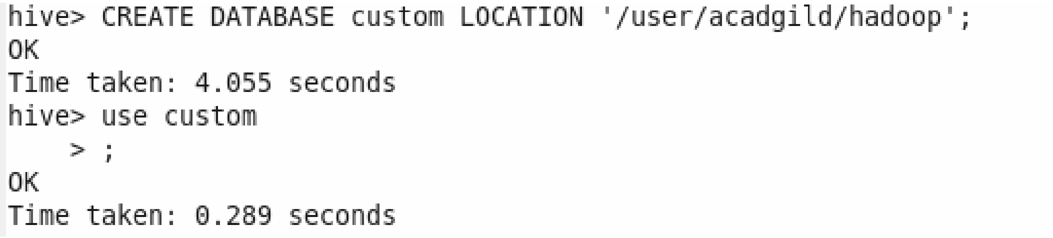
Task 1

Create a database named 'custom'.

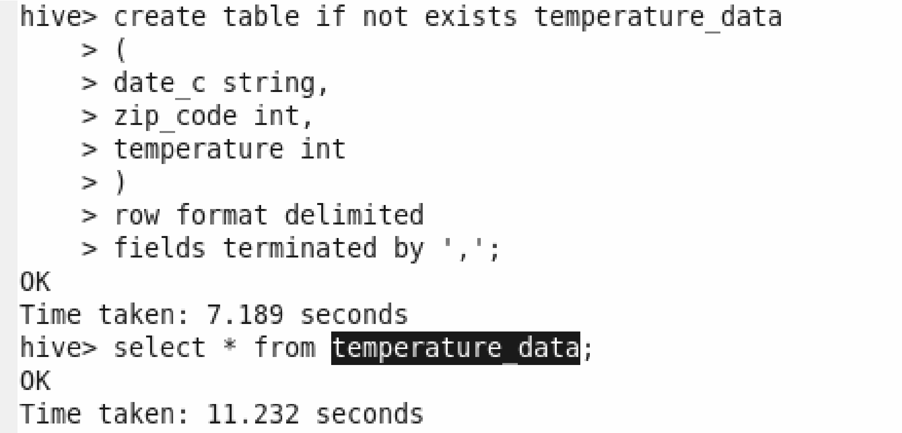


```
hive> CREATE DATABASE custom LOCATION '/user/acadgild/hadoop';
OK
Time taken: 4.055 seconds
hive> use custom
    > ;
OK
Time taken: 0.289 seconds
```

Create a table named temperature_data inside custom having below fields:

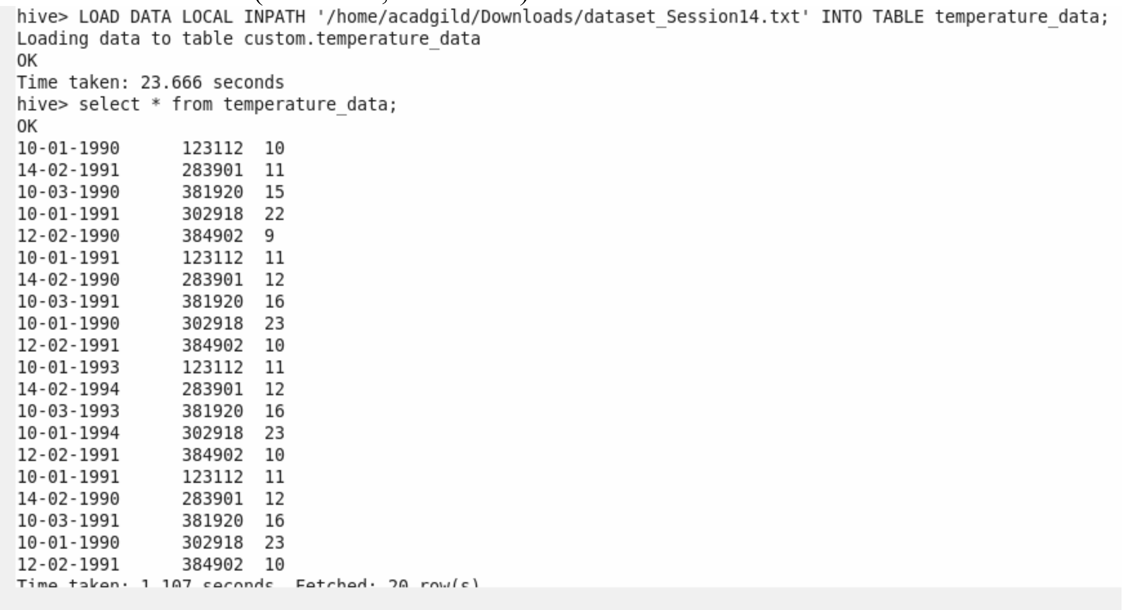
1. date (mm-dd-yyyy) format
2. zip code
3. temperature

The table will be loaded from comma-delimited file.



```
hive> create table if not exists temperature_data
    > (
    > date_c string,
    > zip_code int,
    > temperature int
    > )
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 7.189 seconds
hive> select * from temperature_data;
OK
Time taken: 11.232 seconds
```

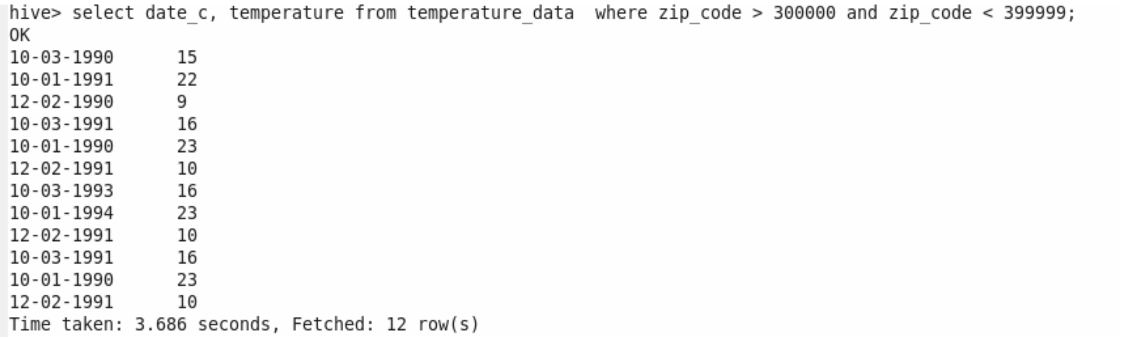
Load the dataset.txt (which is ',' delimited) in the table.



```
hive> LOAD DATA LOCAL INPATH '/home/acadgild/Downloads/dataset_Session14.txt' INTO TABLE temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 23.666 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: 1.107 seconds, Fetched: 20 row(s)
```

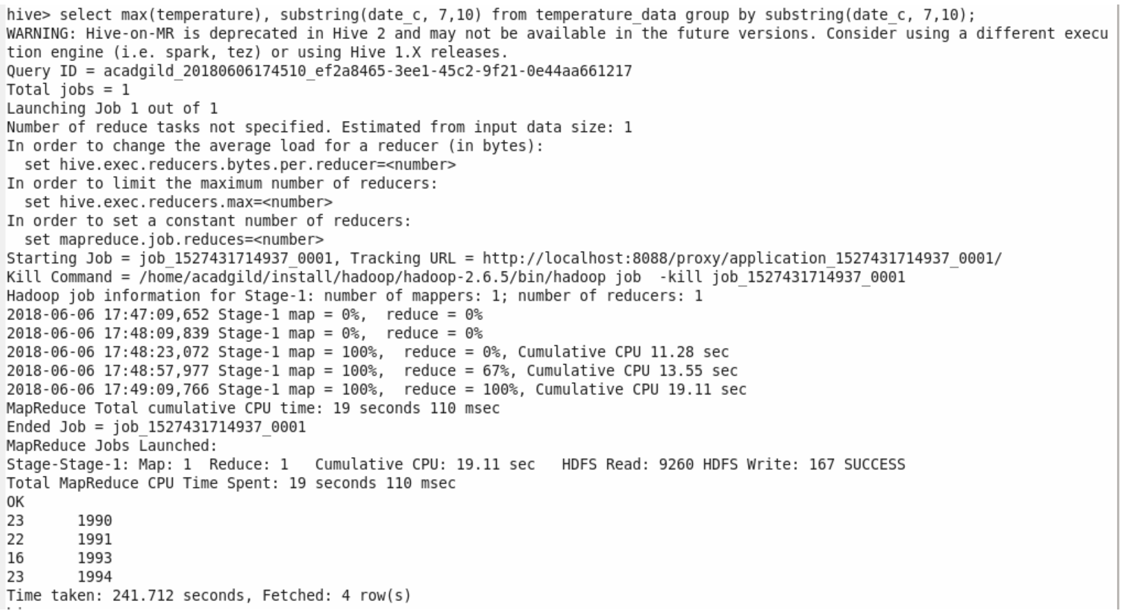
Task 2

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



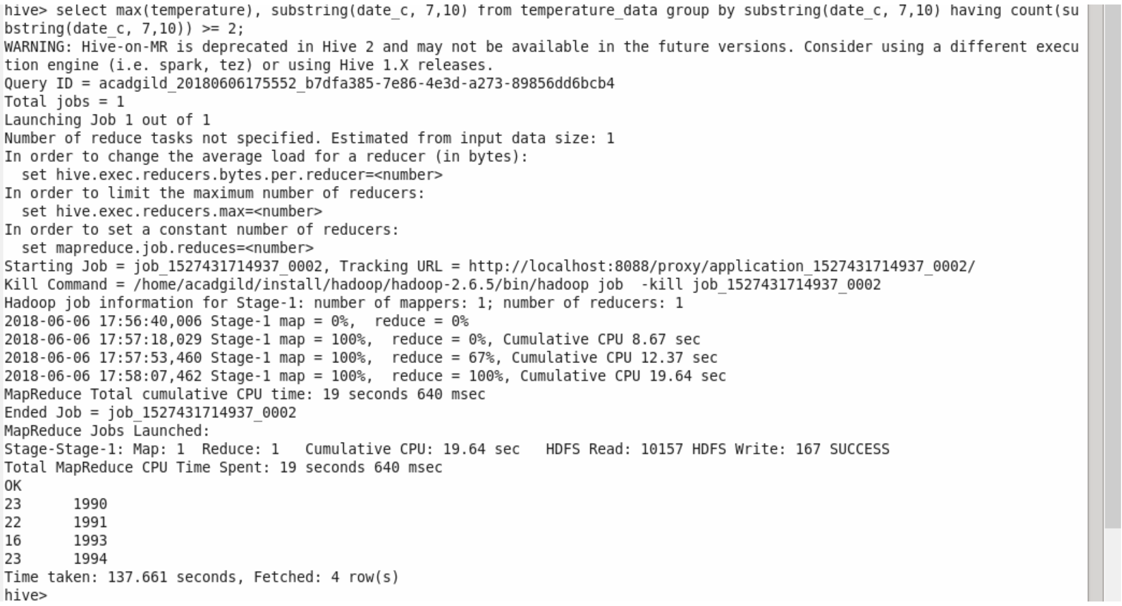
```
hive> select date_c, temperature from temperature_data  where zip_code > 300000 and zip_code < 399999;
OK
10-03-1990      15
10-01-1991      22
12-02-1990      9
10-03-1991      16
10-01-1990      23
12-02-1991      10
10-03-1993      16
10-01-1994      23
12-02-1991      10
10-03-1991      16
10-01-1990      23
12-02-1991      10
Time taken: 3.686 seconds, Fetched: 12 row(s)
```

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



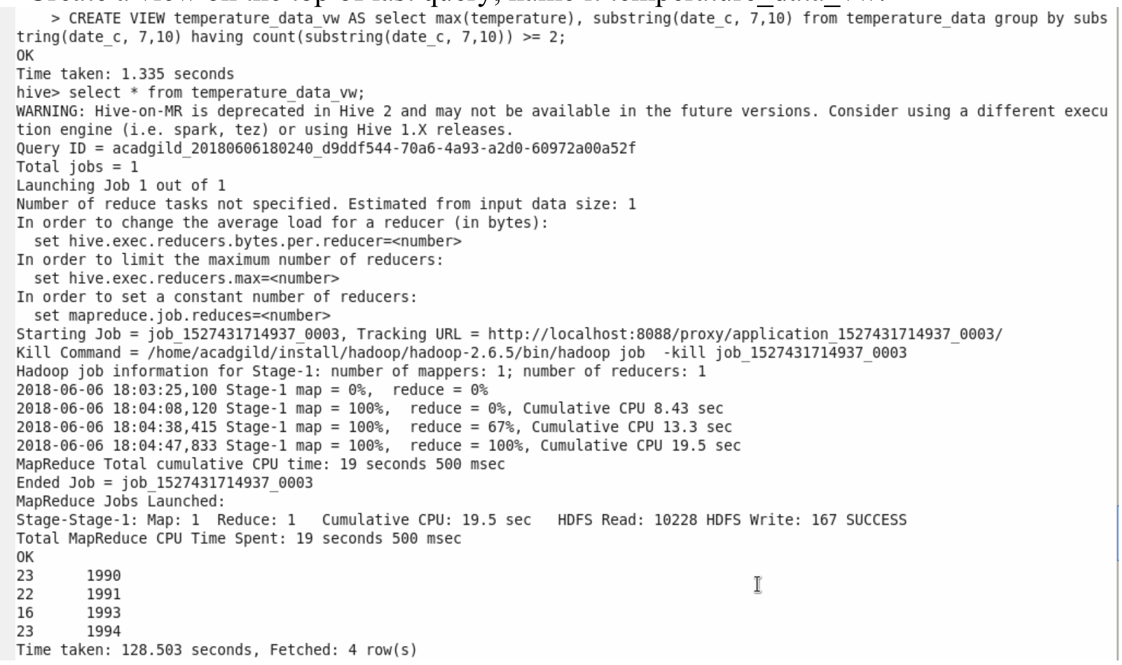
```
hive> select max(temperature), substring(date_c, 7,10) from temperature_data group by substring(date_c, 7,10);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180606174510_ef2a8465-3ee1-45c2-9f21-0e44aa661217
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_1527431714937_0001, Tracking URL = http://localhost:8088/proxy/application_1527431714937_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527431714937_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-06 17:47:09,652 Stage-1 map = 0%,  reduce = 0%
2018-06-06 17:48:09,839 Stage-1 map = 0%,  reduce = 0%
2018-06-06 17:48:23,072 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.28 sec
2018-06-06 17:48:57,977 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 13.55 sec
2018-06-06 17:49:09,766 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.11 sec
MapReduce Total cumulative CPU time: 19 seconds 110 msec
Ended Job = job_1527431714937_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 19.11 sec   HDFS Read: 9260 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 110 msec
OK
23      1990
22      1991
16      1993
23      1994
Time taken: 241.712 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.



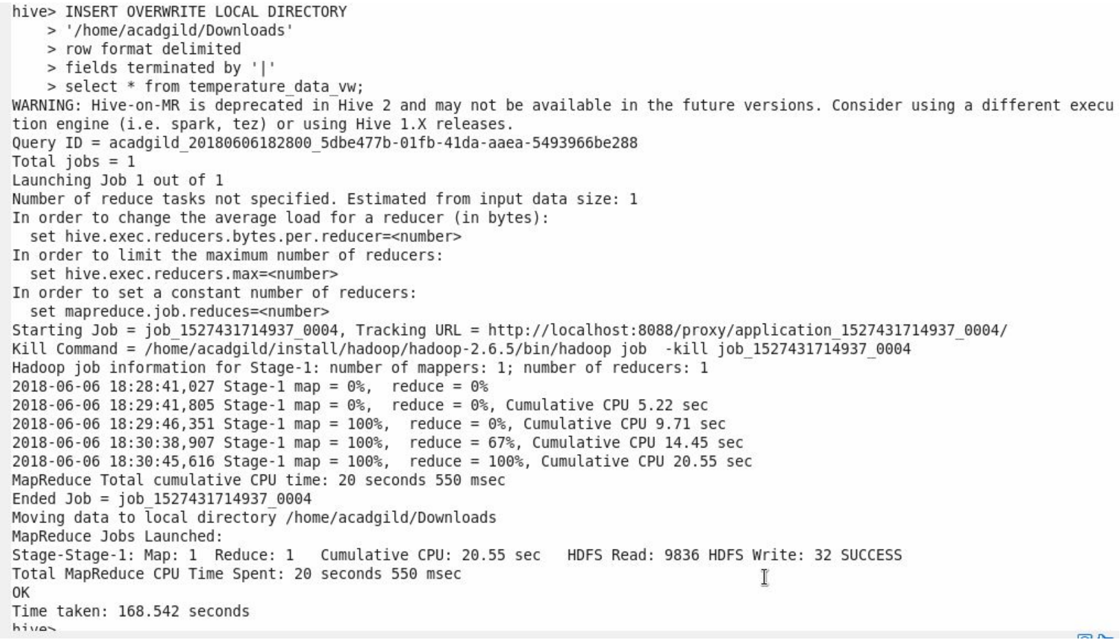
```
hive> select max(temperature), substring(date_c, 7,10) from temperature_data group by substring(date_c, 7,10) having count(su
bstring(date_c, 7,10)) >= 2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180606175552_b7dfa385-7e86-4e3d-a273-89856dd6bcb4
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_1527431714937_0002, Tracking URL = http://localhost:8088/proxy/application_1527431714937_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527431714937_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-06 17:56:40,006 Stage-1 map = 0%,  reduce = 0%
2018-06-06 17:57:18,029 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.67 sec
2018-06-06 17:57:53,460 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 12.37 sec
2018-06-06 17:58:07,462 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.64 sec
MapReduce Total cumulative CPU time: 19 seconds 640 msec
Ended Job = job_1527431714937_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 19.64 sec   HDFS Read: 10157 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 640 msec
OK
23      1990
22      1991
16      1993
23      1994
Time taken: 137.661 seconds, Fetched: 4 row(s)
hive>
```

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



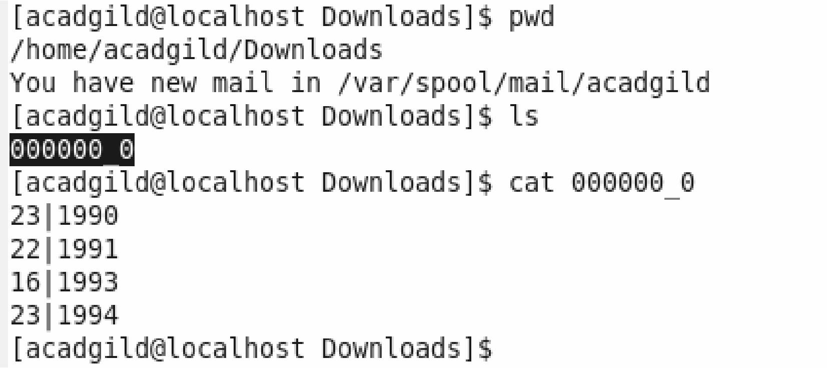
```
    > CREATE VIEW temperature_data_vw AS select max(temperature), substring(date_c, 7,10) from temperature_data group by subs
tring(date_c, 7,10) having count(substring(date_c, 7,10)) >= 2;
OK
Time taken: 1.335 seconds
hive> select * from temperature_data_vw;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180606180240_d9ddf544-70a6-4a93-a2d0-60972a00a52f
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_1527431714937_0003, Tracking URL = http://localhost:8088/proxy/application_1527431714937_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527431714937_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-06 18:03:25,100 Stage-1 map = 0%,  reduce = 0%
2018-06-06 18:04:08,120 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.43 sec
2018-06-06 18:04:38,415 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 13.3 sec
2018-06-06 18:04:47,833 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.5 sec
MapReduce Total cumulative CPU time: 19 seconds 500 msec
Ended Job = job_1527431714937_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 19.5 sec   HDFS Read: 10228 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 500 msec
OK
23      1990
22      1991
16      1993
23      1994
Time taken: 128.503 seconds, Fetched: 4 row(s)
```

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



```
hive> INSERT OVERWRITE LOCAL DIRECTORY
    > '/home/acadgild/Downloads'
    > 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 execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180606182800_5dbe477b-01fb-41da-aaea-5493966be288
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_1527431714937_0004, Tracking URL = http://localhost:8088/proxy/application_1527431714937_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527431714937_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-06 18:28:41,027 Stage-1 map = 0%,  reduce = 0%
2018-06-06 18:29:41,805 Stage-1 map = 0%,  reduce = 0%, Cumulative CPU 5.22 sec
2018-06-06 18:29:46,351 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.71 sec
2018-06-06 18:30:38,907 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 14.45 sec
2018-06-06 18:30:45,616 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 20.55 sec
MapReduce Total cumulative CPU time: 20 seconds 550 msec
Ended Job = job_1527431714937_0004
Moving data to local directory /home/acadgild/Downloads
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 20.55 sec   HDFS Read: 9836 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 550 msec
OK
Time taken: 168.542 seconds
hive>
```

Result



```
[acadgild@localhost Downloads]$ pwd
/home/acadgild/Downloads
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Downloads]$ ls
000000_0
[acadgild@localhost Downloads]$ cat 000000_0
23|1990
22|1991
16|1993
23|1994
[acadgild@localhost Downloads]$
```