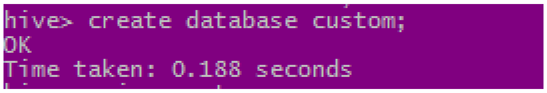
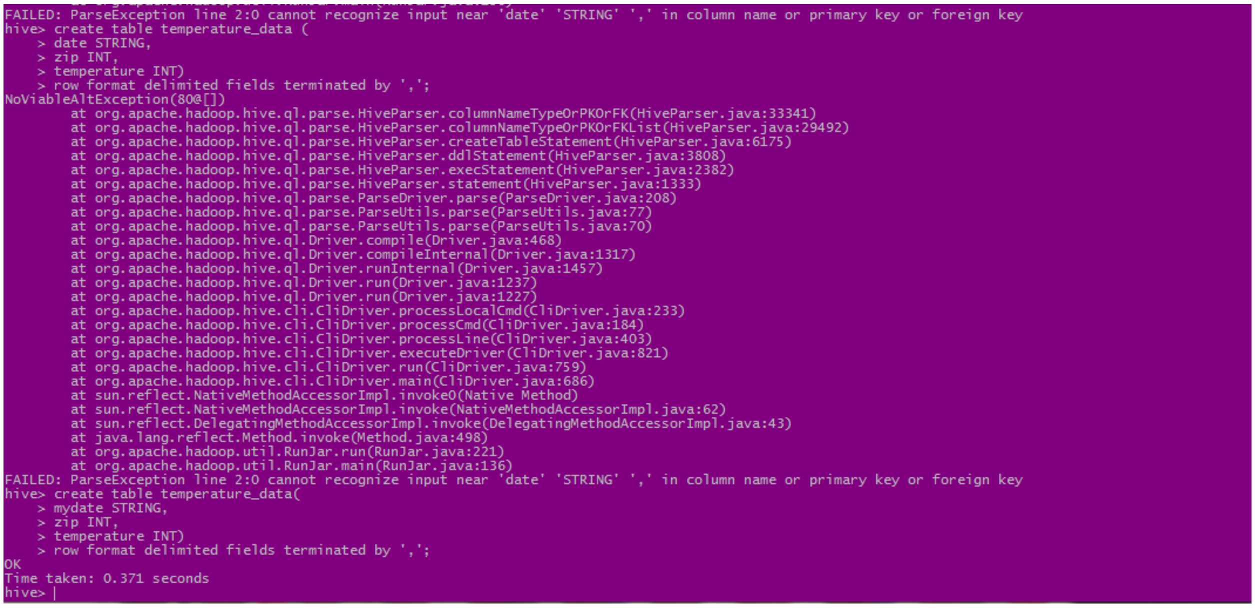
Session 8 Assignment 1

Task 1 Create a database named 'custom'.



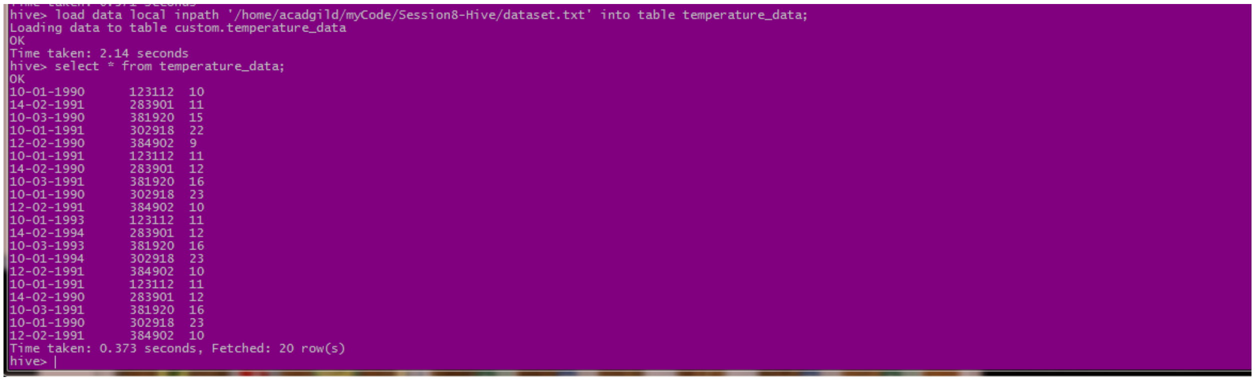
```
hive> create database custom;
OK
Time taken: 0.188 seconds
```

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.



```
FAILED: ParseException line 2:0 cannot recognize input near 'date' 'STRING' ',' in column name or primary key or foreign key
hive> create table temperature_data (
    > date STRING,
    > zip INT,
    > temperature INT)
    > row format delimited fields terminated by ',';
NoViableAltException(80@[])
        at org.apache.hadoop.hive.ql.parse.HiveParser.columnNameTypeOrPKOrFK(HiveParser.java:33341)
        at org.apache.hadoop.hive.ql.parse.HiveParser.columnNameTypeOrPKOrFKList(HiveParser.java:29492)
        at org.apache.hadoop.hive.ql.parse.HiveParser.createTableStatement(HiveParser.java:6175)
        at org.apache.hadoop.hive.ql.parse.HiveParser.ddlStatement(HiveParser.java:3808)
        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 2:0 cannot recognize input near 'date' 'STRING' ',' in column name or primary key or foreign key
hive> create table temperature_data(
    > mydate STRING,
    > zip INT,
    > temperature INT)
    > row format delimited fields terminated by ',';
OK
Time taken: 0.371 seconds
hive> |
```

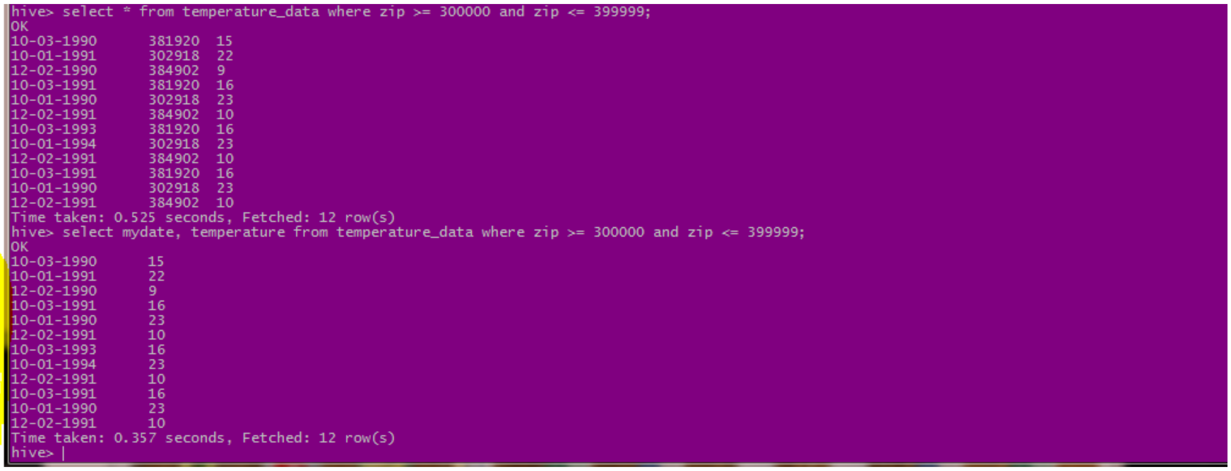
Load the dataset.txt (which is ',' delimited) in the table.



```
hive> load data local inpath '/home/acadgild/myCode/Session8-Hive/dataset.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 2.14 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: 0.373 seconds, Fetched: 20 row(s)
hive> |
```

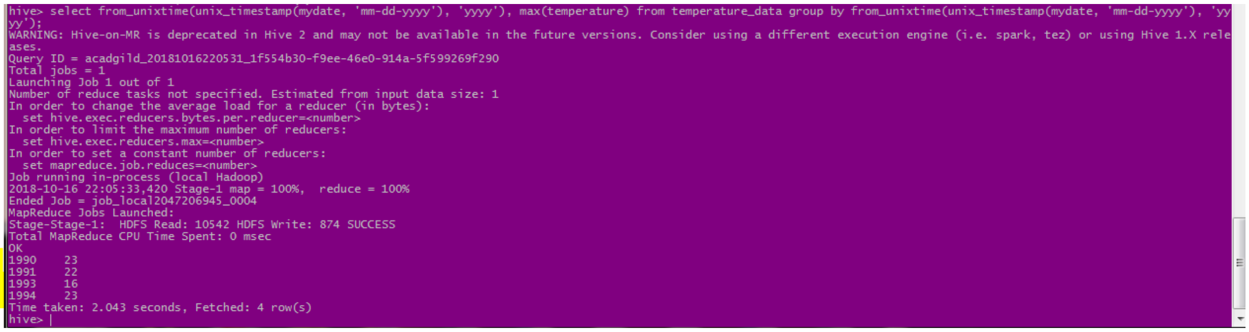
Task 2:

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



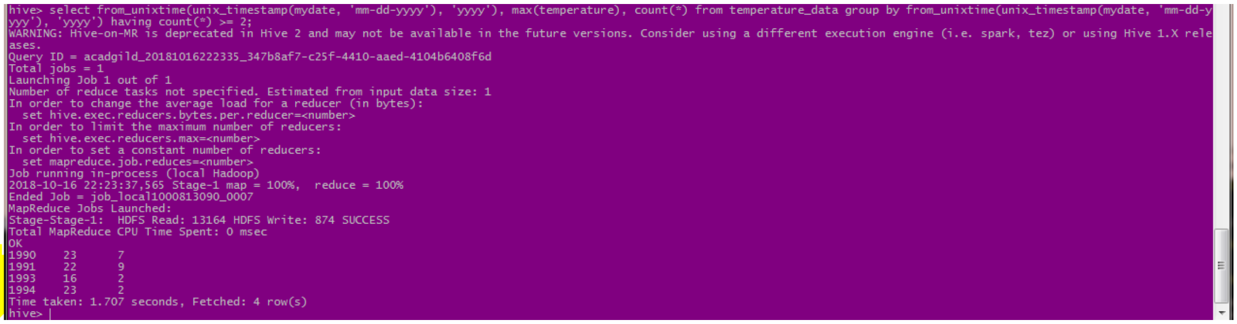
```
hive> select * from temperature_data where zip >= 300000 and zip <= 399999;
OK
10-03-1990      381920  15
10-01-1991      302918  22
12-02-1990      384902  9
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
10-03-1993      381920  16
10-01-1994      302918  23
12-02-1991      384902  10
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
Time taken: 0.525 seconds, Fetched: 12 row(s)
hive> select mydate, temperature from temperature_data where zip >= 300000 and zip <= 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: 0.357 seconds, Fetched: 12 row(s)
hive> |
```

Calculate maximum temperature corresponding to every year from temperature_data table.



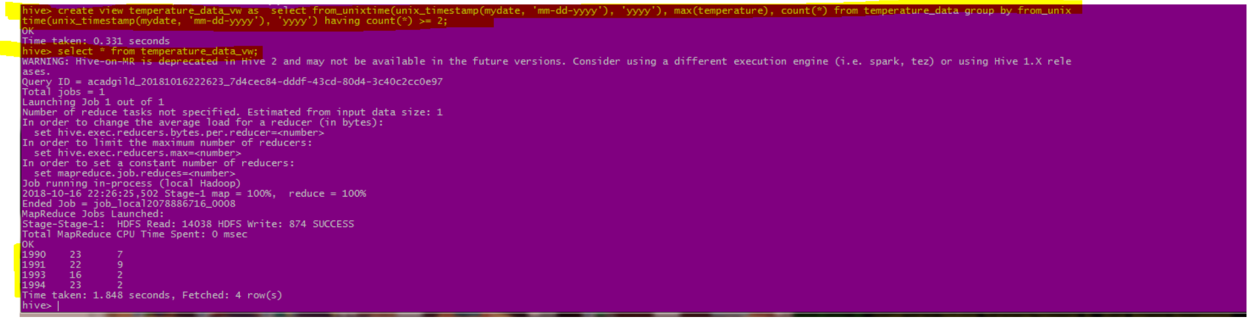
```
hive> select from_unixtime(unix_timestamp(mydate, 'mm-dd-yyyy'), 'yyyy'), max(temperature) from temperature_data group by from_unixtime(unix_timestamp(mydate, 'mm-dd-yyyy'), 'yy
yy');
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X rele
ases.
Query ID = acadgild_20181016220531_1f554b30-f9ee-46e0-914a-5f599269f290
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>
Job running in-process (local Hadoop)
2018-10-16 22:05:33,420 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local2047206945_0004
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 10542 HDFS Write: 874 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 2.043 seconds, Fetched: 4 row(s)
hive> |
```

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



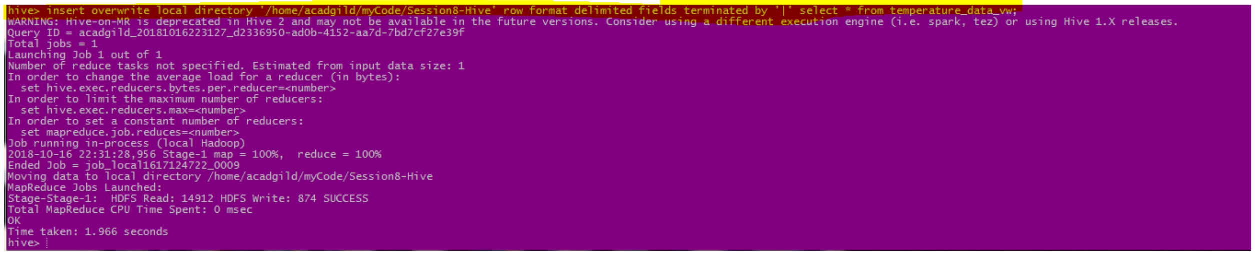
```
hive> select from_unixtime(unix_timestamp(mydate, 'mm-dd-yyyy'), 'yyyy'), max(temperature), count(*) from temperature_data group by from_unixtime(unix_timestamp(mydate, 'mm-dd-y
yyy'), 'yyyy') having count(*) >= 2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X rele
ases.
Query ID = acadgild_20181016222335_347b8af7-c25f-4410-aaed-4104b6408f6d
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>
Job running in-process (local Hadoop)
2018-10-16 22:23:37,565 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1000813090_0007
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 13164 HDFS Write: 874 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1990    23      7
1991    22      9
1993    16      2
1994    23      2
Time taken: 1.707 seconds, Fetched: 4 row(s)
hive> |
```

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

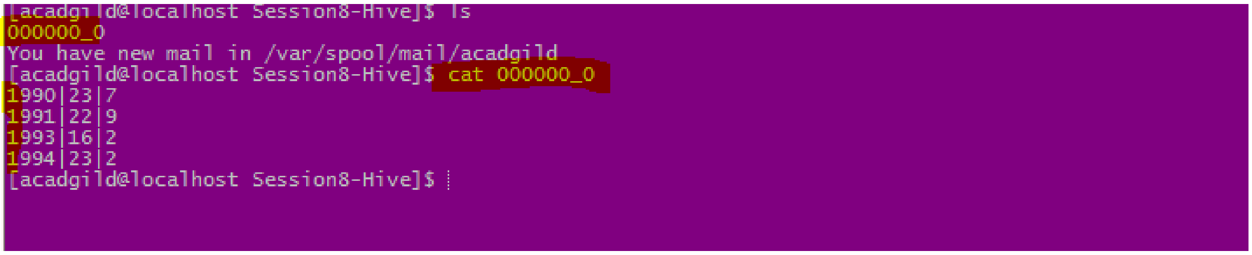


```
hive> create view temperature_data_vw as  select from_unixtime(unix_timestamp(mydate, 'mm-dd-yyyy'), 'yyyy'), max(temperature), count(*) from temperature_data group by from_unix
time(unix_timestamp(mydate, 'mm-dd-yyyy'), 'yyyy') having count(*) >= 2;
OK
Time taken: 0.331 seconds
hive> select * from 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 rele
ases.
Query ID = acadgild_20181016222623_7d4cec84-dddf-43cd-80d4-3c40c2cc0e97
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>
Job running in-process (local Hadoop)
2018-10-16 22:26:25,502 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local2078886716_0008
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 14038 HDFS Write: 874 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1990    23      7
1991    22      9
1993    16      2
1994    23      2
Time taken: 1.848 seconds, Fetched: 4 row(s)
hive> |
```

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/myCode/Session8-Hive' 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_20181016223127_d2336950-ad0b-4152-aa7d-7bd7cf27e39f
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>
Job running in-process (local Hadoop)
2018-10-16 22:31:28,956 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1617124722_0009
Moving data to local directory /home/acadgild/myCode/Session8-Hive
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 14912 HDFS Write: 874 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 1.966 seconds
hive>
```



```
[acadgild@localhost Session8-Hive]$ ls
000000_0
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Session8-Hive]$ cat 000000_0
1990|23|7
1991|22|9
1993|16|2
1994|23|2
[acadgild@localhost Session8-Hive]$
```