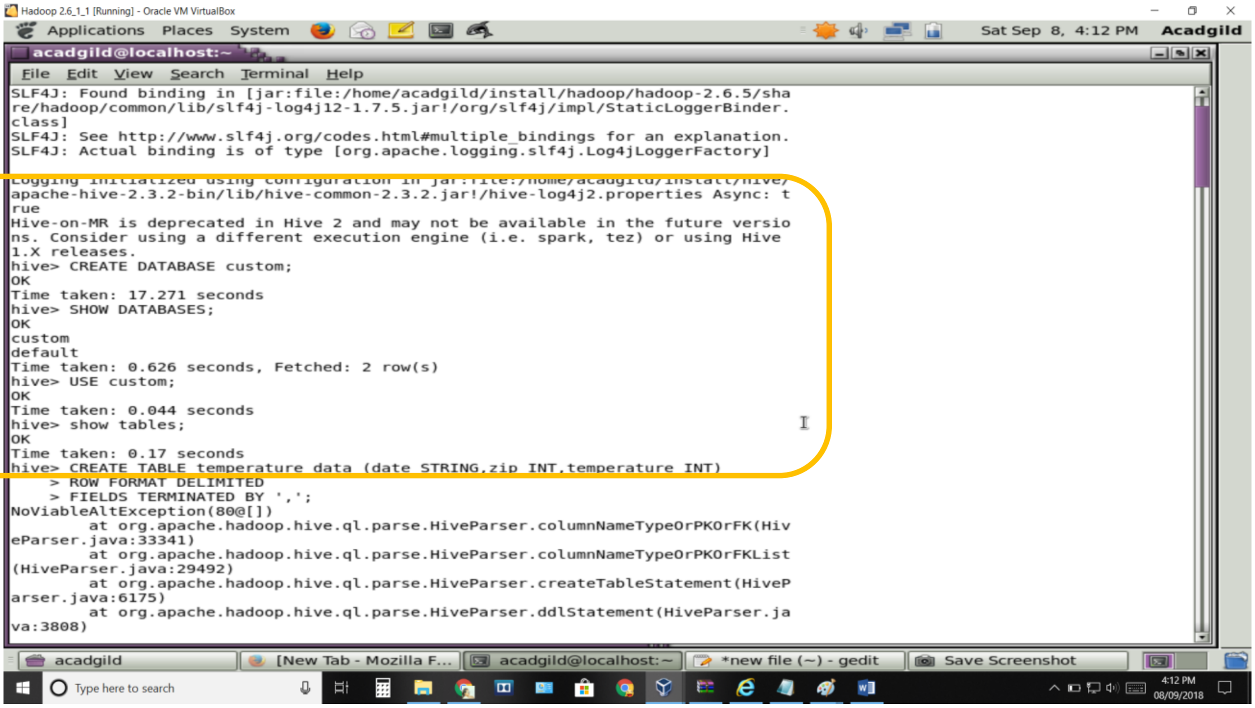
Big Data Hadoop ‘Session 8 HIVE BASICS’

DATASET

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

Task 1

- Create a database named 'custom'



```
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/sha
re/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: t
rue
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
hive> CREATE DATABASE custom;
OK
Time taken: 17.271 seconds
hive> SHOW DATABASES;
OK
custom
default
Time taken: 0.626 seconds, Fetched: 2 row(s)
hive> USE custom;
OK
Time taken: 0.044 seconds
hive> show tables;
OK
Time taken: 0.17 seconds
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(Hiv
eParser.java:33341)
        at org.apache.hadoop.hive.ql.parse.HiveParser.columnNameTypeOrPKOrFKList
(HiveParser.java:29492)
        at org.apache.hadoop.hive.ql.parse.HiveParser.createTableStatement(HiveP
arser.java:6175)
        at org.apache.hadoop.hive.ql.parse.HiveParser.ddlStatement(HiveParser.ja
va:3808)
```

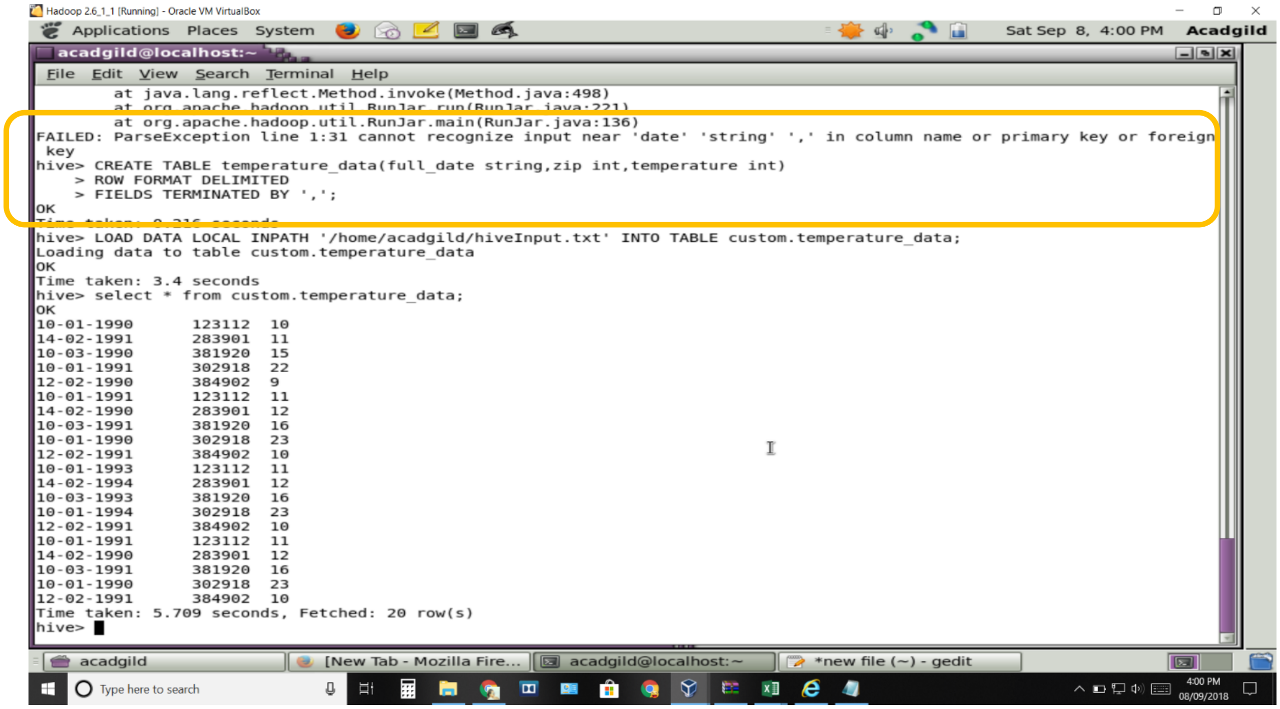
- Create a table named temperature_data inside custom having below fields:

  1. date (mm-dd-yyyy) format

  2. zip code

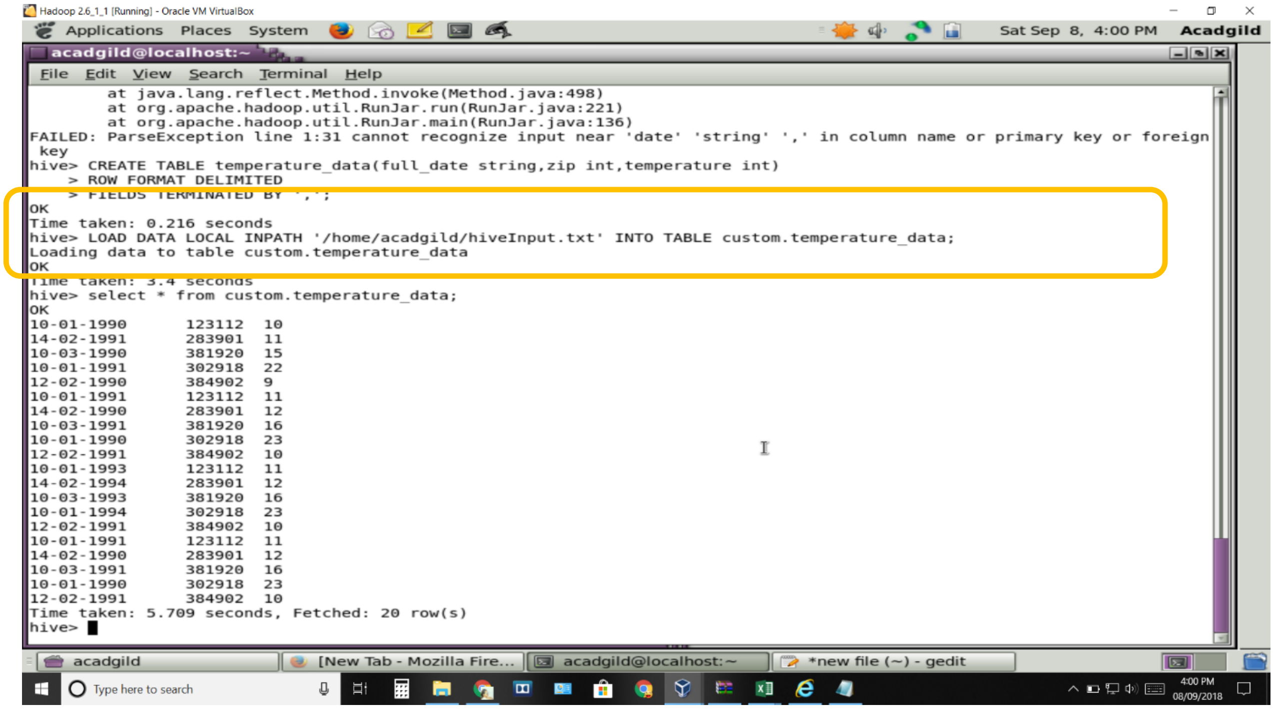
  3. temperature

  The table will be loaded from comma-delimited file.


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Sat Sep 8, 4:00 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
acadgild
[New Tab - Mozilla Fire...
acadgild@localhost:~
*new file (~) - gedit
Type here to search
4:00 PM
08/09/2018

```
        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 1:31 cannot recognize input near 'date' 'string' ',' in column name or primary key or foreign
 key
hive> CREATE TABLE temperature_data(full_date string,zip int,temperature int)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.216 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/hiveInput.txt' INTO TABLE custom.temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 3.4 seconds
hive> select * from custom.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: 5.709 seconds, Fetched: 20 row(s)
hive>
```

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

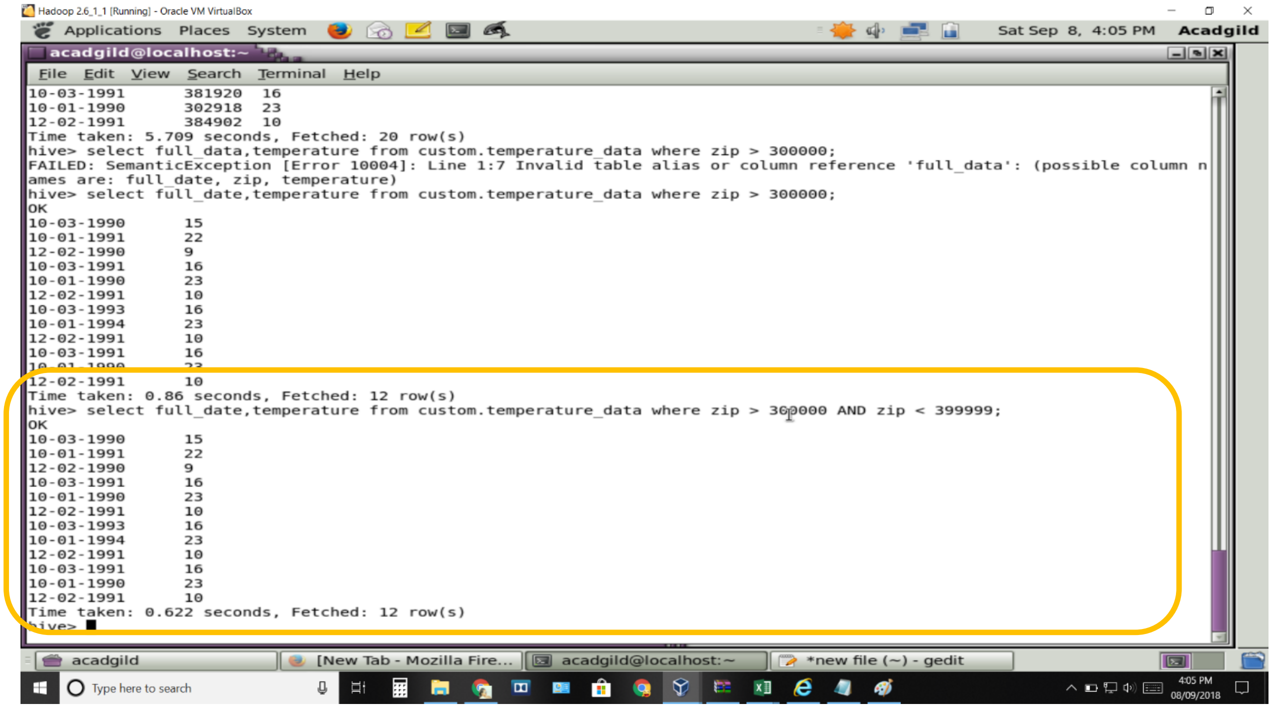

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Sat Sep 8, 4:00 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
acadgild
[New Tab - Mozilla Fire...
acadgild@localhost:~
*new file (~) - gedit
Type here to search
4:00 PM
08/09/2018

```
        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 1:31 cannot recognize input near 'date' 'string' ',' in column name or primary key or foreign
 key
hive> CREATE TABLE temperature_data(full_date string,zip int,temperature int)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 0.216 seconds
hive> LOAD DATA LOCAL INPATH '/home/acadgild/hiveInput.txt' INTO TABLE custom.temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 3.4 seconds
hive> select * from custom.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: 5.709 seconds, Fetched: 20 row(s)
hive>
```

Task 2

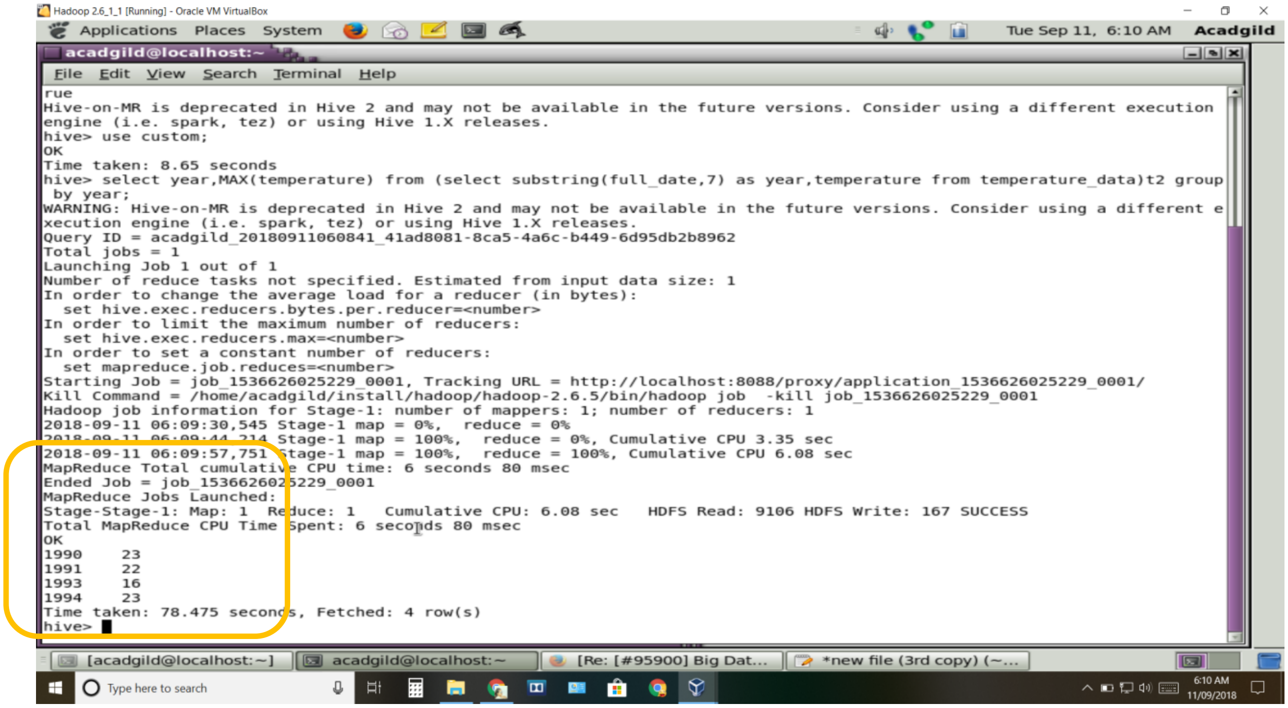
● Fetch date and temperature from temperature_data where zip code is greater than

300000 and less than 399999.


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Sat Sep 8, 4:05 PM
Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
acadgild
[New Tab - Mozilla Fire...
acadgild@localhost:~
*new file (~) - gedit
Type here to search
4:05 PM
08/09/2018

```
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
Time taken: 5.709 seconds, Fetched: 20 row(s)
hive> select full_data,temperature from custom.temperature_data where zip > 300000;
FAILED: SemanticException [Error 10004]: Line 1:7 Invalid table alias or column reference 'full_data': (possible column n
ames are: full_date, zip, temperature)
hive> select full_date,temperature from custom.temperature_data where zip > 300000;
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: 0.86 seconds, Fetched: 12 row(s)
hive> select full_date,temperature from custom.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: 0.622 seconds, Fetched: 12 row(s)
hive>
```

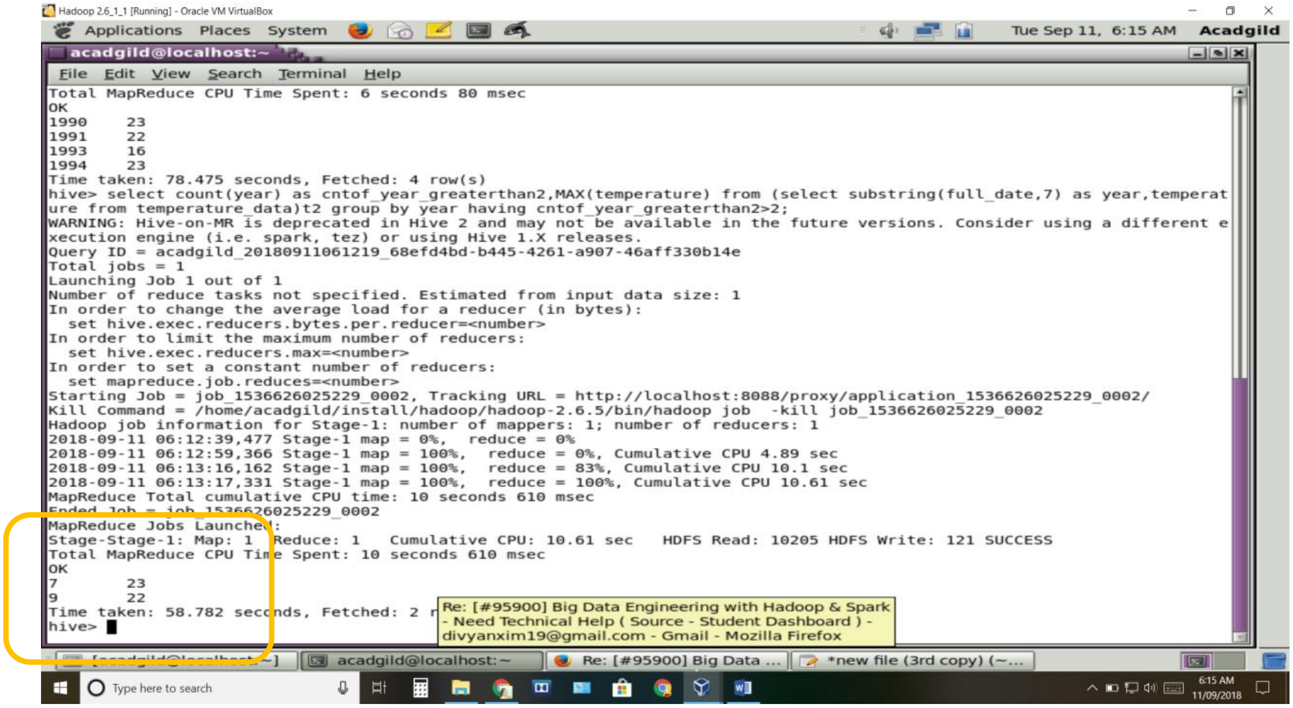
- Calculate maximum temperature corresponding to every year from temperature_data table.


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Tue Sep 11, 6:10 AM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost:~]
acadgild@localhost:~
[Re: [#95900] Big Dat...
*new file (3rd copy) (~...
Type here to search
6:10 AM
11/09/2018

```
rue
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> use custom;
OK
Time taken: 8.65 seconds
hive> select year,MAX(temperature) from (select substring(full_date,7) as year,temperature from temperature_data)t2 group
 by year;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different e
xecution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180911060841_41ad8081-8ca5-4a6c-b449-6d95db2b8962
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_1536626025229_0001, Tracking URL = http://localhost:8088/proxy/application_1536626025229_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1536626025229_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-11 06:09:30,545 Stage-1 map = 0%,  reduce = 0%
2018-09-11 06:09:44,214 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.35 sec
2018-09-11 06:09:57,751 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.08 sec
MapReduce Total cumulative CPU time: 6 seconds 80 msec
Ended Job = job_1536626025229_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.08 sec   HDFS Read: 9106 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 80 msec
OK
1990	23
1991	22
1993	16
1994	23
Time taken: 78.475 seconds, Fetched: 4 row(s)
hive> 
```

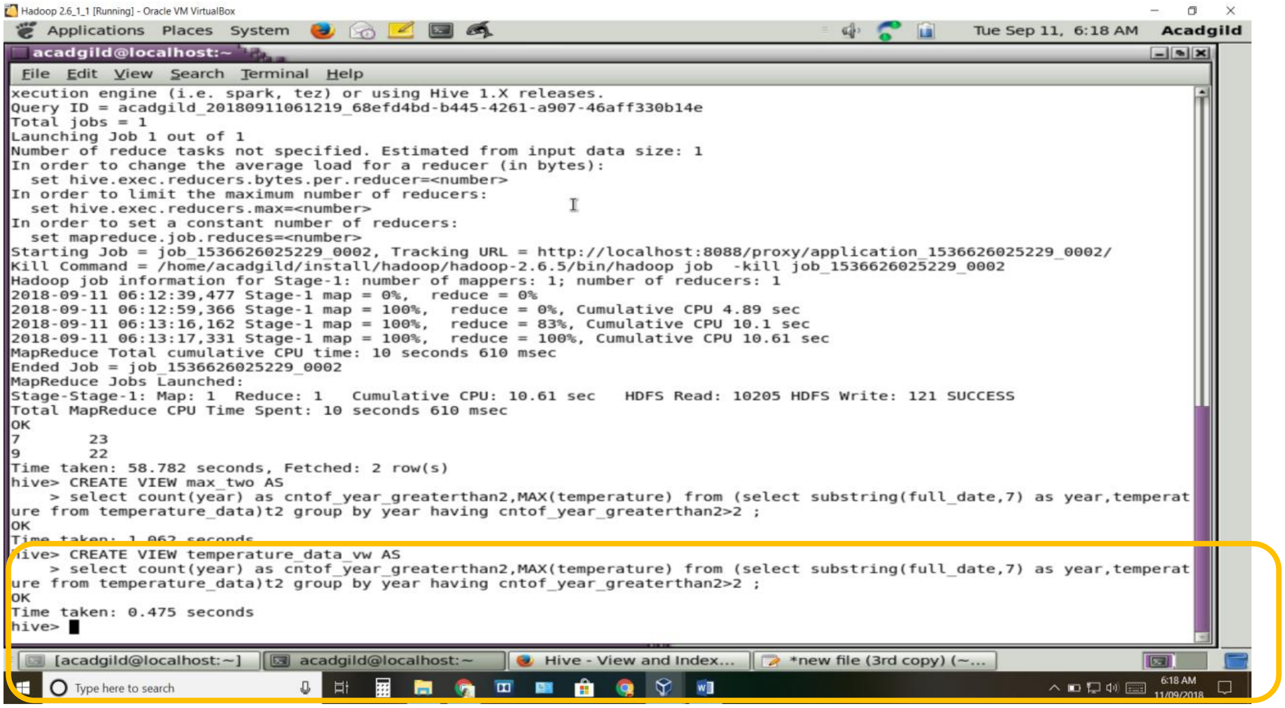
- Calculate maximum temperature from temperature_data table corresponding to those

years which have at least 2 entries in the table.


Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Tue Sep 11, 6:15 AM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
Re: [#95900] Big Data Engineering with Hadoop & Spark - Need Technical Help (Source - Student Dashboard) - divyanxim19@gmail.com - Gmail - Mozilla Firefox
[acadgild@localhost:~]
acadgild@localhost:~
Re: [#95900] Big Data ...
*new file (3rd copy) (~...
Type here to search
6:15 AM
11/09/2018

```
Total MapReduce CPU Time Spent: 6 seconds 80 msec
OK
1990	23
1991	22
1993	16
1994	23
Time taken: 78.475 seconds, Fetched: 4 row(s)
hive> select count(year) as cntof_year_greaterthan2,MAX(temperature) from (select substring(full_date,7) as year,temperat
ure from temperature_data)t2 group by year having cntof_year_greaterthan2>2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different e
xecution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180911061219_68efd4bd-b445-4261-a907-46aff330b14e
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_1536626025229_0002, Tracking URL = http://localhost:8088/proxy/application_1536626025229_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1536626025229_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-11 06:12:39,477 Stage-1 map = 0%,  reduce = 0%
2018-09-11 06:12:59,366 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.89 sec
2018-09-11 06:13:16,162 Stage-1 map = 100%,  reduce = 83%, Cumulative CPU 10.1 sec
2018-09-11 06:13:17,331 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.61 sec
MapReduce Total cumulative CPU time: 10 seconds 610 msec
Ended Job = job_1536626025229_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.61 sec   HDFS Read: 10205 HDFS Write: 121 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 610 msec
OK
7	23
9	22
Time taken: 58.782 seconds, Fetched: 2 r
hive> 
```

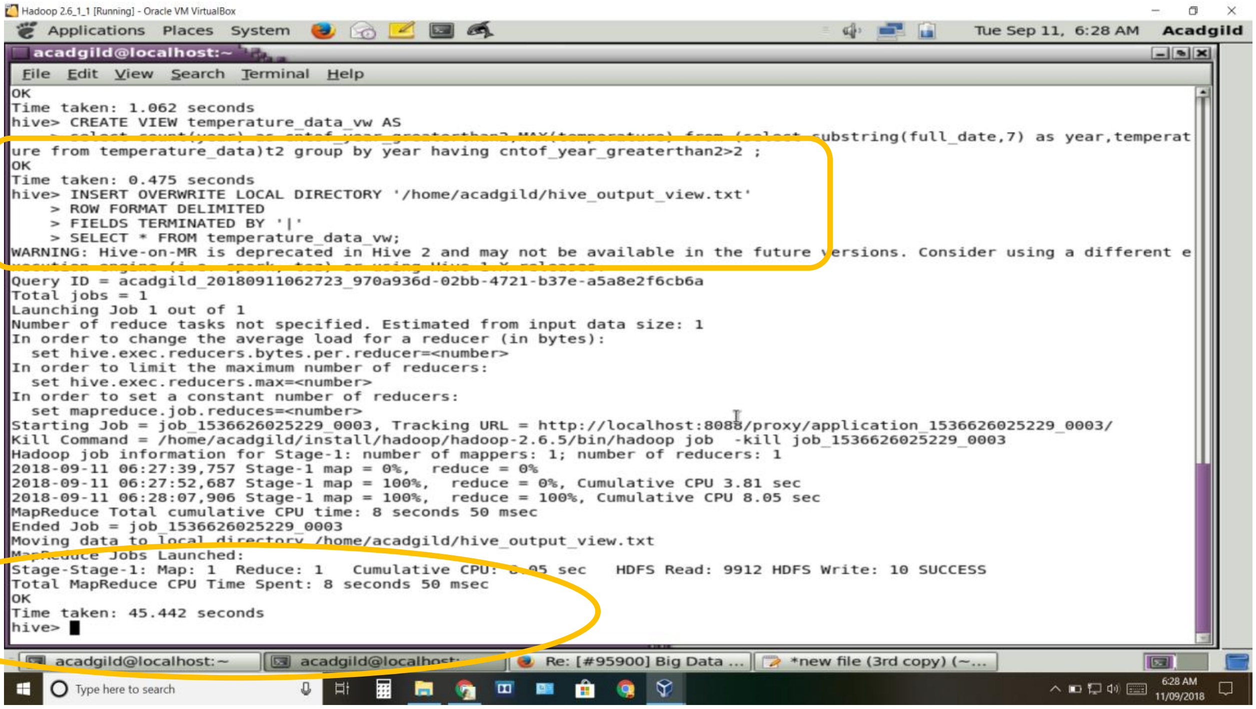
- Create a view on the top of last query, name it temperature_data_vw.



Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Tue Sep 11, 6:18 AM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
[acadgild@localhost:~]
acadgild@localhost:~
Hive - View and Index...
*new file (3rd copy) (~...
Type here to search
6:18 AM
11/09/2018

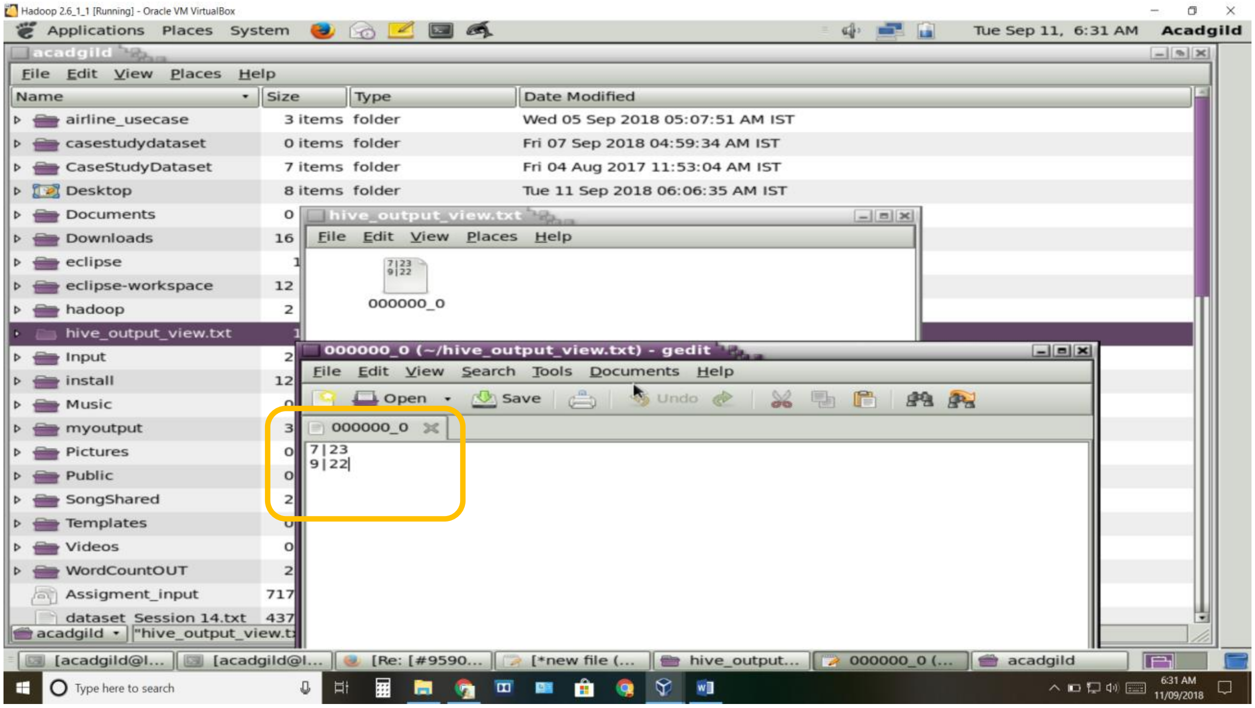
```
xecution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180911061219_68efd4bd-b445-4261-a907-46aff330b14e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in 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_1536626025229_0002, Tracking URL = http://localhost:8088/proxy/application_1536626025229_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1536626025229_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-11 06:12:39,477 Stage-1 map = 0%,  reduce = 0%
2018-09-11 06:12:59,366 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.89 sec
2018-09-11 06:13:16,162 Stage-1 map = 100%,  reduce = 83%, Cumulative CPU 10.1 sec
2018-09-11 06:13:17,331 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.61 sec
MapReduce Total cumulative CPU time: 10 seconds 610 msec
Ended Job = job_1536626025229_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.61 sec   HDFS Read: 10205 HDFS Write: 121 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 610 msec
OK
7       23
9       22
Time taken: 58.782 seconds, Fetched: 2 row(s)
hive> CREATE VIEW max_two AS
    > select count(year) as cntof_year_greaterthan2,MAX(temperature) from (select substring(full_date,7) as year,temperat
ure from temperature_data)t2 group by year having cntof_year_greaterthan2>2 ;
OK
Time taken: 1.062 seconds
hive> CREATE VIEW temperature_data_vw AS
    > select count(year) as cntof_year_greaterthan2,MAX(temperature) from (select substring(full_date,7) as year,temperat
ure from temperature_data)t2 group by year having cntof_year_greaterthan2>2 ;
OK
Time taken: 0.475 seconds
hive> 
```

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



```
OK
Time taken: 1.062 seconds
hive> CREATE VIEW temperature_data_vw AS
    > select count(year) as cntof_year_greaterthan2,MAX(temperature) from (select substring(full_date,7) as year,temperat
ure from temperature_data)t2 group by year having cntof_year_greaterthan2>2 ;
OK
Time taken: 0.475 seconds
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/hive_output_view.txt'
    > 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 e
xecution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180911062723_970a936d-02bb-4721-b37e-a5a8e2f6cb6a
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_1536626025229_0003, Tracking URL = http://localhost:8088/proxy/application_1536626025229_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1536626025229_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-09-11 06:27:39,757 Stage-1 map = 0%,  reduce = 0%
2018-09-11 06:27:52,687 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.81 sec
2018-09-11 06:28:07,906 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.05 sec
MapReduce Total cumulative CPU time: 8 seconds 50 msec
Ended Job = job_1536626025229_0003
Moving data to local directory /home/acadgild/hive_output_view.txt
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.05 sec   HDFS Read: 9912 HDFS Write: 10 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 50 msec
OK
Time taken: 45.442 seconds
hive> 
```

Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
Applications Places System
Tue Sep 11, 6:28 AM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Help
acadgild@localhost:~
Re: [#95900] Big Data ...
*new file (3rd copy) (~...
Type here to search
6:28 AM
11/09/2018



```
Name                      Size      Type     Date Modified
airline_usecase           3 items   folder   Wed 05 Sep 2018 05:07:51 AM IST
casestudydataset          0 items   folder   Fri 07 Sep 2018 04:59:34 AM IST
CaseStudyDataset          7 items   folder   Fri 04 Aug 2017 11:53:04 AM IST
Desktop                   8 items   folder   Tue 11 Sep 2018 06:06:35 AM IST
Documents                 0
Downloads                 16
eclipse                   1
eclipse-workspace         12
hadoop                    2
hive_output_view.txt      1
Input                     2
install                   12
Music                     0
myoutput                  3
Pictures                  0
Public                    0
SongShared                2
Templates                 0
Videos                    0
WordCountOUT              2
Assigment_input           717
dataset_Session 14.txt    437
```

000000_0 (~/hive_output_view.txt) - gedit

```
7|23
9|22
```