Assignment 8:

Submitted by Jyoti Jha
DOS: 8th Aug'18

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 with screenshot:

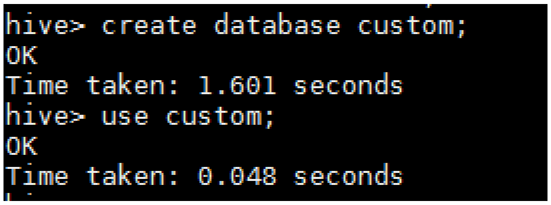
Create database custom;

Use custom;

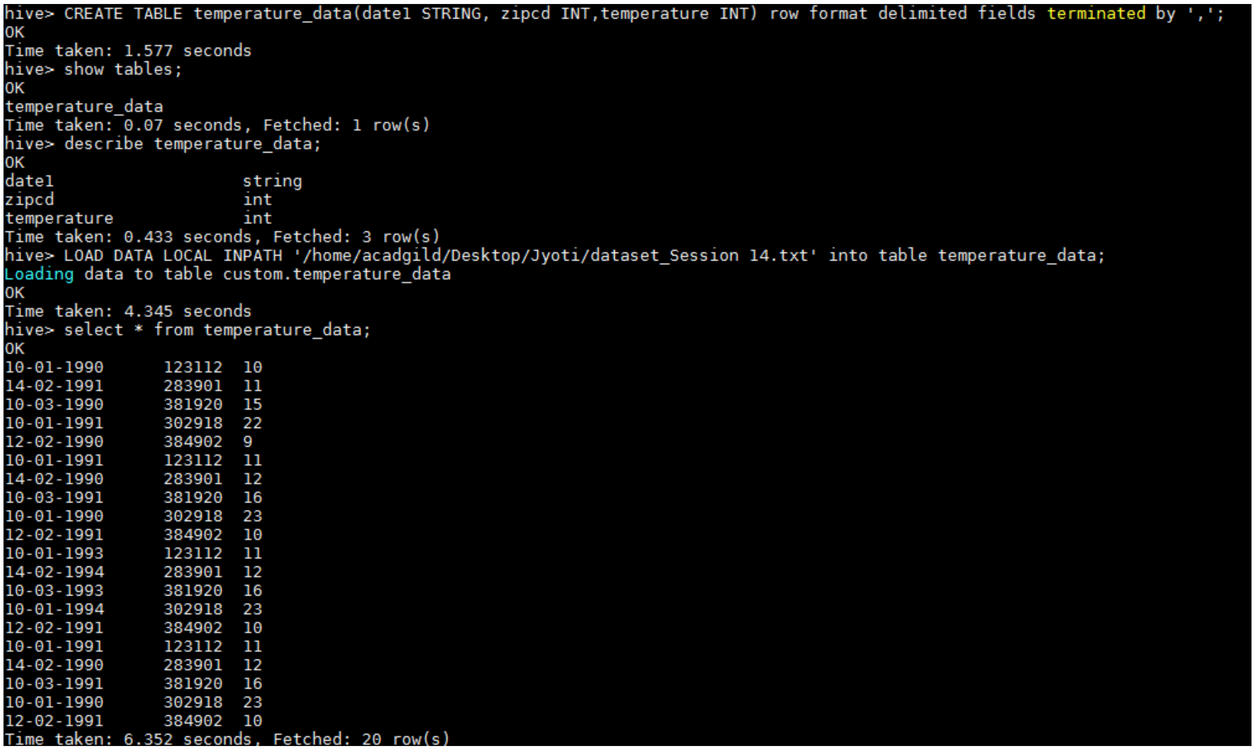
CREATE TABLE temperature_data(date1 STRING, zipcd INT,temperature INT) row format delimited fields terminated by ',';

LOAD DATA LOCAL INPATH '/home/acadgild/Desktop/Jyoti/dataset_Session 14.txt' into table temperature_data;

select * from temperature_data;



```
hive> create database custom;
OK
Time taken: 1.601 seconds
hive> use custom;
OK
Time taken: 0.048 seconds
```



```
hive> CREATE TABLE temperature_data(date1 STRING, zipcd INT,temperature INT) row format delimited fields terminated by ',';
OK
Time taken: 1.577 seconds
hive> show tables;
OK
temperature_data
Time taken: 0.07 seconds, Fetched: 1 row(s)
hive> describe temperature_data;
OK
date1                   string
zipcd                   int
temperature             int
Time taken: 0.433 seconds, Fetched: 3 row(s)
hive> LOAD DATA LOCAL INPATH '/home/acadgild/Desktop/Jyoti/dataset_Session 14.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 4.345 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: 6.352 seconds, Fetched: 20 row(s)
```

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.

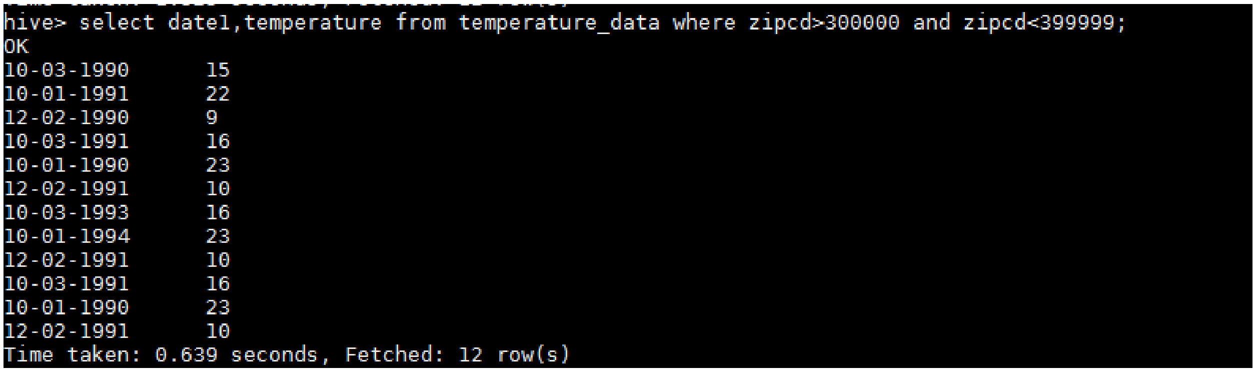
● Create a view on the top of last query, name it temperature_data_vw.

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

Solution with O/P and Screenshot:

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

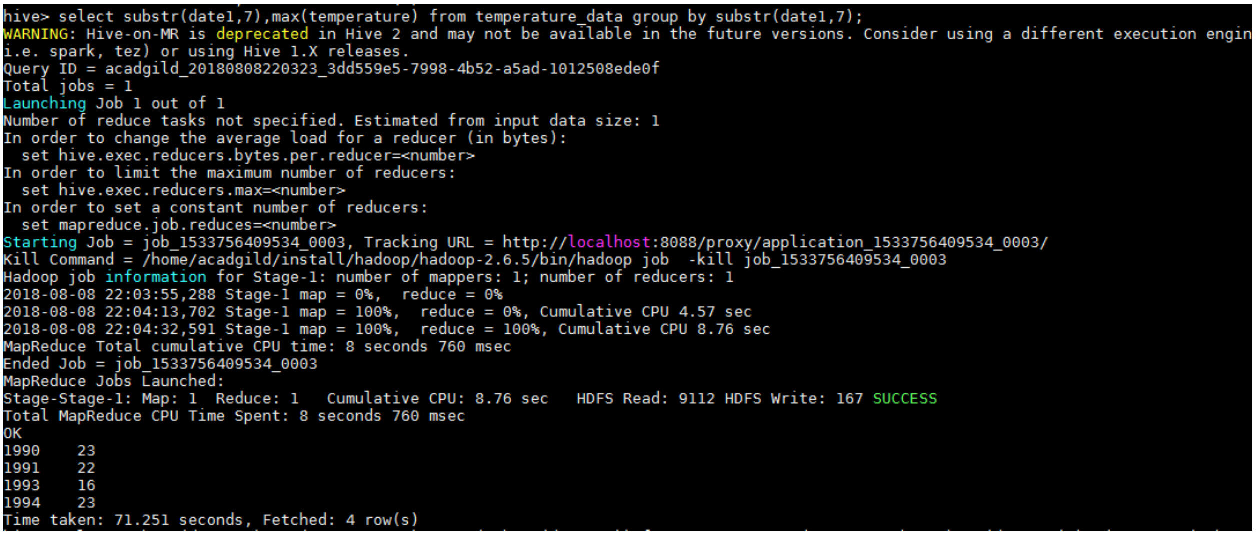
select date1,temperature from temperature_data where zipcd>300000 and zipcd<399999;



```
hive> select date1,temperature from temperature_data where zipcd>300000 and zipcd<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: 0.639 seconds, Fetched: 12 row(s)
```

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

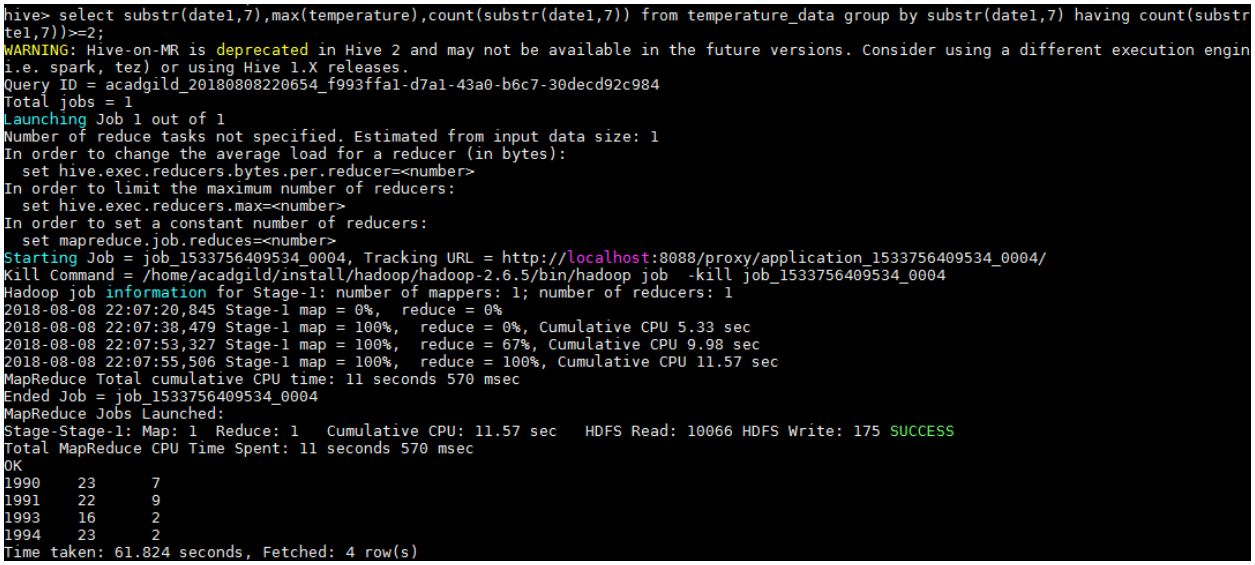
select substr(date1,7),max(temperature) from temperature_data group by substr(date1,7);



```
hive> select substr(date1,7),max(temperature) from temperature_data group by substr(date1,7);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engin
i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180808220323_3dd559e5-7998-4b52-a5ad-1012508ede0f
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_1533756409534_0003, Tracking URL = http://localhost:8088/proxy/application_1533756409534_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533756409534_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-08 22:03:55,288 Stage-1 map = 0%,  reduce = 0%
2018-08-08 22:04:13,702 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.57 sec
2018-08-08 22:04:32,591 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.76 sec
MapReduce Total cumulative CPU time: 8 seconds 760 msec
Ended Job = job_1533756409534_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.76 sec   HDFS Read: 9112 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 760 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 71.251 seconds, Fetched: 4 row(s)
```

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

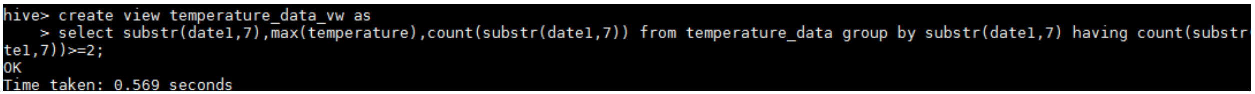
select substr(date1,7),max(temperature),count(substr(date1,7)) from temperature_data group by substr(date1,7) having count(substr(date1,7))>=2;



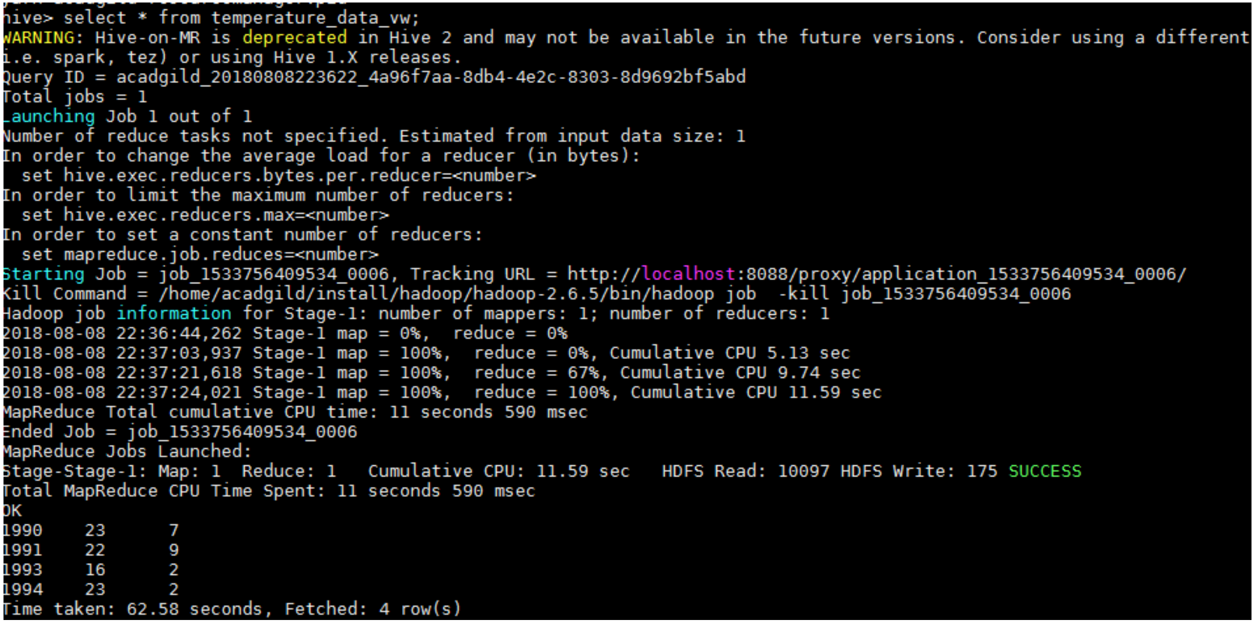
```
hive> select substr(date1,7),max(temperature),count(substr(date1,7)) from temperature_data group by substr(date1,7) having count(substr
te1,7))>=2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engin
i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180808220654_f993ffa1-d7a1-43a0-b6c7-30decd92c984
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_1533756409534_0004, Tracking URL = http://localhost:8088/proxy/application_1533756409534_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533756409534_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-08 22:07:20,845 Stage-1 map = 0%,  reduce = 0%
2018-08-08 22:07:38,479 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.33 sec
2018-08-08 22:07:53,327 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 9.98 sec
2018-08-08 22:07:55,506 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.57 sec
MapReduce Total cumulative CPU time: 11 seconds 570 msec
Ended Job = job_1533756409534_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.57 sec   HDFS Read: 10066 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 570 msec
OK
1990    23      7
1991    22      9
1993    16      2
1994    23      2
Time taken: 61.824 seconds, Fetched: 4 row(s)
```

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

create view temperature_data_vw as
select substr(date1,7),max(temperature),count(substr(date1,7)) from temperature_data group by substr(date1,7) having count(substr(date1,7))>=2;



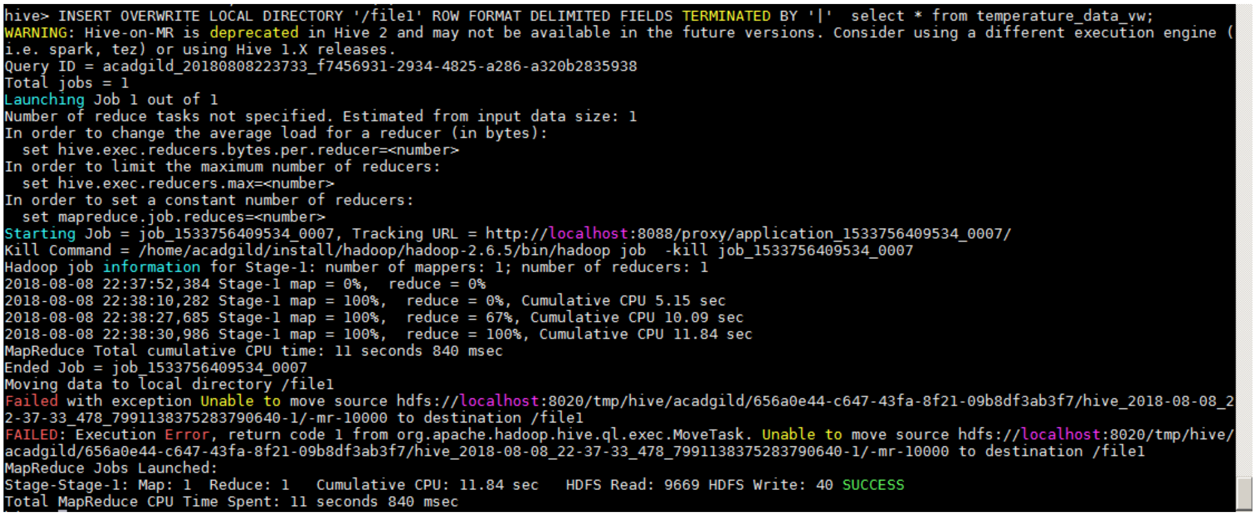
```
hive> create view temperature_data_vw as
    > select substr(date1,7),max(temperature),count(substr(date1,7)) from temperature_data group by substr(date1,7) having count(substr
te1,7))>=2;
OK
Time taken: 0.569 seconds
```



```
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
i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180808223622_4a96f7aa-8db4-4e2c-8303-8d9692bf5abd
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_1533756409534_0006, Tracking URL = http://localhost:8088/proxy/application_1533756409534_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533756409534_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-08 22:36:44,262 Stage-1 map = 0%,  reduce = 0%
2018-08-08 22:37:03,937 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.13 sec
2018-08-08 22:37:21,618 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 9.74 sec
2018-08-08 22:37:24,021 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.59 sec
MapReduce Total cumulative CPU time: 11 seconds 590 msec
Ended Job = job_1533756409534_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.59 sec   HDFS Read: 10097 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 590 msec
OK
1990    23      7
1991    22      9
1993    16      2
1994    23      2
Time taken: 62.58 seconds, Fetched: 4 row(s)
```

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

INSERT OVERWRITE LOCAL DIRECTORY '/file1' ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' select * from temperature_data_vw;



```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/file1' 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_20180808223733_f7456931-2934-4825-a286-a320b2835938
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_1533756409534_0007, Tracking URL = http://localhost:8088/proxy/application_1533756409534_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533756409534_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-08 22:37:52,384 Stage-1 map = 0%,  reduce = 0%
2018-08-08 22:38:10,282 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.15 sec
2018-08-08 22:38:27,685 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 10.09 sec
2018-08-08 22:38:30,986 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.84 sec
MapReduce Total cumulative CPU time: 11 seconds 840 msec
Ended Job = job_1533756409534_0007
Moving data to local directory /file1
Failed with exception Unable to move source hdfs://localhost:8020/tmp/hive/acadgild/656a0e44-c647-43fa-8f21-09b8df3ab3f7/hive_2018-08-08_2
2-37-33_478_7991138375283790640-1/-mr-10000 to destination /file1
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.MoveTask. Unable to move source hdfs://localhost:8020/tmp/hive/
acadgild/656a0e44-c647-43fa-8f21-09b8df3ab3f7/hive_2018-08-08_22-37-33_478_7991138375283790640-1/-mr-10000 to destination /file1
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.84 sec   HDFS Read: 9669 HDFS Write: 40 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 840 msec
```