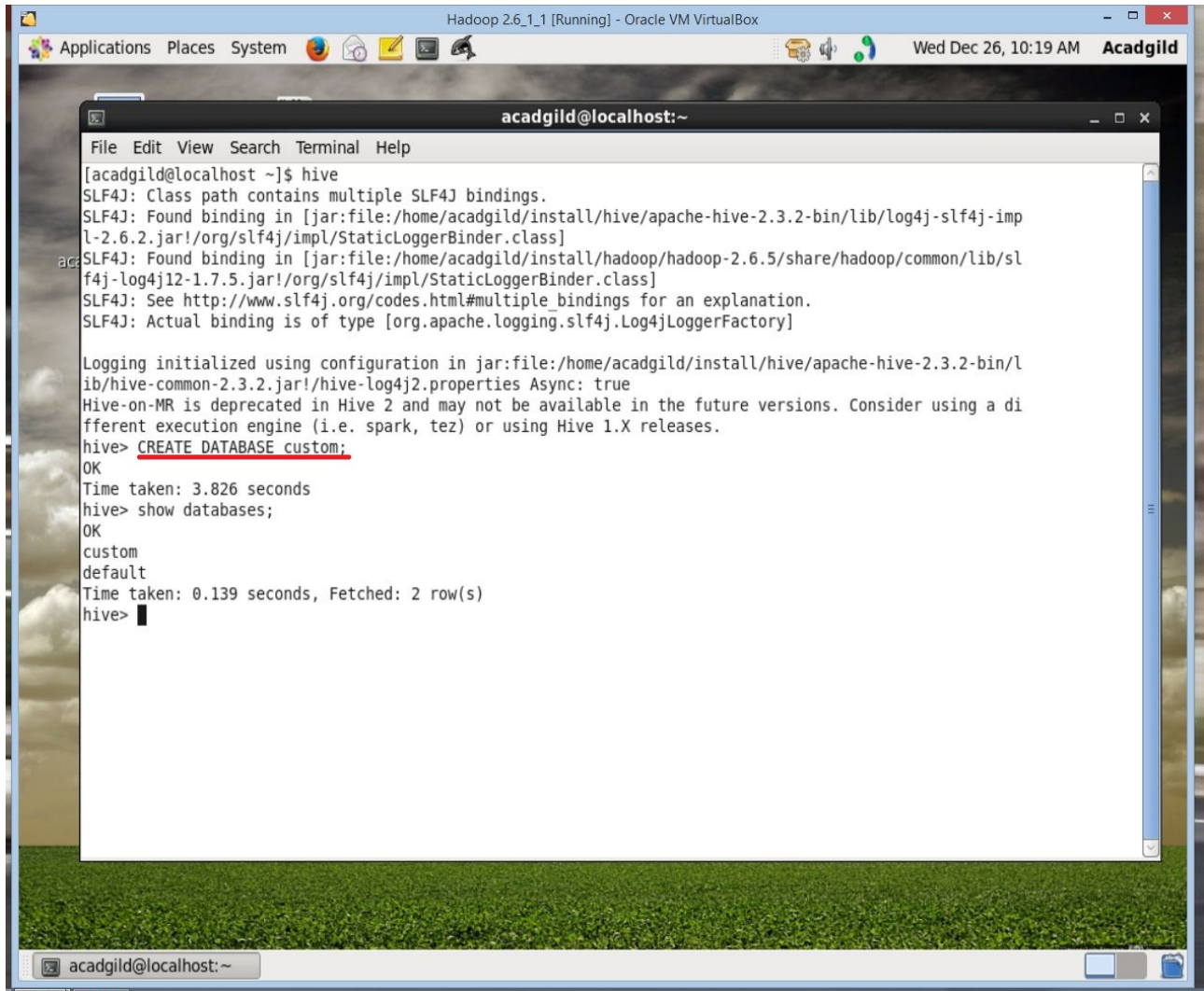


Assignment 1

TASK 1

1) Database is created and displayed

CREATE DATABASE custom;

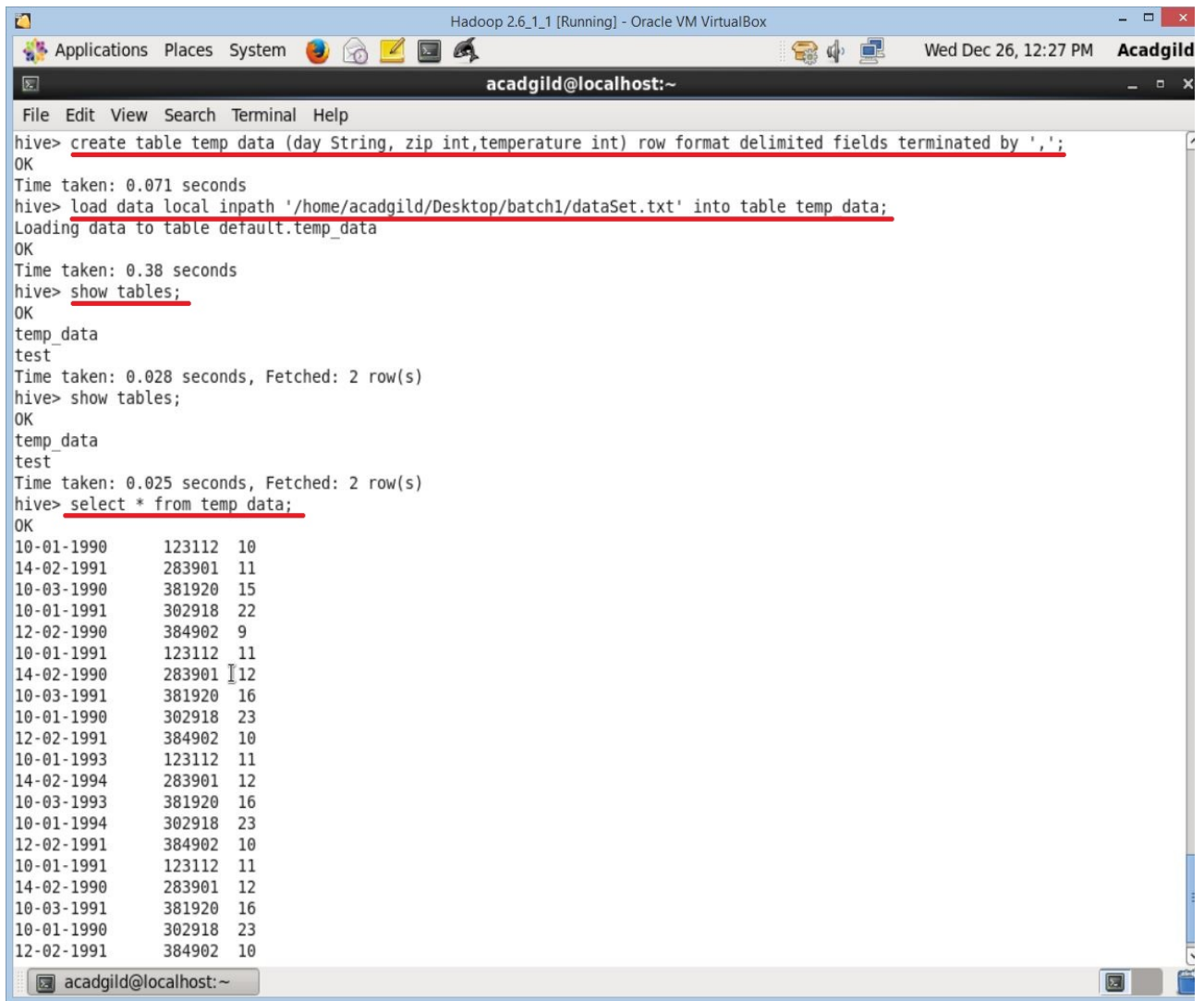


The screenshot shows a terminal window titled 'acadgild@localhost:~' within an Oracle VM VirtualBox environment. The terminal displays the execution of the 'hive' command, which outputs SLF4J logging information and initializes the Hive environment. Subsequently, the user enters 'hive> CREATE DATABASE custom;', followed by 'hive> show databases;', which lists 'custom' and 'default' databases. The terminal output includes the following text:

```
[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/log4j-slf4j-imp
l-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/sl
f4j-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/l
ib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a di
fferent execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> CREATE DATABASE custom;
OK
Time taken: 3.826 seconds
hive> show databases;
OK
custom
default
Time taken: 0.139 seconds, Fetched: 2 row(s)
hive> 
```

Table is created and the data is loaded from the local file system



The screenshot shows a terminal window titled "Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox" with the user "Acadgild" on "Wed Dec 26, 12:27 PM". The terminal prompt is "acadgild@localhost:~". The user enters several Hive commands, which are underlined in the original image. The output shows the creation of a table, loading of data from a local file, and a query result displaying a list of dates and associated numerical values.

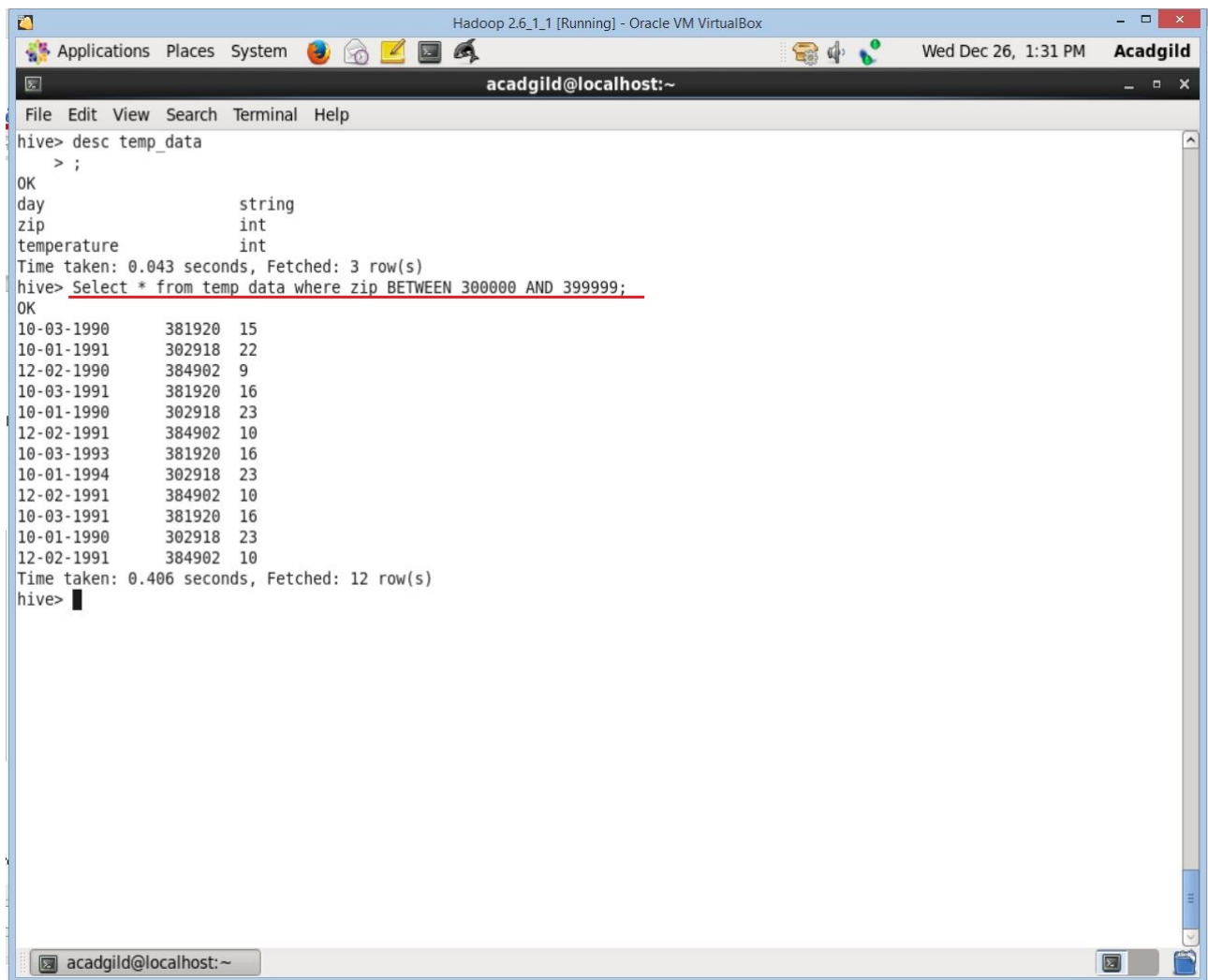
```
File Edit View Search Terminal Help
hive> create table temp_data (day String, zip int, temperature int) row format delimited fields terminated by ',';
OK
Time taken: 0.071 seconds
hive> load data local inpath '/home/acadgild/Desktop/batch1/dataSet.txt' into table temp_data;
Loading data to table default.temp_data
OK
Time taken: 0.38 seconds
hive> show tables;
OK
temp_data
test
Time taken: 0.028 seconds, Fetched: 2 row(s)
hive> show tables;
OK
temp_data
test
Time taken: 0.025 seconds, Fetched: 2 row(s)
hive> select * from temp_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
```

TASK 1

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

HIVE Command:

Select * From temp_data where zip BETWEEN 300000 AND 399999;



The screenshot shows a terminal window titled 'Hadoop 2.6_1_1 [Running] - Oracle VM VirtualBox' with the user 'Acadgild'. The terminal displays the following Hive commands and their outputs:

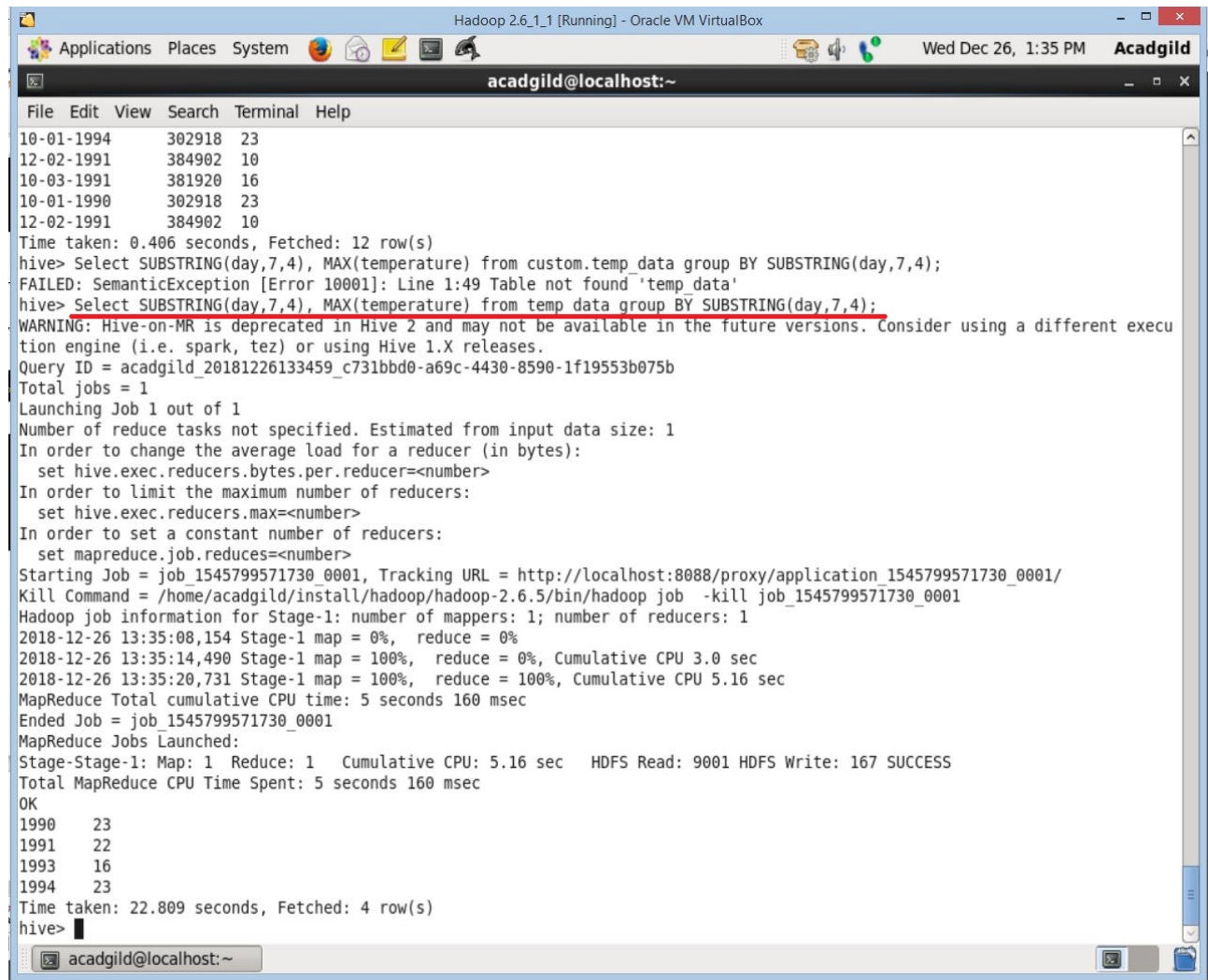
```
hive> desc temp_data
> ;
OK
day                string
zip                int
temperature        int
Time taken: 0.043 seconds, Fetched: 3 row(s)
hive> Select * from temp_data where zip BETWEEN 300000 AND 399999;
OK
10-03-1990      381920  15
10-01-1991      302918  22
12-02-1990      384902   9
10-03-1991      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: 0.406 seconds, Fetched: 12 row(s)
hive> █
```

The terminal window also shows a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The status bar at the bottom indicates the user is 'acadgild@localhost:~'.

b) maximum temperature corresponding to every year from temp_data table is

HIVE Command:

```
SELECT SUBSTRING(day,7,4), MAX(temperature) FROM temperature_data GROUP BY SUBSTRING(day,7,4);
```

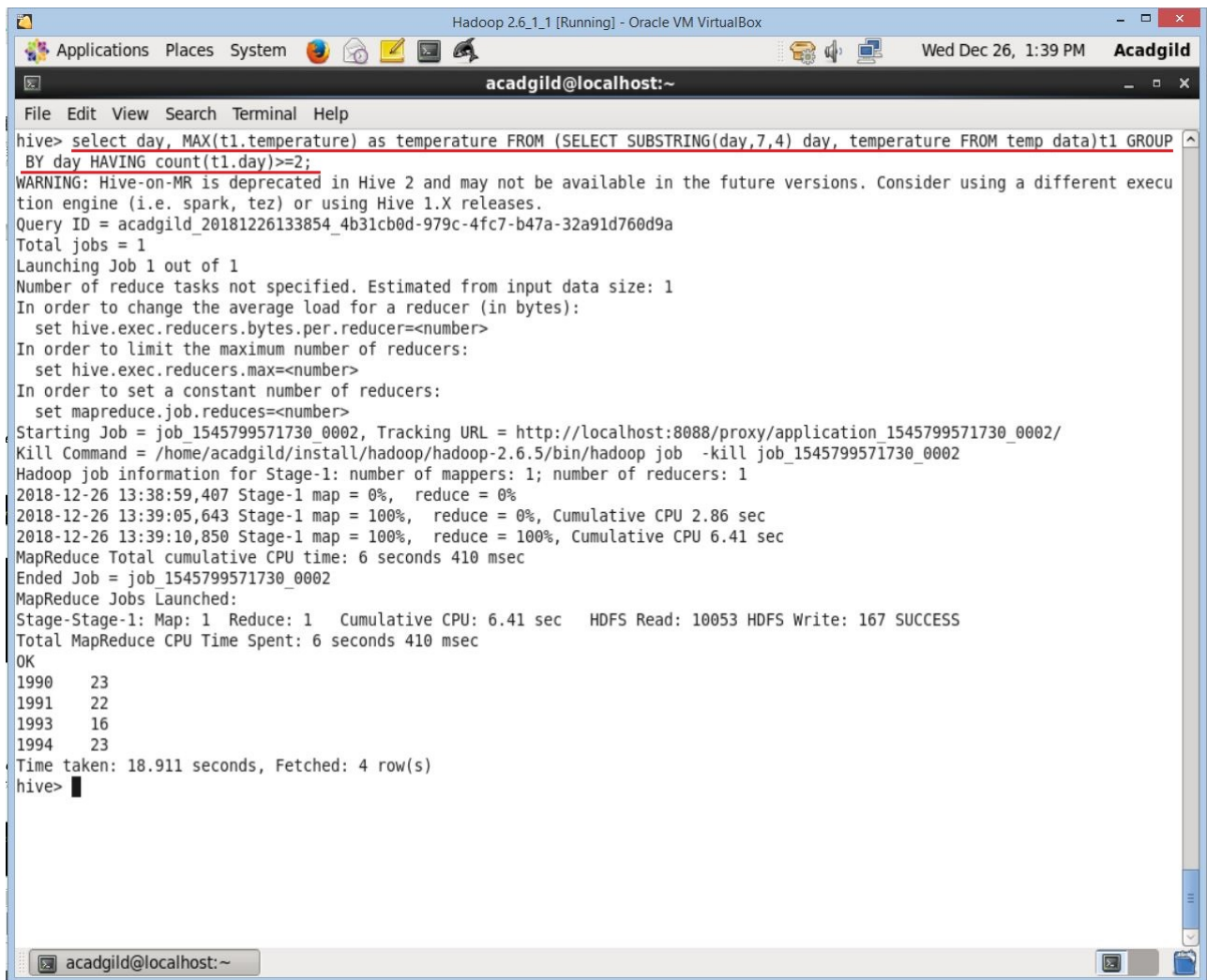


```
Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
acadgild@localhost:~
File Edit View Search Terminal Help
10-01-1994      302918  23
12-02-1991      384902  10
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
Time taken: 0.406 seconds, Fetched: 12 row(s)
hive> Select SUBSTRING(day,7,4), MAX(temperature) from custom.temp_data group BY SUBSTRING(day,7,4);
FAILED: SemanticException [Error 10001]: Line 1:49 Table not found 'temp_data'
hive> Select SUBSTRING(day,7,4), MAX(temperature) from temp_data group BY SUBSTRING(day,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, tez) or using Hive 1.X releases.
Query ID = acadgild_20181226133459_c731bbd0-a69c-4430-8590-1f19553b075b
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_1545799571730_0001, Tracking URL = http://localhost:8088/proxy/application_1545799571730_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1545799571730_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-26 13:35:08,154 Stage-1 map = 0%, reduce = 0%
2018-12-26 13:35:14,490 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.0 sec
2018-12-26 13:35:20,731 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.16 sec
MapReduce Total cumulative CPU time: 5 seconds 160 msec
Ended Job = job_1545799571730_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.16 sec HDFS Read: 9001 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 160 msec
OK
1990      23
1991      22
1993      16
1994      23
Time taken: 22.809 seconds, Fetched: 4 row(s)
hive>
```

c) Maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.

HIVE Command:

```
SELECT day, MAX(t1.temperature) as temperature FROM (SELECT  
SUBSTRING(day,7,4) day, temperature FROM temp_data)t1 GROUP BY full_date  
HAVING count(t1.day)>=2;
```

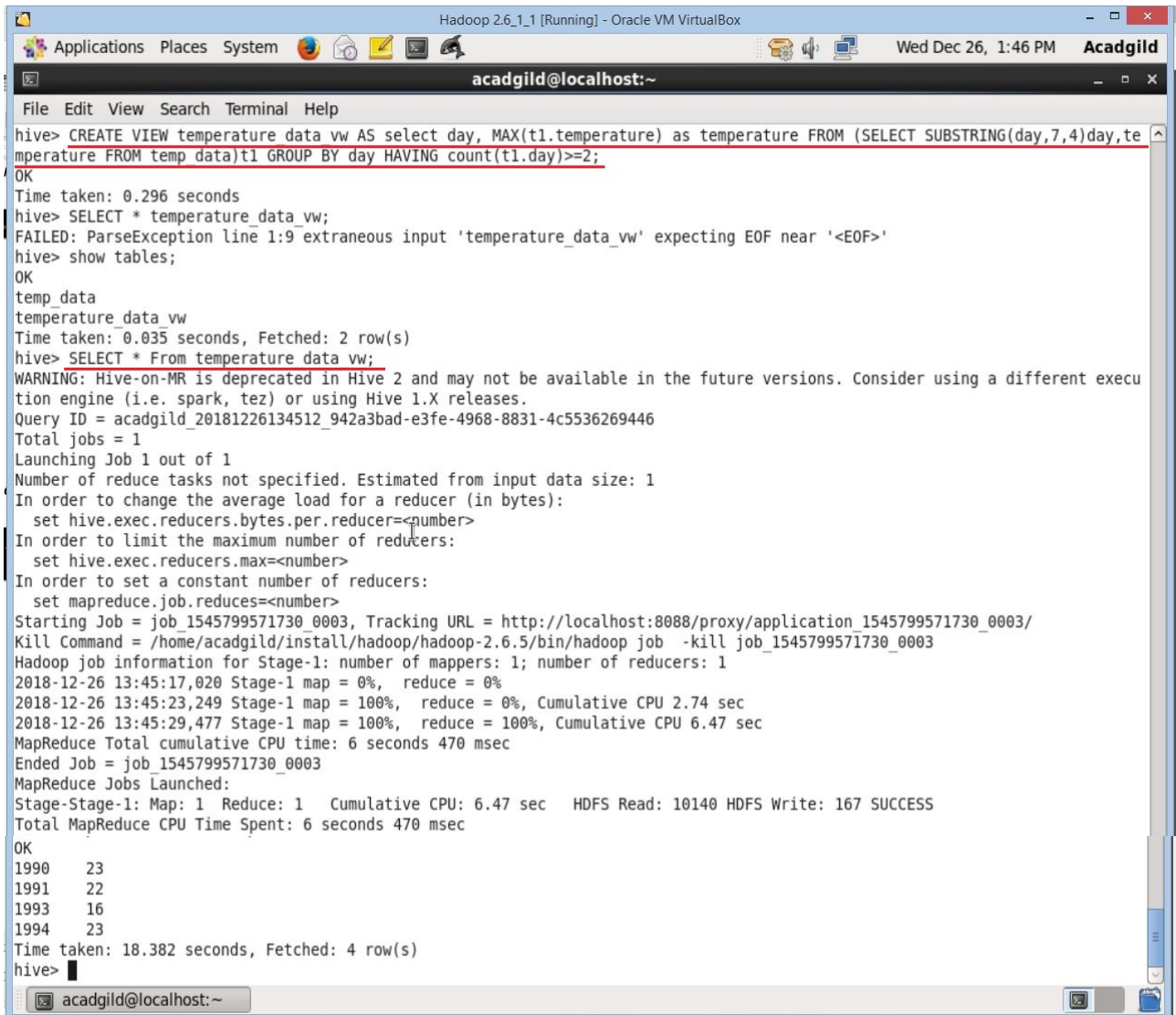


```
Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadgild@localhost:~
File Edit View Search Terminal Help
hive> select day, MAX(t1.temperature) as temperature FROM (SELECT SUBSTRING(day,7,4) day, temperature FROM temp_data)t1 GROUP
BY day HAVING count(t1.day)>=2;
WARNING: 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_20181226133854_4b31cb0d-979c-4fc7-b47a-32a91d760d9a
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_1545799571730_0002, Tracking URL = http://localhost:8088/proxy/application_1545799571730_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1545799571730_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-26 13:38:59,407 Stage-1 map = 0%, reduce = 0%
2018-12-26 13:39:05,643 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.86 sec
2018-12-26 13:39:10,850 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.41 sec
MapReduce Total cumulative CPU time: 6 seconds 410 msec
Ended Job = job_1545799571730_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.41 sec HDFS Read: 10053 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 410 msec
OK
1990      23
1991      22
1993      16
1994      23
Time taken: 18.911 seconds, Fetched: 4 row(s)
hive>
```


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

HIVE Command:

```
CREATE VIEW temperature_data_vw AS SELECT day, MAX(t1.temperature) as
temperature FROM (SELECT SUBSTRING(day,7,4) day, temperature FROM temp_data)t1
GROUP BY day HAVING count(t1.day)>=2;
```



The screenshot shows a terminal window titled 'Hadoop 2.6.1.1 [Running] - Oracle VM VirtualBox' with the user 'Acadgild'. The terminal prompt is 'acadgild@localhost:~'. The user enters the following Hive commands:

```
hive> CREATE VIEW temperature_data_vw AS select day, MAX(t1.temperature) as temperature FROM (SELECT SUBSTRING(day,7,4)day,te
mperature FROM temp_data)t1 GROUP BY day HAVING count(t1.day)>=2;
```

The output shows the command was successful, with a time taken of 0.296 seconds. The user then enters:

```
hive> SELECT * temperature_data_vw;
```

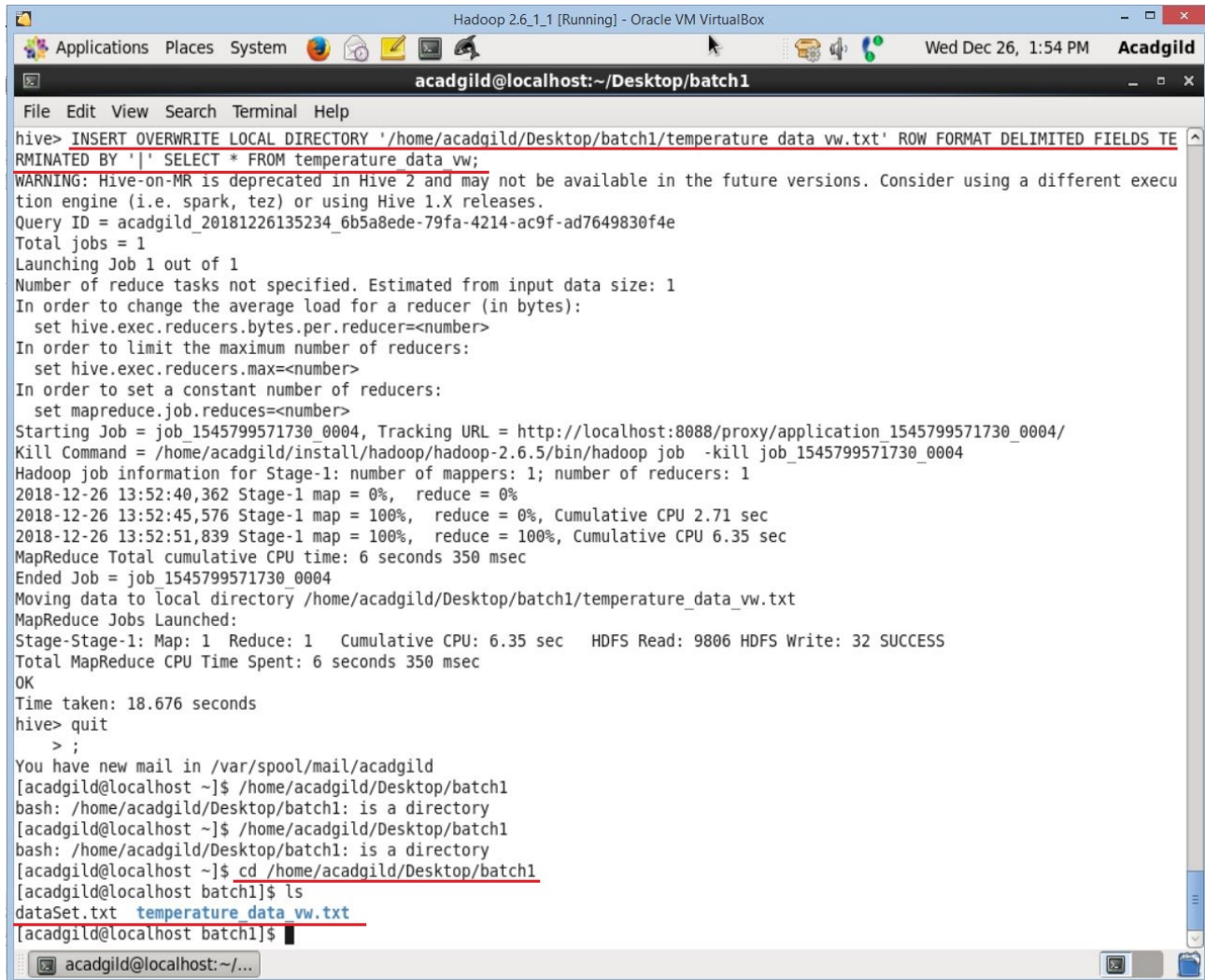
The output shows a warning about Hive-on-MR being deprecated, followed by job information and a successful result set:

```
Time taken: 0.035 seconds, Fetched: 2 row(s)
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_20181226134512_942a3bad-e3fe-4968-8831-4c5536269446
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_1545799571730_0003, Tracking URL = http://localhost:8088/proxy/application_1545799571730_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1545799571730_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-26 13:45:17,020 Stage-1 map = 0%, reduce = 0%
2018-12-26 13:45:23,249 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.74 sec
2018-12-26 13:45:29,477 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.47 sec
MapReduce Total cumulative CPU time: 6 seconds 470 msec
Ended Job = job_1545799571730_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.47 sec HDFS Read: 10140 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 470 msec
OK
1990 23
1991 22
1993 16
1994 23
Time taken: 18.382 seconds, Fetched: 4 row(s)
hive>
```

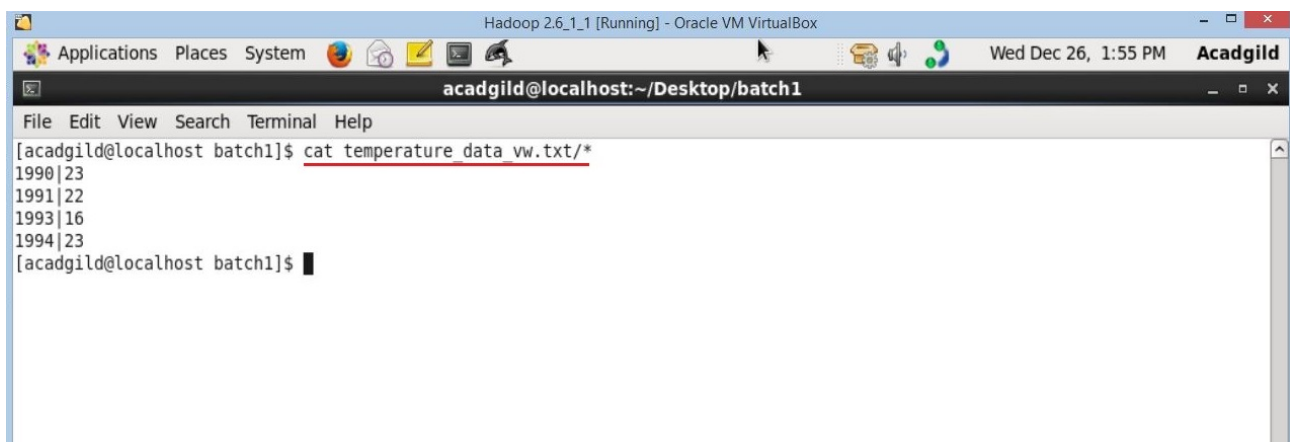
e) Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.
HIVE Command:

INSERT OVERWRITE LOCAL DIRECTORY

'/home/acadgild/Desktop/batch1/temperature_data_vw.txt' ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|' SELECT * FROM temperature_data_vw;



```
Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadmild@localhost:~/Desktop/batch1
File Edit View Search Terminal Help
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/Desktop/batch1/temperature data vw.txt' ROW FORMAT DELIMITED FIELDS TE
RMINATED 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_20181226135234_6b5a8ede-79fa-4214-ac9f-ad7649830f4e
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_1545799571730_0004, Tracking URL = http://localhost:8088/proxy/application 1545799571730_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1545799571730_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-26 13:52:40,362 Stage-1 map = 0%, reduce = 0%
2018-12-26 13:52:45,576 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.71 sec
2018-12-26 13:52:51,839 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.35 sec
MapReduce Total cumulative CPU time: 6 seconds 350 msec
Ended Job = job_1545799571730_0004
Moving data to local directory /home/acadgild/Desktop/batch1/temperature_data_vw.txt
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.35 sec HDFS Read: 9806 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 350 msec
OK
Time taken: 18.676 seconds
hive> quit
> ;
You have new mail in /var/spool/mail/acadmild
[acadmild@localhost ~]$ /home/acadgild/Desktop/batch1
bash: /home/acadgild/Desktop/batch1: is a directory
[acadmild@localhost ~]$ /home/acadgild/Desktop/batch1
bash: /home/acadgild/Desktop/batch1: is a directory
[acadmild@localhost ~]$ cd /home/acadgild/Desktop/batch1
[acadmild@localhost batch1]$ ls
dataSet.txt temperature data vw.txt
[acadmild@localhost batch1]$
```



```
Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadmild@localhost:~/Desktop/batch1
File Edit View Search Terminal Help
[acadmild@localhost batch1]$ cat temperature_data vw.txt/*
1990|23
1991|22
1993|16
1994|23
[acadmild@localhost batch1]$
```