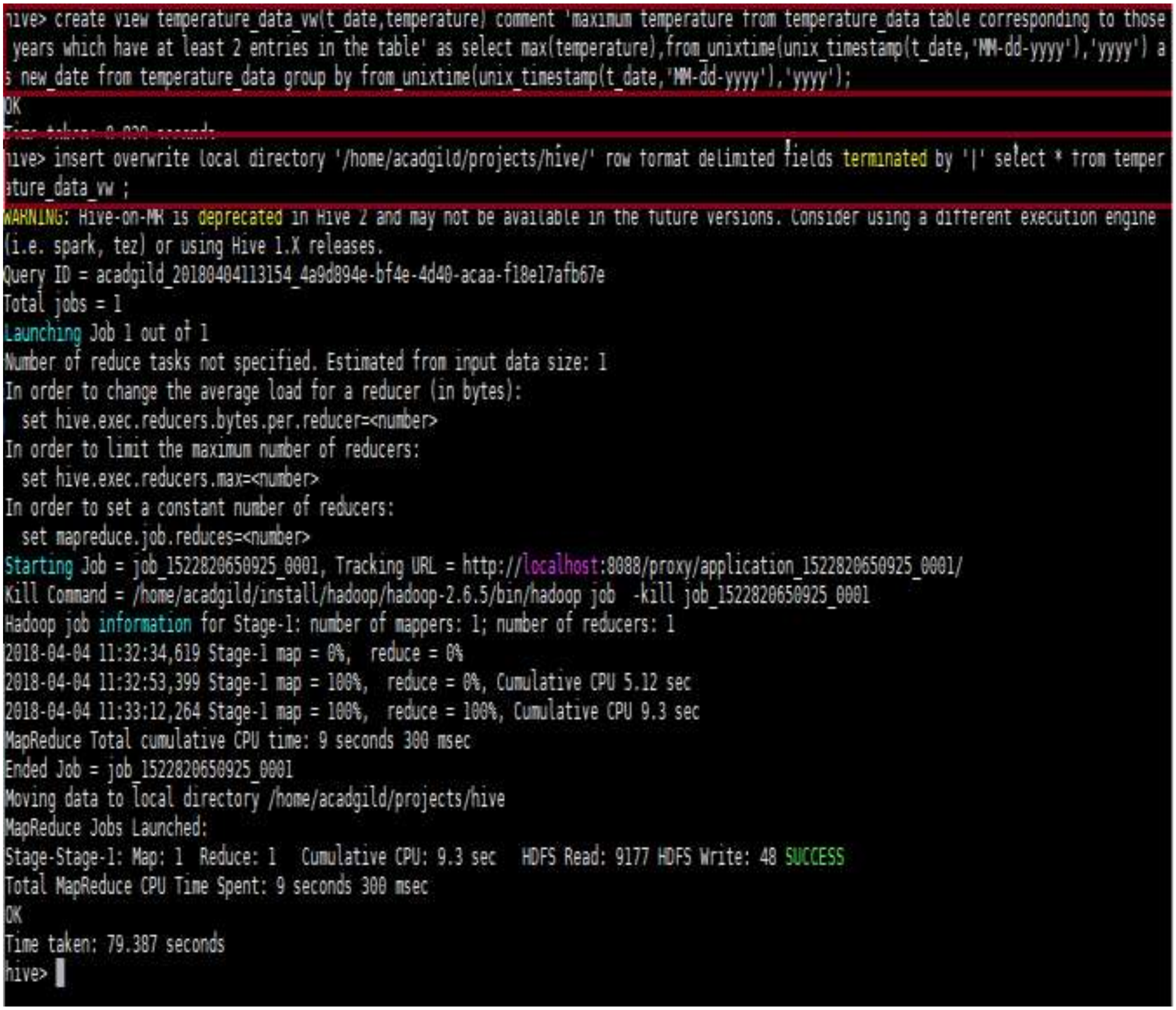


```
hive> create view temperature_data_vw(t_date,temperature) comment 'maximum temperature from temperature_data table corresponding to those
 years which have at least 2 entries in the table' as select max(temperature),from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy') a
s new_date from temperature_data group by from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy');
OK
Time taken: 0.409 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_20180403203343_e8e79cba-8504-4f4a-8abb-7b29145ed1e1
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_1522750572897_0032, Tracking URL = http://localhost:8088/proxy/application_1522750572897_0032/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1522750572897_0032
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-04-03 20:33:58,376 Stage-1 map = 0%,  reduce = 0%
2018-04-03 20:34:11,199 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.08 sec
2018-04-03 20:34:24,615 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.95 sec
MapReduce Total cumulative CPU time: 6 seconds 950 msec
Ended Job = job_1522750572897_0032
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.95 sec   HDFS Read: 9626 HDFS Write: 207 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 950 msec
OK
23      1990
22      1991
11      1992
16      1993
23      1994
12      1995
Time taken: 42.724 seconds, Fetched: 6 row(s)
hive>
```

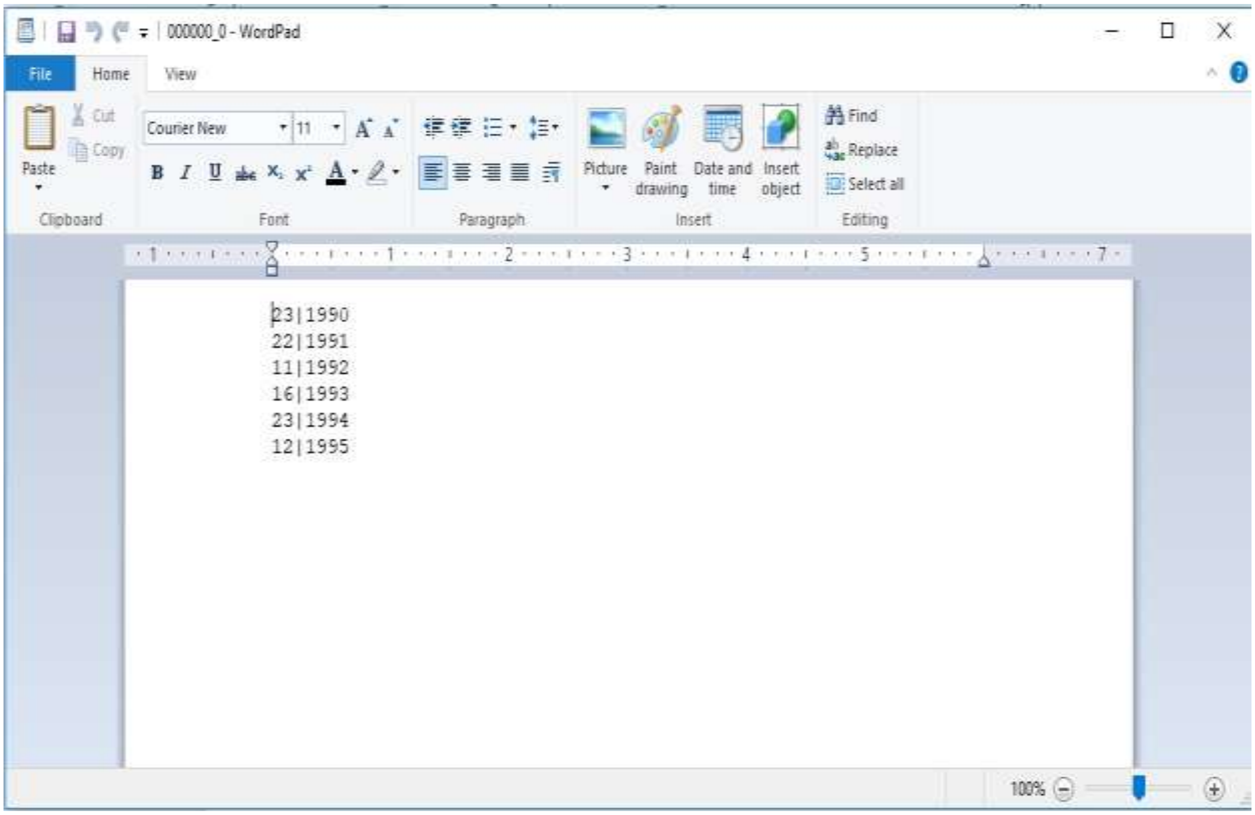
- Export contents from temperature_data_vw to a file in local file system, such that each

file is '|' delimited.

Exported the contents of the view to local file system using the command:
insert overwrite local directory '/home/acadgild/projects/hive/' row format delimited fields terminated by '|' select * from temperature_data_vw ;



```
hive> create view temperature_data_vw(t_date,temperature) comment 'maximum temperature from temperature_data table corresponding to those
 years which have at least 2 entries in the table' as select max(temperature),from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy') a
s new_date from temperature_data group by from_unixtime(unix_timestamp(t_date,'MM-dd-yyyy'),'yyyy');
OK
Time taken: 0.039 seconds
hive> insert overwrite local directory '/home/acadgild/projects/hive/' row format delimited fields terminated by '|' select * from temper
ature_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_20180404113154_4a9d894e-bf4e-4d40-acaa-f18e17afb67e
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_1522820650925_0001, Tracking URL = http://localhost:8088/proxy/application_1522820650925_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1522820650925_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-04-04 11:32:34,619 Stage-1 map = 0%,  reduce = 0%
2018-04-04 11:32:53,399 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.12 sec
2018-04-04 11:33:12,264 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.3 sec
MapReduce Total cumulative CPU time: 9 seconds 300 msec
Ended Job = job_1522820650925_0001
Moving data to local directory /home/acadgild/projects/hive
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.3 sec   HDFS Read: 9177 HDFS Write: 48 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 300 msec
OK
Time taken: 79.387 seconds
hive> 
```



```
23|1990
22|1991
11|1992
16|1993
23|1994
12|1995
```