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

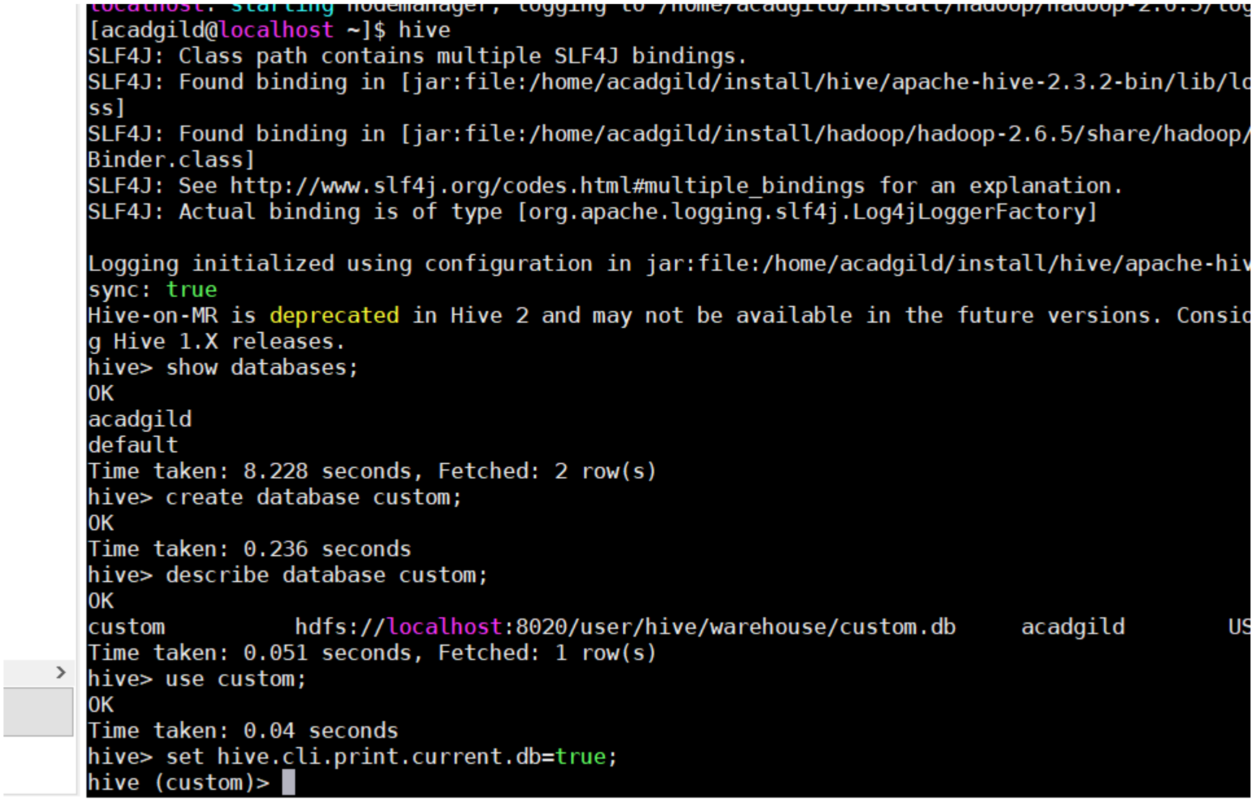
3. temperature

The table will be loaded from comma-delimited file.

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

[acadgild@localhost ~]$ **hive**

hive> **create database custom;**
OK
Time taken: 0.236 seconds
hive> **describe database custom;**
OK
custom	hdfs://localhost:8020/user/hive/warehouse/custom.db	acadgild	USER
Time taken: 0.051 seconds, Fetched: 1 row(s)
hive> **use custom;**
OK
Time taken: 0.04 seconds
hive> **set hive.cli.print.current.db=true;**
hive (custom)>



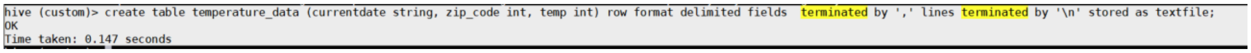
```
localhost: starting nodemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/log
[acadgild@localhost ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.2-bin/lib/lo
ss]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/
Binder.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-hiv
sync: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consic
g Hive 1.X releases.
hive> show databases;
OK
acadgild
default
Time taken: 8.228 seconds, Fetched: 2 row(s)
hive> create database custom;
OK
Time taken: 0.236 seconds
hive> describe database custom;
OK
custom          hdfs://localhost:8020/user/hive/warehouse/custom.db     acadgild        US
Time taken: 0.051 seconds, Fetched: 1 row(s)
hive> use custom;
OK
Time taken: 0.04 seconds
hive> set hive.cli.print.current.db=true;
hive (custom)>
```

***hive (custom)*>** **create table temperature_data (currentdate string, zip_code int, temp int) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;**

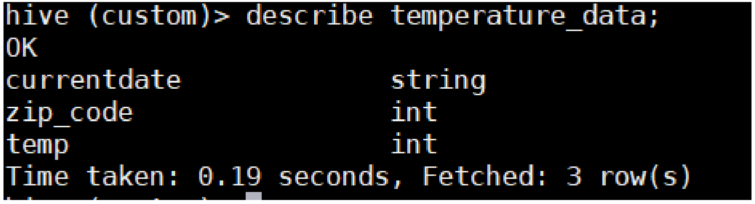
OK

Time taken: 0.147 seconds



```
hive (custom)> create table temperature_data (currentdate string, zip_code int, temp int) row format delimited fields  terminated by ',' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.147 seconds
```

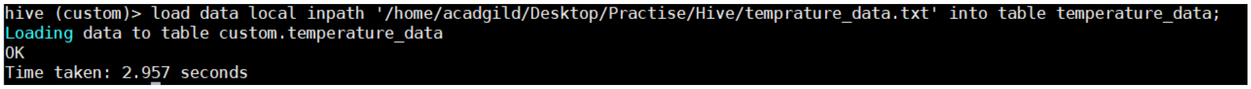
***hive (custom)*>** **describe temperature_data;**



```
hive (custom)> describe temperature_data;
OK
currentdate             string
zip_code                int
temp                    int
Time taken: 0.19 seconds, Fetched: 3 row(s)
```

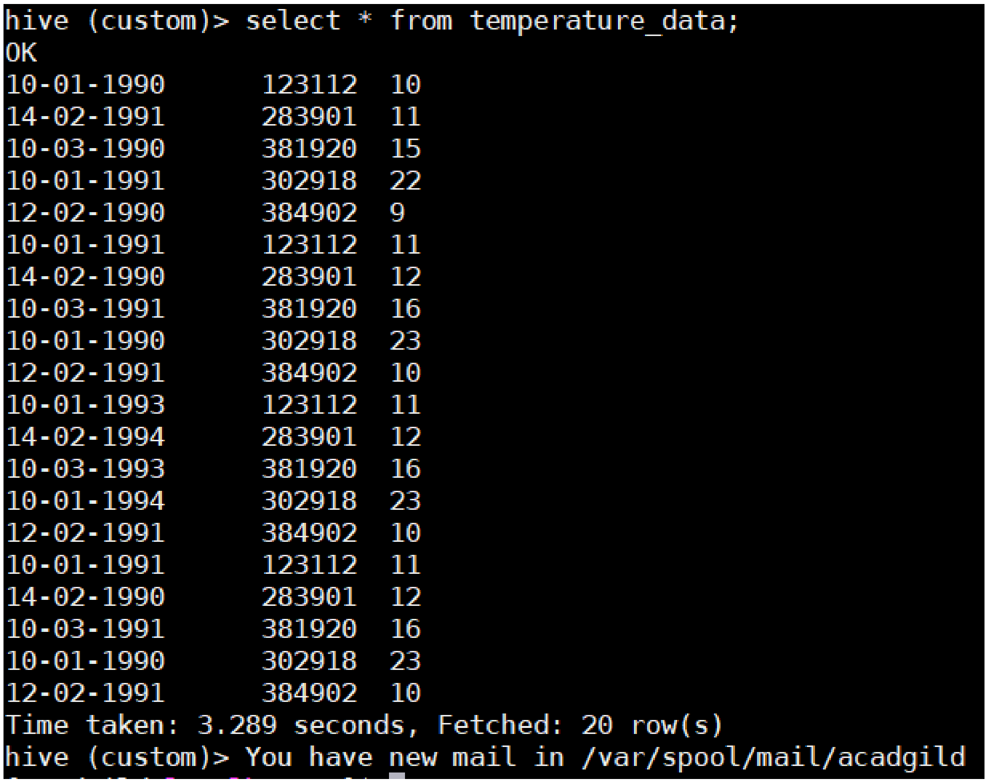
LOAD DATA to the new table temperature_data

***hive (custom)*>** **load data local inpath '/home/acadgild/Desktop/Practise/Hive/temprature_data.txt' into table temperature_data;**



```
hive (custom)> load data local inpath '/home/acadgild/Desktop/Practise/Hive/temprature_data.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 2.957 seconds
```

***hive (custom)*>** **select * from temperature_data;**

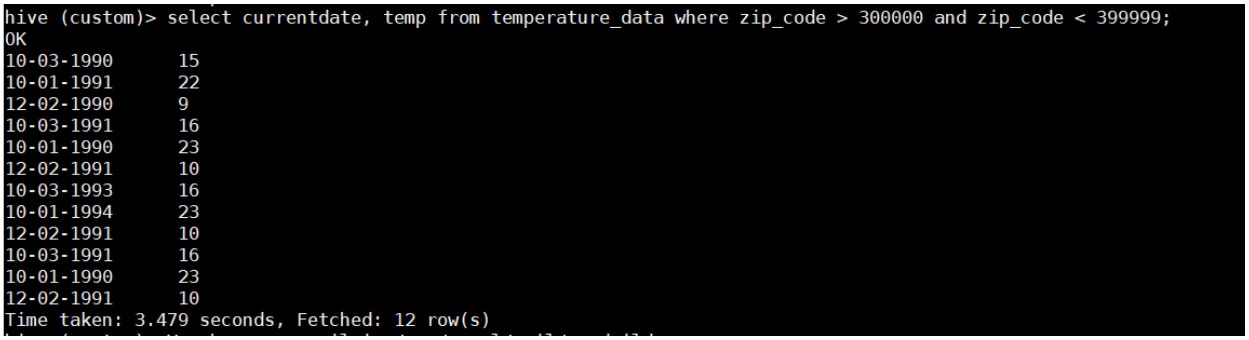


```
hive (custom)> select * from temperature_data;
OK
10-01-1990      123112  10
14-02-1991      283901  11
10-03-1990      381920  15
10-01-1991      302918  22
12-02-1990      384902  9
10-01-1991      123112  11
14-02-1990      283901  12
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      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: 3.289 seconds, Fetched: 20 row(s)
hive (custom)> You have new mail in /var/spool/mail/acadgild
```

Task 2

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

***hive (custom)*>** **select currentdate, temp from temperature_data where zip_code > 300000 and zip_code < 399999;**

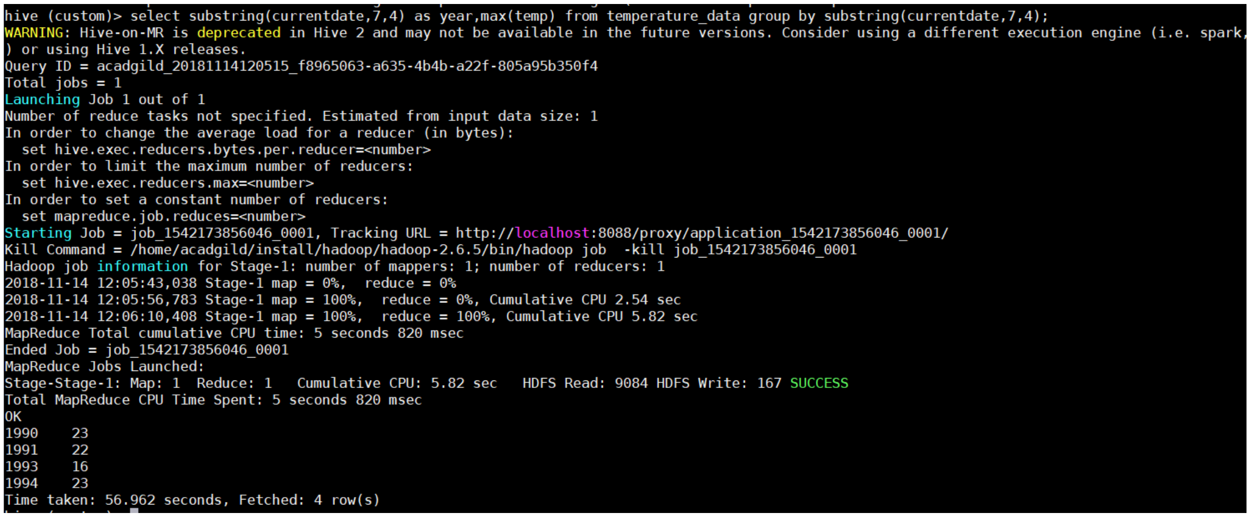


```
hive (custom)> select currentdate, temp from temperature_data where zip_code > 300000 and zip_code < 399999;
OK
10-03-1990      15
10-01-1991      22
12-02-1990      9
10-03-1991      16
10-01-1990      23
12-02-1991      10
10-03-1993      16
10-01-1994      23
12-02-1991      10
10-03-1991      16
10-01-1990      23
12-02-1991      10
Time taken: 3.479 seconds, Fetched: 12 row(s)
```

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

***hive (custom)*>** **select substring(currentdate,7,4) as year,max(temp) from temperature_data group by substring(currentdate,7,4);**

1990 23
1991 22
1993 16
1994 23
Time taken: 56.962 seconds, Fetched: 4 row(s)



```
hive (custom)> select substring(currentdate,7,4) as year,max(temp) from temperature_data group by substring(currentdate,7,4);
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,
) or using Hive 1.X releases.
Query ID = acadgild_20181114120515_f8965063-a635-4b4b-a22f-805a95b350f4
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_1542173856046_0001, Tracking URL = http://localhost:8088/proxy/application_1542173856046_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542173856046_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-14 12:05:43,038 Stage-1 map = 0%,  reduce = 0%
2018-11-14 12:05:56,783 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.54 sec
2018-11-14 12:06:10,408 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.82 sec
MapReduce Total cumulative CPU time: 5 seconds 820 msec
Ended Job = job_1542173856046_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.82 sec   HDFS Read: 9084 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 820 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 56.962 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**

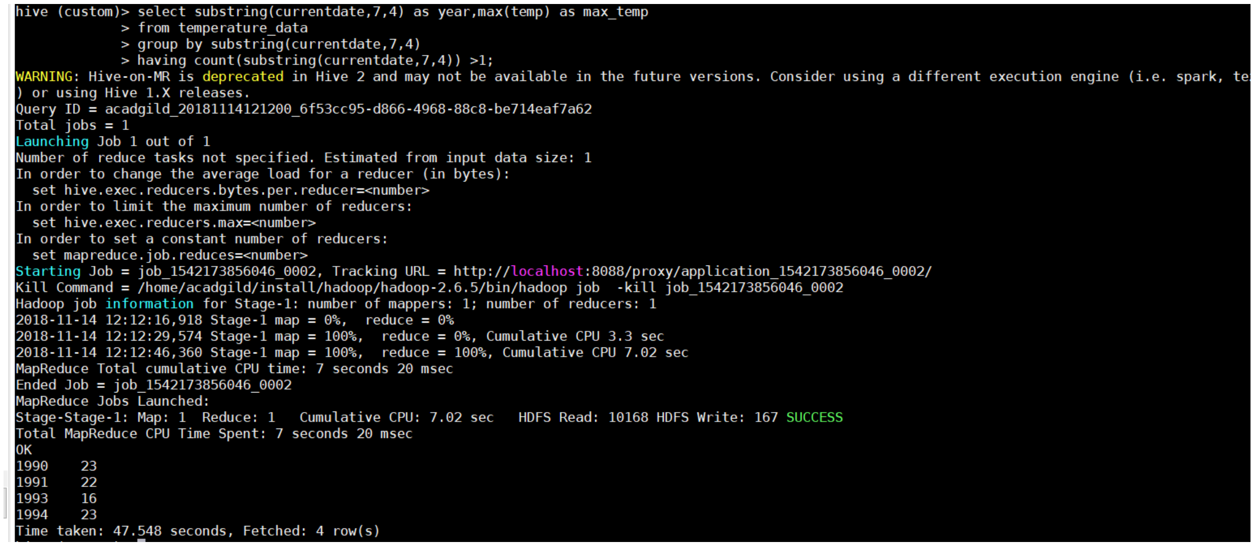
hive (custom)> **select substring(currentdate,7,4) as year,max(temp) as max_temp**

> from temperature_data

> group by substring(currentdate,7,4)

> having count(substring(currentdate,7,4)) >1;

```
1990    23
1991    22
1993    16
1994    23
Time taken: 47.548 seconds, Fetched: 4 row(s)
```



```
hive (custom)> select substring(currentdate,7,4) as year,max(temp) as max_temp
             > from temperature_data
             > group by substring(currentdate,7,4)
             > having count(substring(currentdate,7,4)) >1;
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, te
) or using Hive 1.X releases.
Query ID = acadgild_20181114121200_6f53cc95-d866-4968-88c8-be714eaf7a62
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_1542173856046_0002, Tracking URL = http://localhost:8088/proxy/application_1542173856046_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542173856046_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-14 12:12:16,918 Stage-1 map = 0%,  reduce = 0%
2018-11-14 12:12:29,574 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.3 sec
2018-11-14 12:12:46,360 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.02 sec
MapReduce Total cumulative CPU time: 7 seconds 20 msec
Ended Job = job_1542173856046_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.02 sec   HDFS Read: 10168 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 20 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 47.548 seconds, Fetched: 4 row(s)
```

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

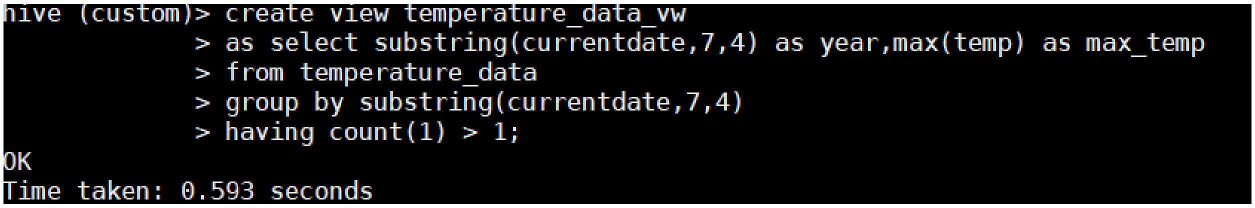
hive (custom)> **create view temperature_data_vw**

> as select substring(currentdate,7,4) as year,max(temp) as max_temp

> from temperature_data

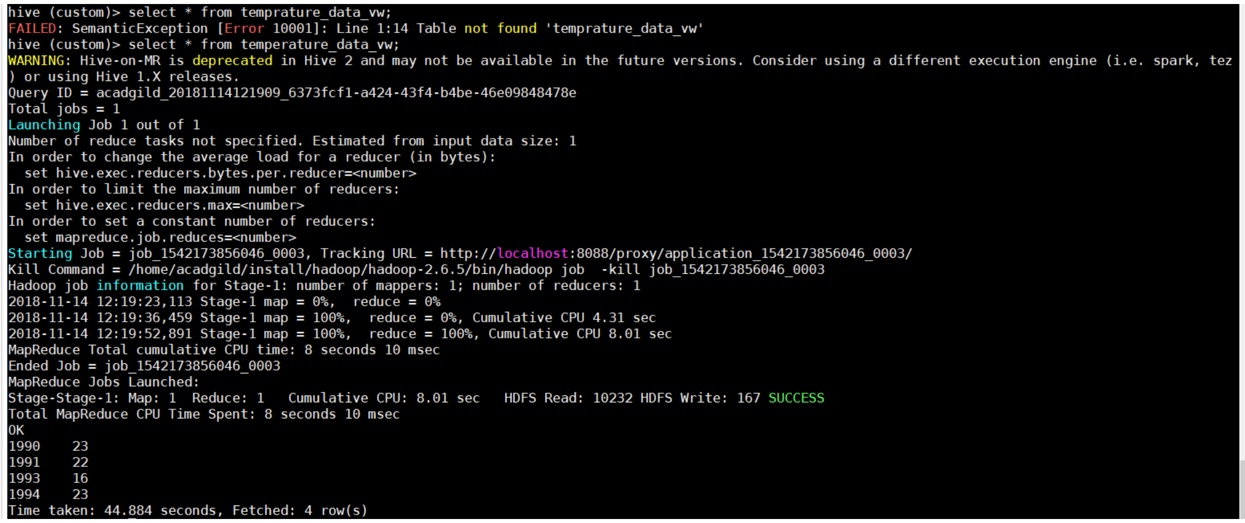
> group by substring(currentdate,7,4)

> having count(1) > 1;



```
hive (custom)> create view temperature_data_vw
             > as select substring(currentdate,7,4) as year,max(temp) as max_temp
             > from temperature_data
             > group by substring(currentdate,7,4)
             > having count(1) > 1;
OK
Time taken: 0.593 seconds
```

hive (custom)> **select * from temperature_data_vw;**

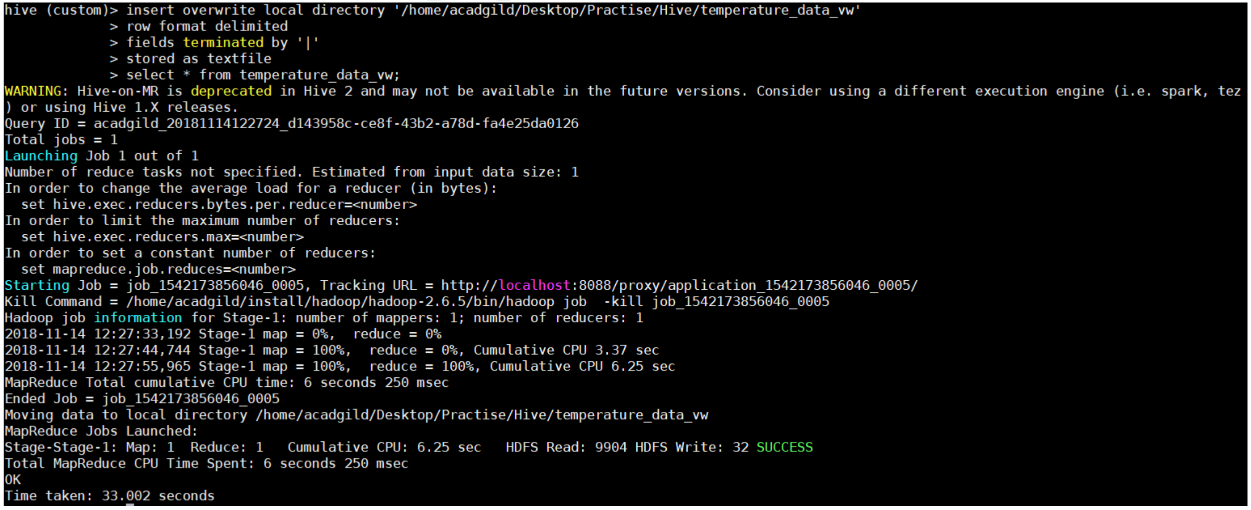


```
hive (custom)> select * from temprature_data_vw;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'temprature_data_vw'
hive (custom)> select * from temperature_data_vw;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez
) or using Hive 1.X releases.
Query ID = acadgild_20181114121909_6373fcf1-a424-43f4-b4be-46e09848478e
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_1542173856046_0003, Tracking URL = http://localhost:8088/proxy/application_1542173856046_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542173856046_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-14 12:19:23,113 Stage-1 map = 0%,  reduce = 0%
2018-11-14 12:19:36,459 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.31 sec
2018-11-14 12:19:52,891 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.01 sec
MapReduce Total cumulative CPU time: 8 seconds 10 msec
Ended Job = job_1542173856046_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.01 sec   HDFS Read: 10232 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 10 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 44.884 seconds, Fetched: 4 row(s)
```

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

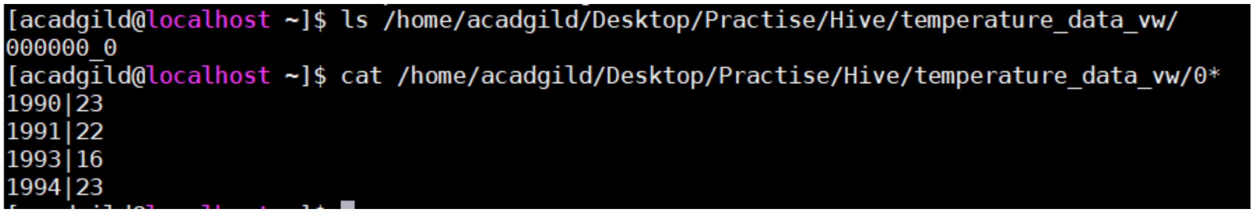
hive (custom)> **insert overwrite local directory '/home/acadgild/Desktop/Practise/Hive/temperature_data_vw'**

```
        > row format delimited
        > fields terminated by '|'
        > stored as textfile
        > select * from temperature_data_vw;
```



```
hive (custom)> insert overwrite local directory '/home/acadgild/Desktop/Practise/Hive/temperature_data_vw'
             > row format delimited
             > fields terminated by '|'
             > stored as textfile
             > 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_20181114122724_d143958c-ce8f-43b2-a78d-fa4e25da0126
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_1542173856046_0005, Tracking URL = http://localhost:8088/proxy/application_1542173856046_0005/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1542173856046_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-11-14 12:27:33,192 Stage-1 map = 0%,  reduce = 0%
2018-11-14 12:27:44,744 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.37 sec
2018-11-14 12:27:55,965 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.25 sec
MapReduce Total cumulative CPU time: 6 seconds 250 msec
Ended Job = job_1542173856046_0005
Moving data to local directory /home/acadgild/Desktop/Practise/Hive/temperature_data_vw
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.25 sec   HDFS Read: 9904 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 250 msec
OK
Time taken: 33.002 seconds
```

[acadgild@localhost ~]$ **cat /home/acadgild/Desktop/Practise/Hive/temperature_data_vw/0***



```
[acadgild@localhost ~]$ ls /home/acadgild/Desktop/Practise/Hive/temperature_data_vw/
000000_0
[acadgild@localhost ~]$ cat /home/acadgild/Desktop/Practise/Hive/temperature_data_vw/0*
1990|23
1991|22
1993|16
1994|23
```