Basic Hive Assignment

Task 1

Create a database named 'custom'.

Create a table named temperature_data inside custom having below fields:

1. date (mm-dd-yyyy) format

2. zip code

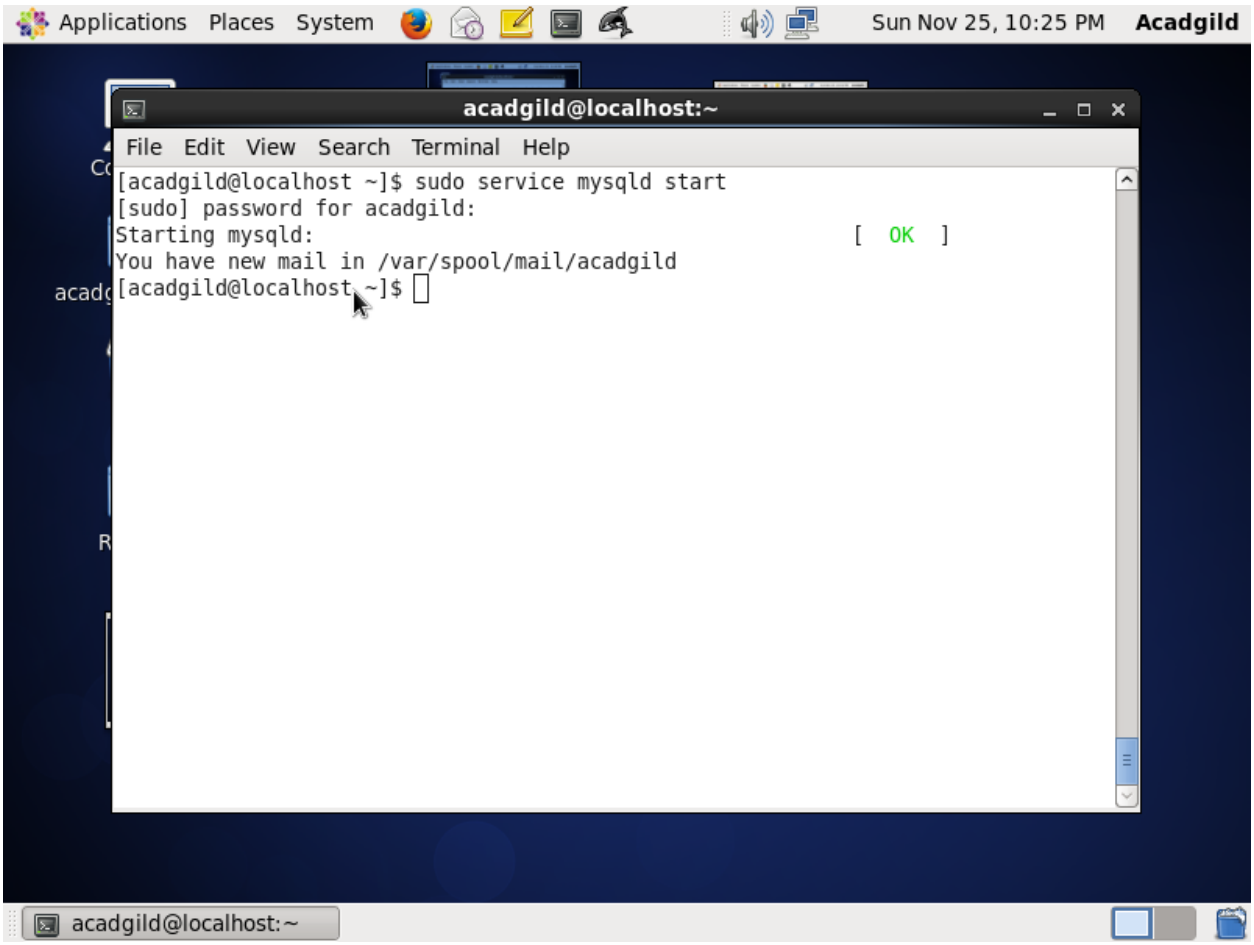
3. temperature

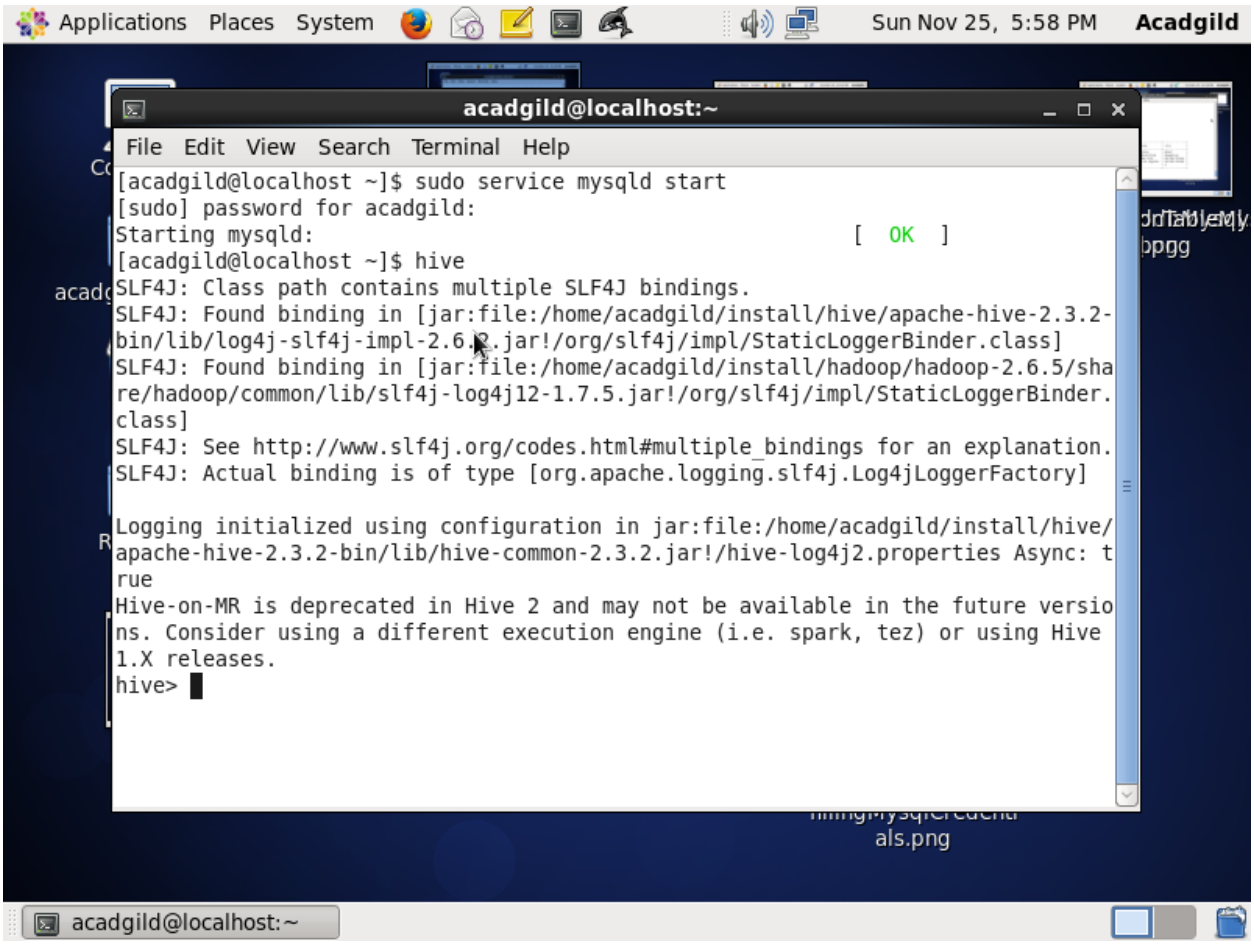
The table will be loaded from comma-delimited file.

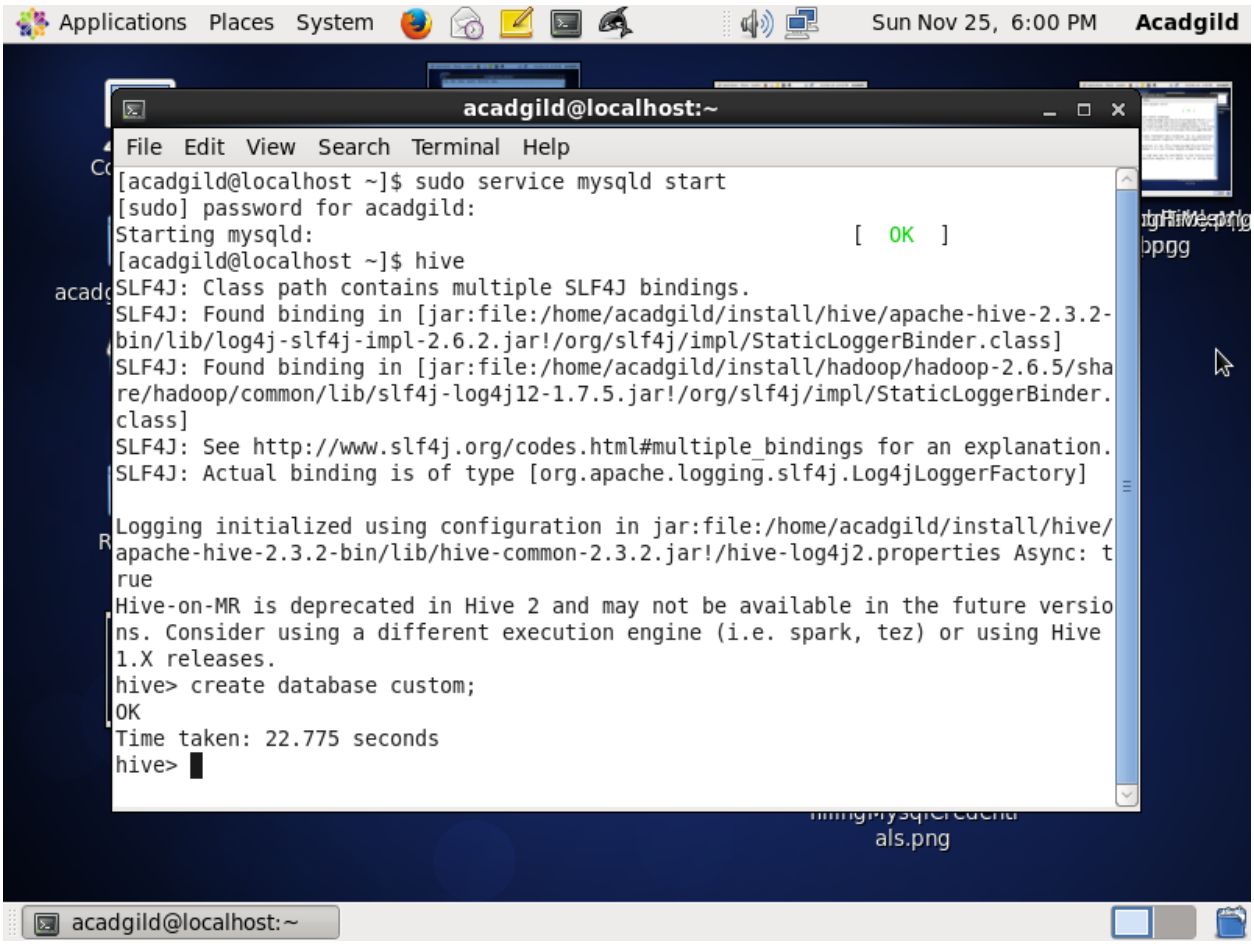
Load the dataset.txt (which is ',' delimited) in the table.

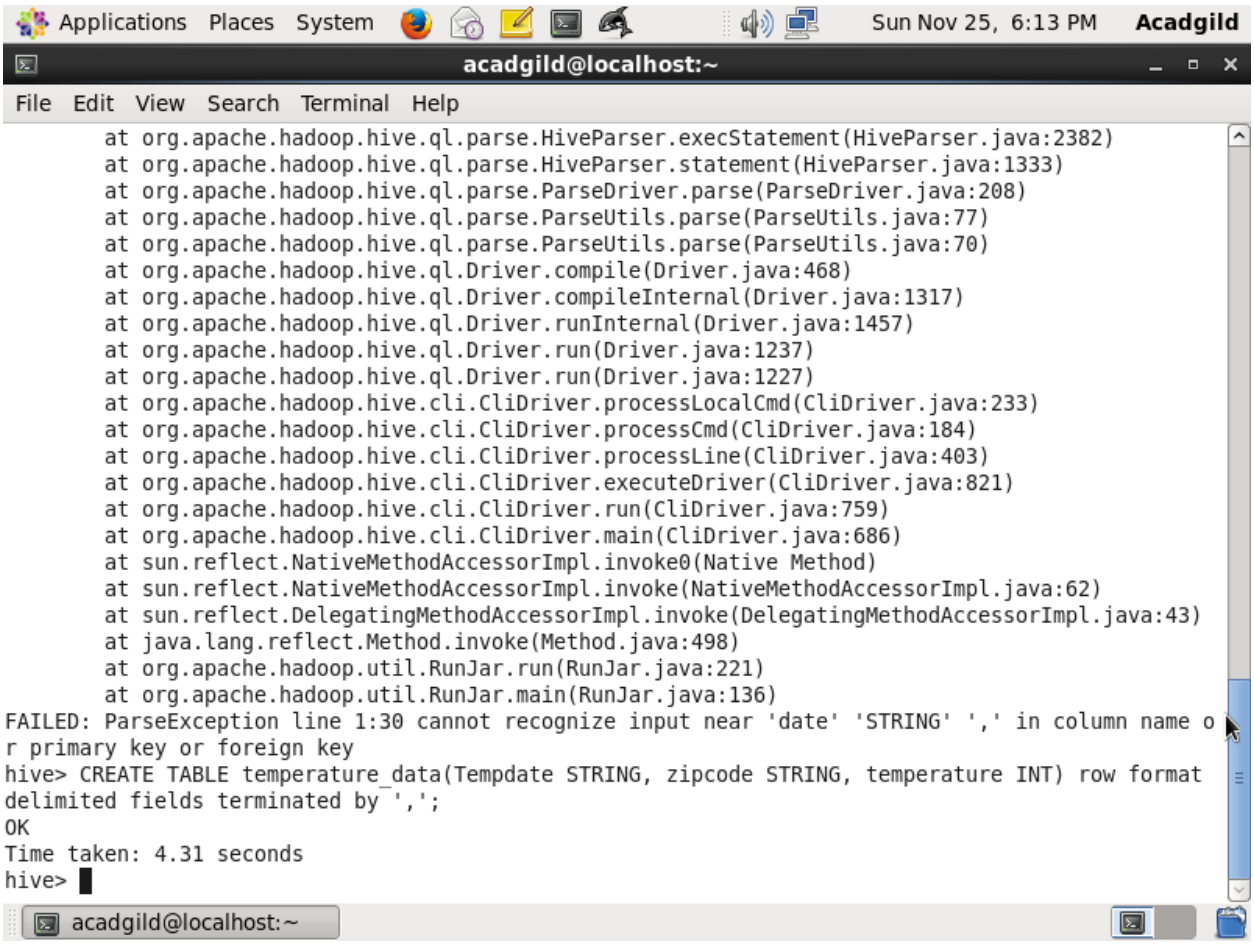
Solution:

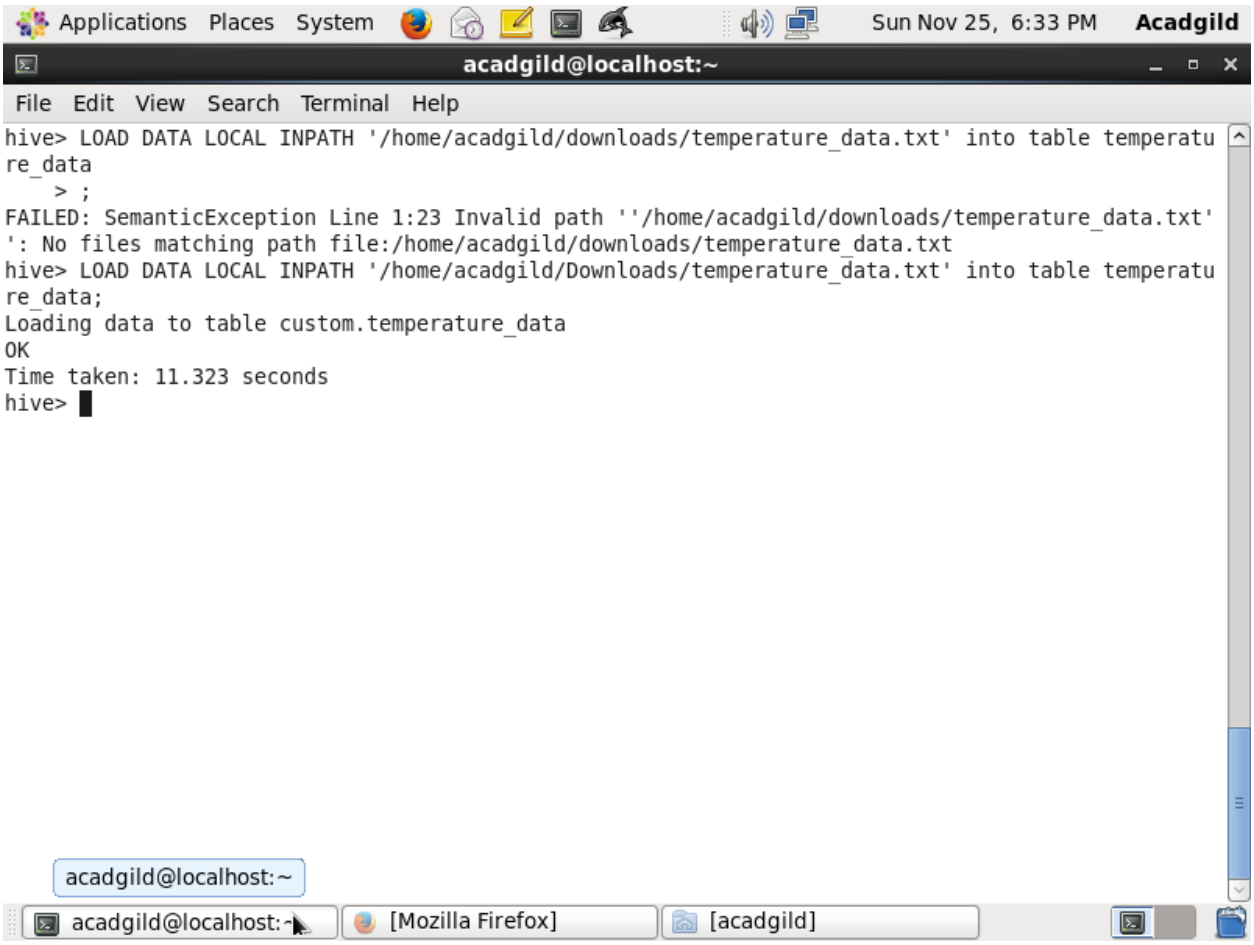
The screen shots of the task is below:




Applications Places System
Sun Nov 25, 5:58 PM
Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ sudo service mysqld start
[sudo] password for acadgild:
Starting mysqld: [OK]
[acadgild@localhost ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Logging initialized using configuration in jar:file:/home/acadgild/install/hive/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 versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
als.png
acadgild@localhost:~


Applications Places System
Sun Nov 25, 6:00 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost ~]$ sudo service mysqld start
[sudo] password for acadgild:
Starting mysqld: [OK]
[acadgild@localhost ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Logging initialized using configuration in jar:file:/home/acadgild/install/hive/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 versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> create database custom;
OK
Time taken: 22.775 seconds
hive>
als.png
acadgild@localhost:~


Applications Places System
Sun Nov 25, 6:13 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
at org.apache.hadoop.hive.ql.parse.HiveParser.execStatement(HiveParser.java:2382)
at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1333)
at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:208)
at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:77)
at org.apache.hadoop.hive.ql.parse.ParseUtils.parse(ParseUtils.java:70)
at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:468)
at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1317)
at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1457)
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1237)
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1227)
at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:233)
at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:184)
at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:403)
at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:30 cannot recognize input near 'date' 'STRING' ',' in column name o
r primary key or foreign key
hive> CREATE TABLE temperature_data(Tempdate STRING, zipcode STRING, temperature INT) row format
delimited fields terminated by ',';
OK
Time taken: 4.31 seconds
hive>
acadgild@localhost:~



Task 2

● Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.

● Calculate maximum temperature corresponding to every year from temperature_data table.

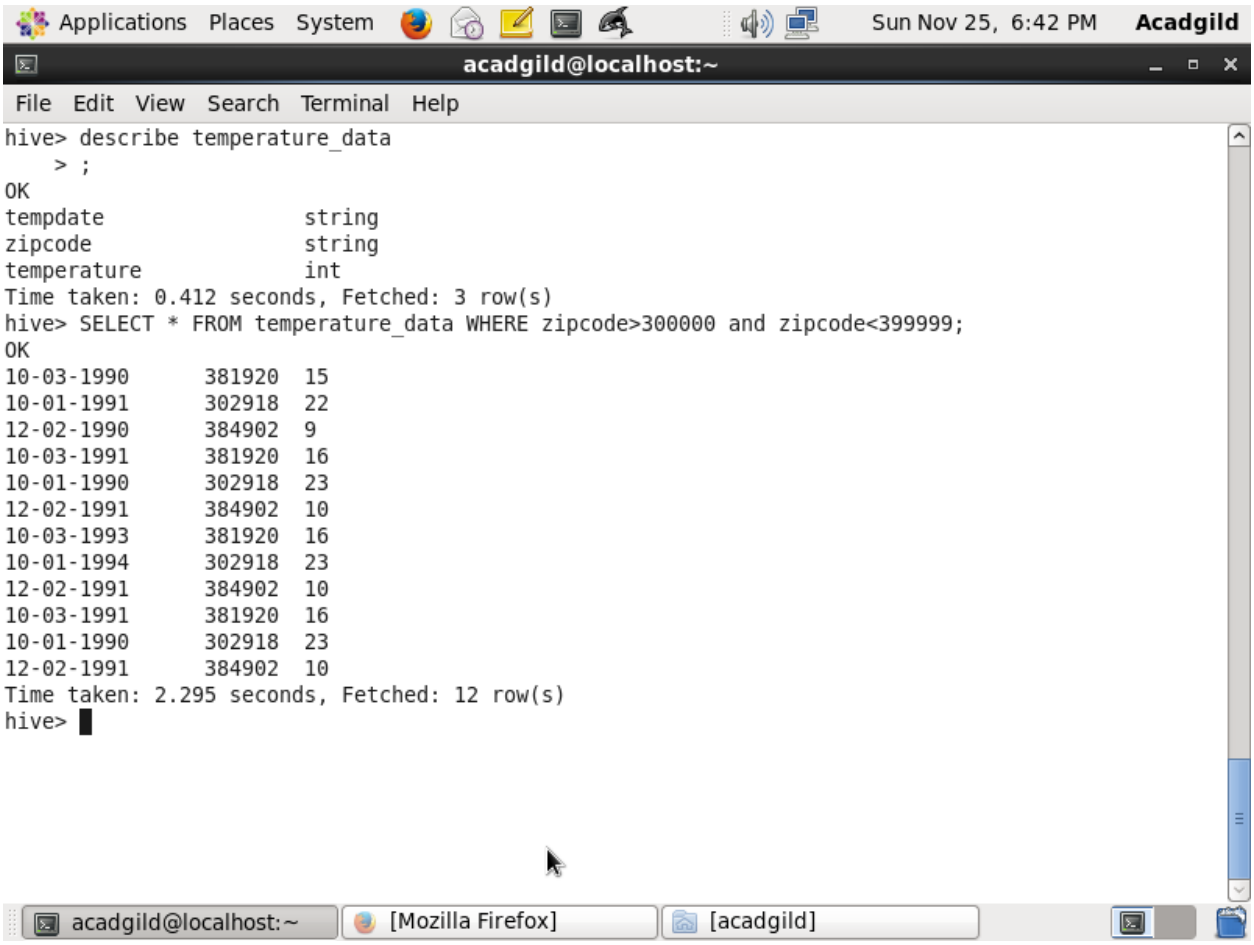
● Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.

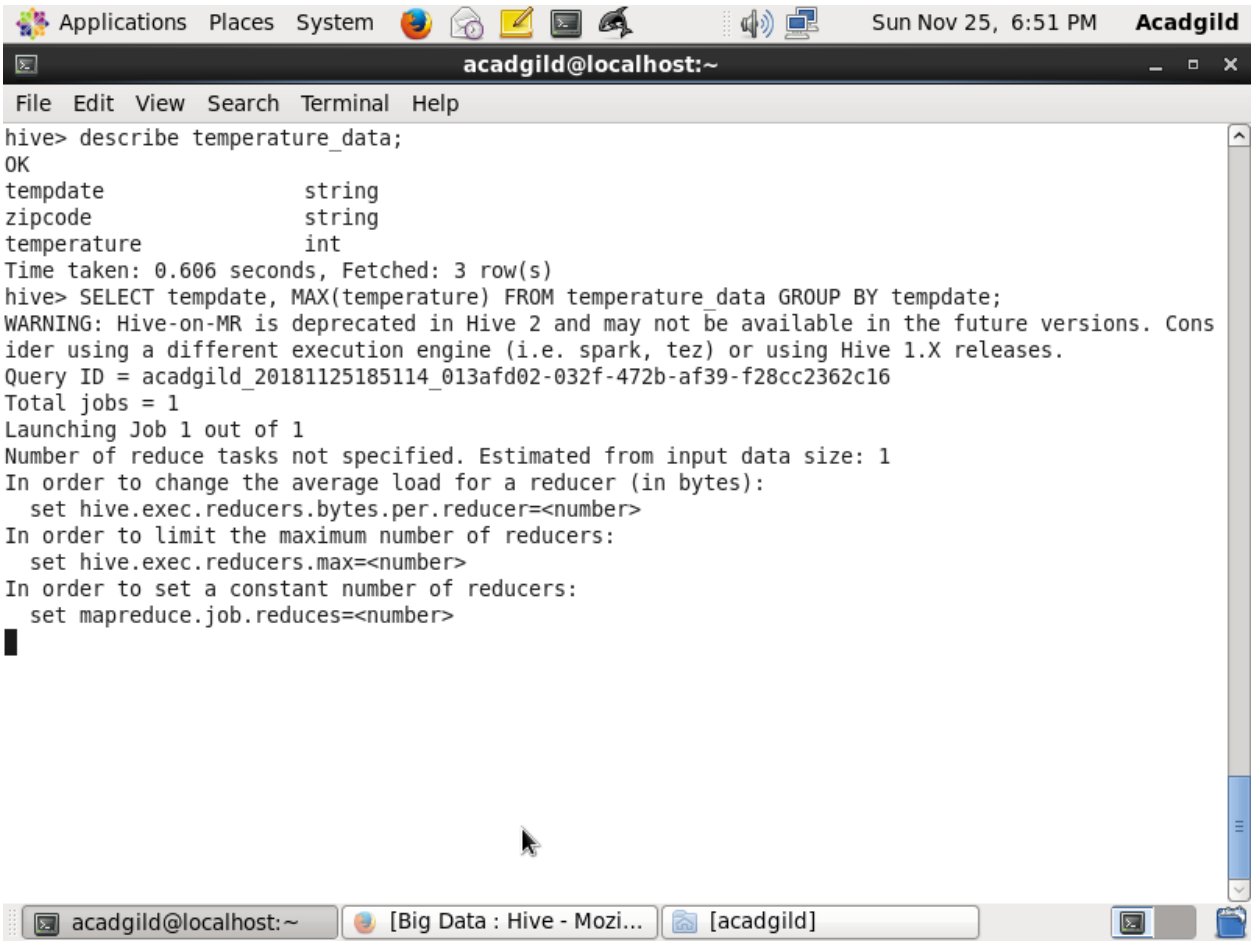
● Create a view on the top of last query, name it temperature_data_vw.

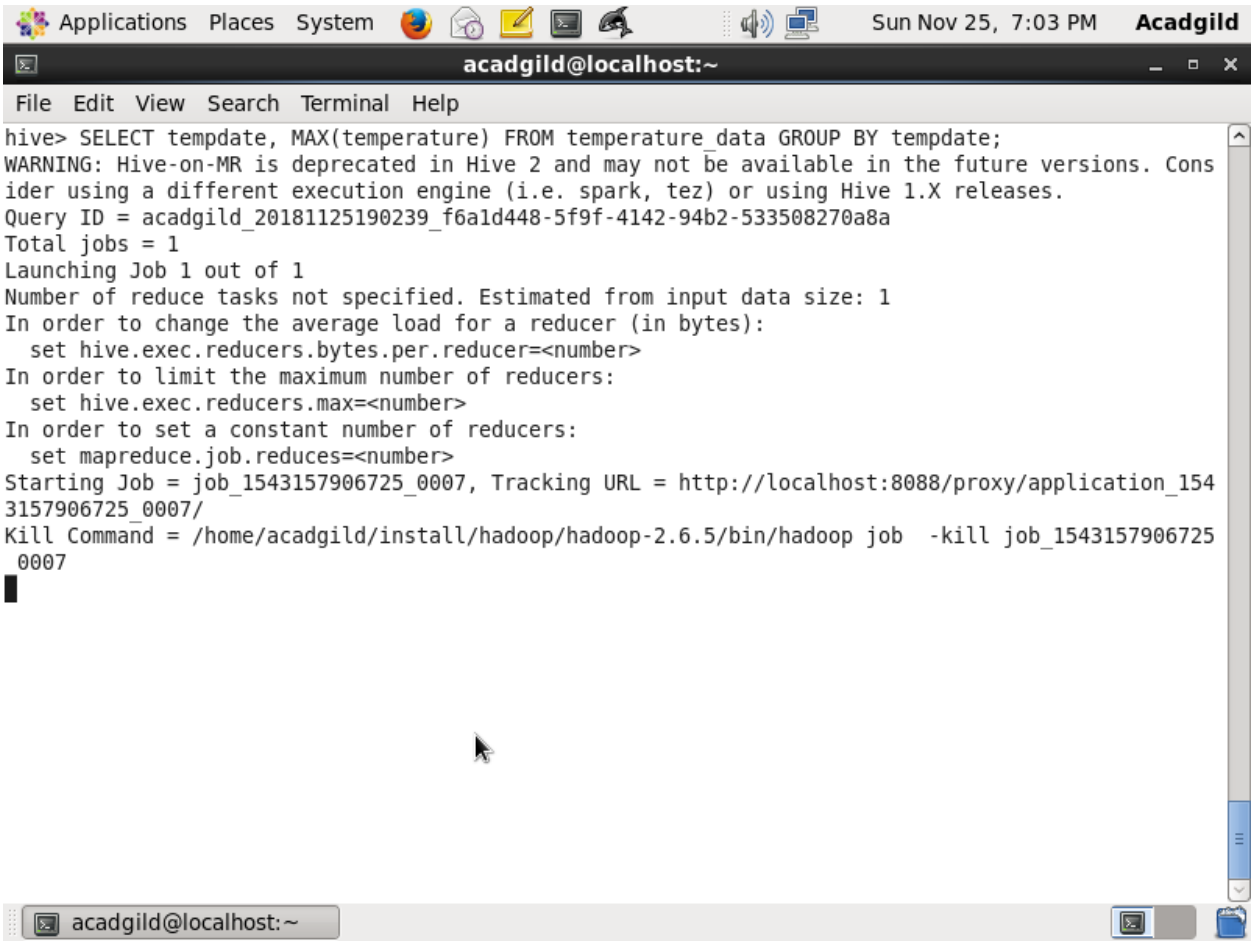
● Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

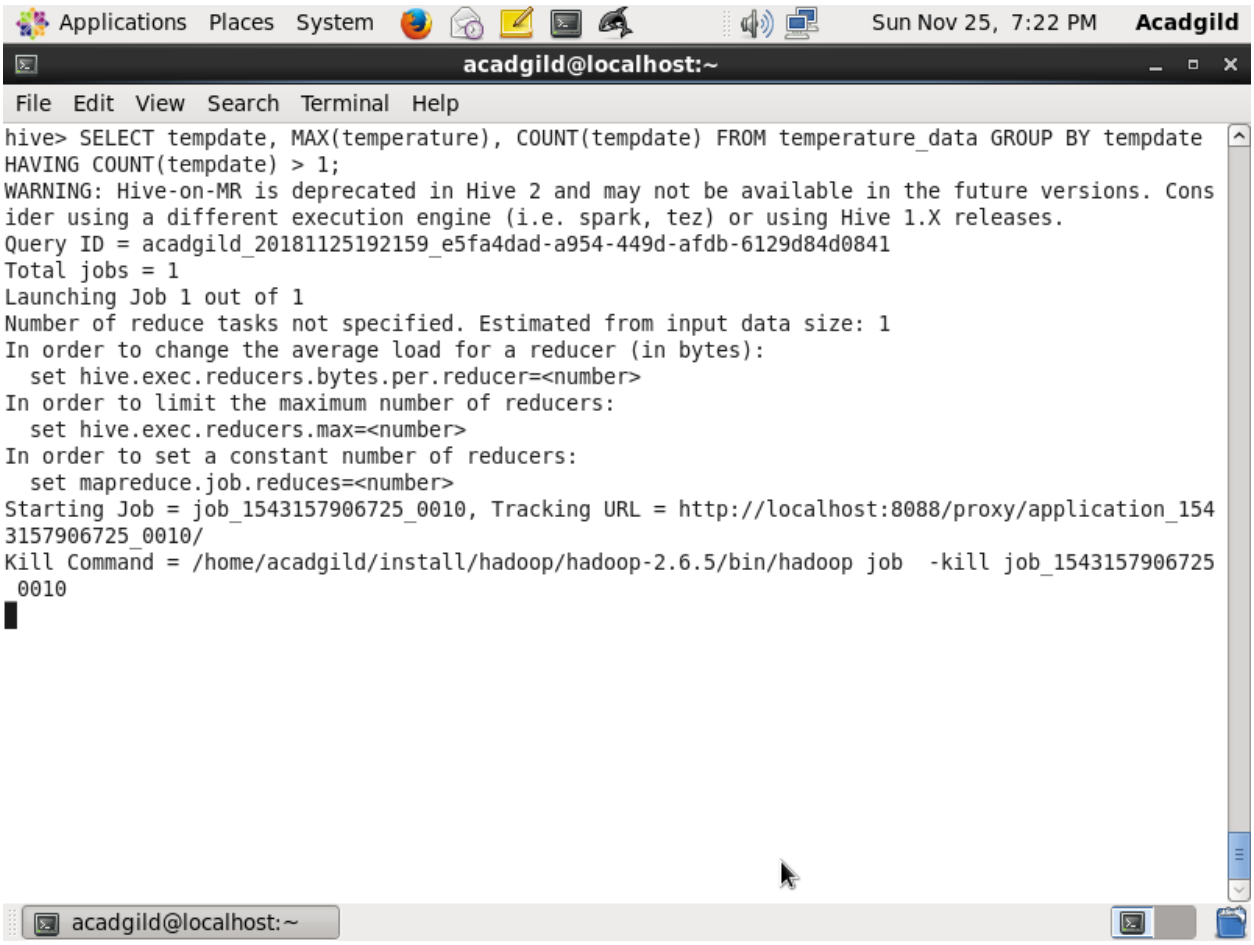
Solution:

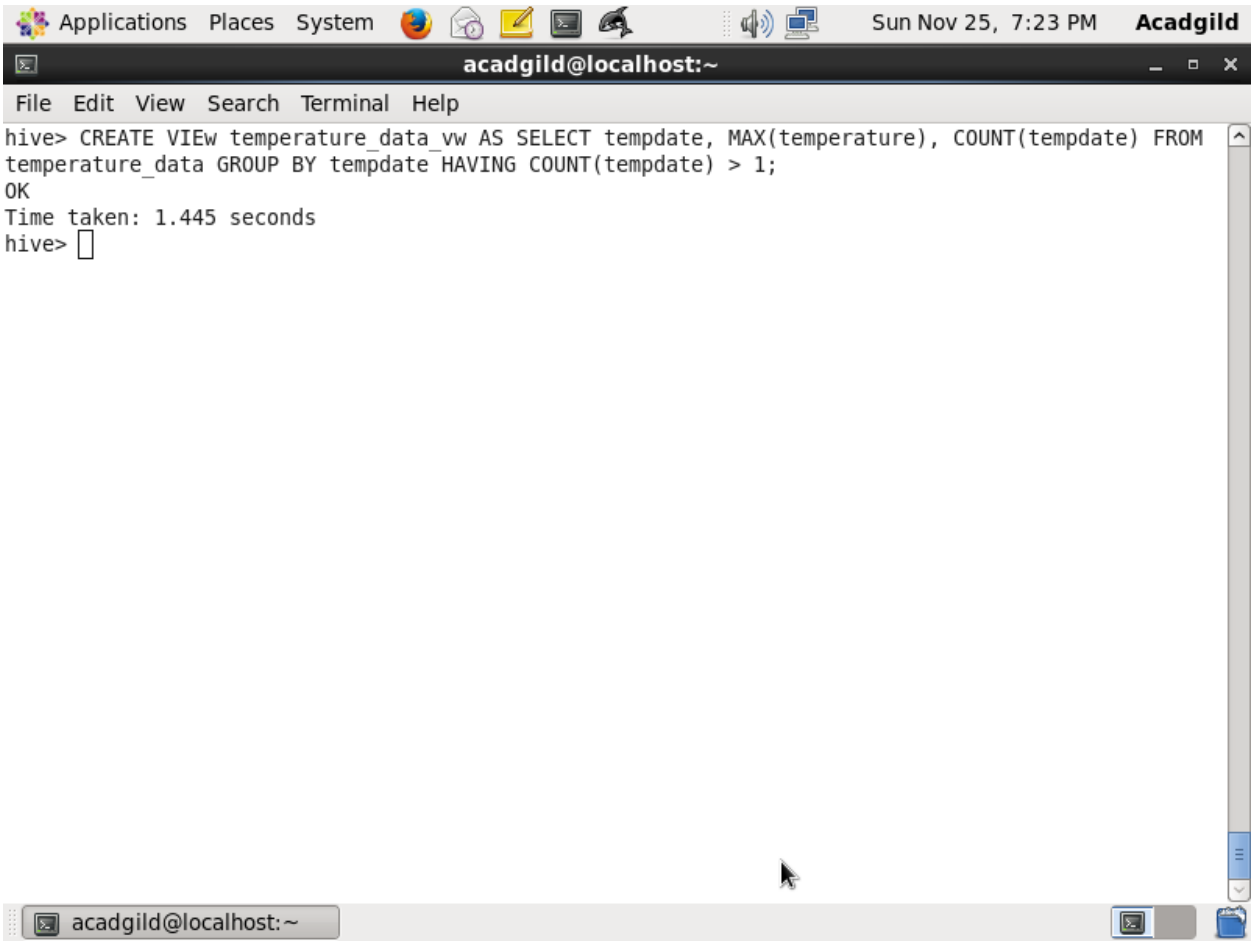
The screen shots of the task is below:

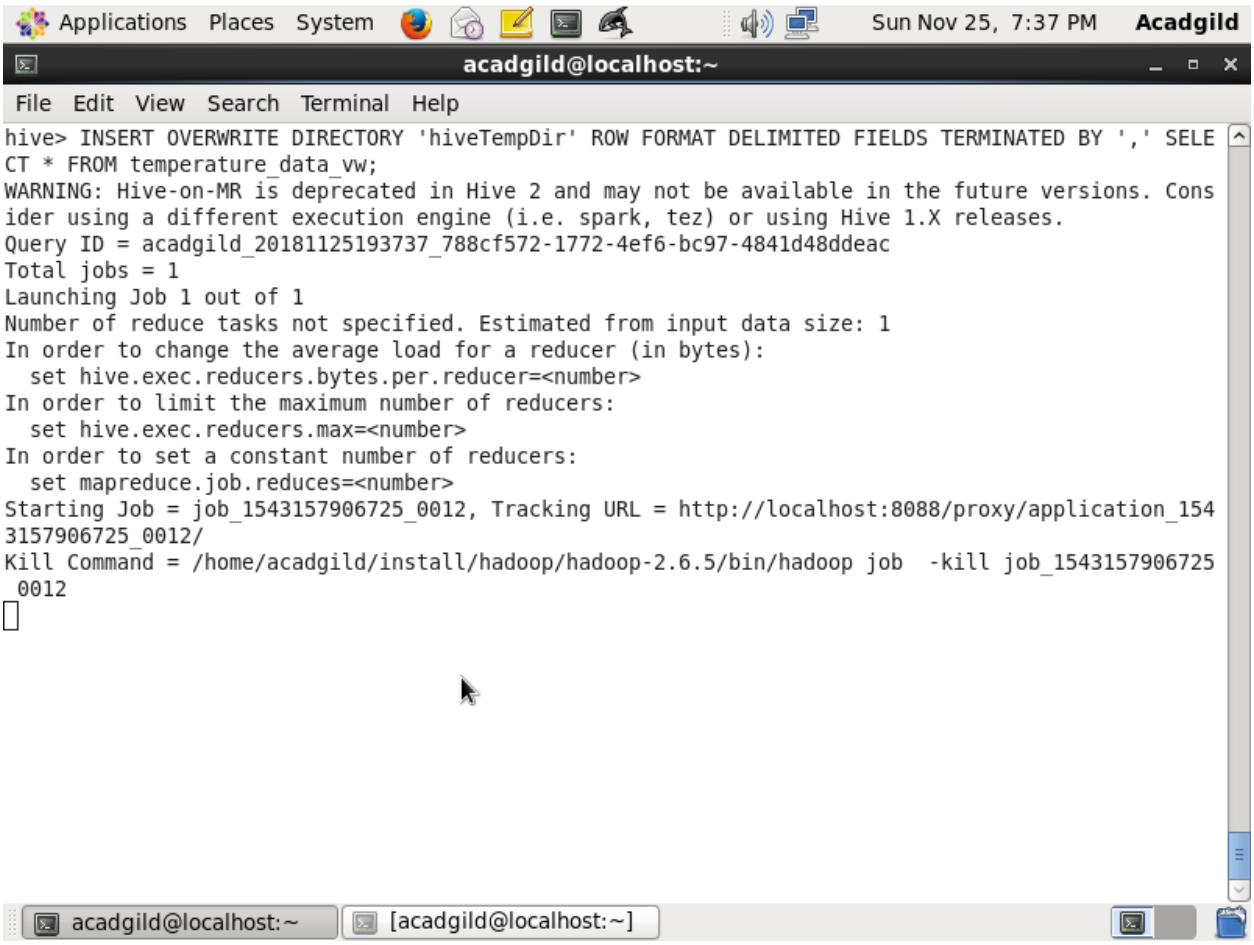

Applications Places System
Sun Nov 25, 6:42 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
hive> describe temperature_data
> ;
OK
tempdate string
zipcode string
temperature int
Time taken: 0.412 seconds, Fetched: 3 row(s)
hive> SELECT * FROM temperature_data WHERE zipcode>300000 and zipcode<399999;
OK
10-03-1990 381920 15
10-01-1991 302918 22
12-02-1990 384902 9
10-03-1991 381920 16
10-01-1990 302918 23
12-02-1991 384902 10
10-03-1993 381920 16
10-01-1994 302918 23
12-02-1991 384902 10
10-03-1991 381920 16
10-01-1990 302918 23
12-02-1991 384902 10
Time taken: 2.295 seconds, Fetched: 12 row(s)
hive>
acadgild@localhost:~
[Mozilla Firefox]
[acadgild]


Applications Places System
Sun Nov 25, 6:51 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
hive> describe temperature_data;
OK
tempdate string
zipcode string
temperature int
Time taken: 0.606 seconds, Fetched: 3 row(s)
hive> SELECT tempdate, MAX(temperature) FROM temperature_data GROUP BY tempdate;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Cons
ider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181125185114_013afd02-032f-472b-af39-f28cc2362c16
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>
acadgild@localhost:~
[Big Data : Hive - Mozi...
[acadgild]


Applications Places System
Sun Nov 25, 7:03 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
hive> SELECT tempdate, MAX(temperature) FROM temperature_data GROUP BY tempdate;
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_20181125190239_f6a1d448-5f9f-4142-94b2-533508270a8a
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_1543157906725_0007, Tracking URL = http://localhost:8088/proxy/application_1543157906725_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1543157906725_0007
acadgild@localhost:~


Applications Places System
Sun Nov 25, 7:22 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
hive> SELECT tempdate, MAX(temperature), COUNT(tempdate) FROM temperature_data GROUP BY tempdate
HAVING COUNT(tempdate) > 1;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Cons
ider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181125192159_e5fa4dad-a954-449d-afdb-6129d84d0841
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_1543157906725_0010, Tracking URL = http://localhost:8088/proxy/application_154
3157906725_0010/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1543157906725
0010
acadgild@localhost:~


Applications Places System
Sun Nov 25, 7:23 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
hive> CREATE VIEw temperature_data_vw AS SELECT tempdate, MAX(temperature), COUNT(tempdate) FROM temperature_data GROUP BY tempdate HAVING COUNT(tempdate) > 1;
OK
Time taken: 1.445 seconds
hive>
acadgild@localhost:~


Applications Places System
Sun Nov 25, 7:37 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
hive> INSERT OVERWRITE DIRECTORY 'hiveTempDir' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELE
CT * FROM temperature_data_vw;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Cons
ider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181125193737_788cf572-1772-4ef6-bc97-4841d48ddeac
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_1543157906725_0012, Tracking URL = http://localhost:8088/proxy/application_154
3157906725_0012/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1543157906725
_0012
acadgild@localhost:~
[acadgild@localhost:~]