Assignment 8 : Hive Basic Assignment Problems (Updated with terminal execution)

Note: Due to non working of the Hive terminal in my VM I am just uploading the programs on git hub. The execution for the same will be uploaded soon.

Problem Statement

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.

Terminal Execution :

```
[acadgild@localhost ~]$ jps
3179 Jps
[acadgild@localhost ~]$ sudo service sshd start
[sudo] password for acadgild:
[acadgild@localhost ~]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
18/07/11 21:36:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to
/home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-namenode-localhost.localdomain.ou
t
localhost: starting datanode, logging to
/home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to
/home/acadgild/install/hadoop/hadoop-2.6.5/logs/hadoop-acadgild-secondarynamenode-localhost.local
domain.out
18/07/11 21:37:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your
platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to
```

```
/home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /home/acadgild/install/hadoop/hadoop-2.6.5/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
[acadgild@localhost ~]$ jps
3428 DataNode
3556 SecondaryNameNode
3766 ResourceManager
3334 NameNode
3864 NodeManager
3903 Jps

[acadgild@localhost ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.3-bin/lib/log4j-slf4j-impl-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/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-2.3.3-bin/lib/hive-common-2.3.3.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 different execution engine (i.e. spark, tez) or using Hive 1.X releases.
[acadgild@localhost ~]$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hive/apache-hive-2.3.3-bin/lib/log4j-slf4j-impl-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/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-2.3.3-bin/lib/hive-common-2.3.3.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 different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> show databases;
OK
default
Time taken: 28.494 seconds, Fetched: 1 row(s)
```

Problem 1: Create a database named 'custom'.

Commands :

1. show databases;
2. create database custom;
3. use custom;

Terminal Execution:

```
hive> show databases;
OK
default
Time taken: 28.494 seconds, Fetched: 1 row(s)

hive> create database custom;
OK
Time taken: 0.633 seconds

hive> use custom;
OK
Time taken: 0.042 seconds

hive> show databases;
OK
custom
default
Time taken: 0.055 seconds, Fetched: 2 row(s)
```

Problem 2: Create a table named temperature_data inside custom having below fields:
1. date (mm-dd-yyyy) format
2. zip code
3. temperature

Commands:

```
Create table IF NOT EXISTS  temperature_data
( date timestamp(MM-DD-YYYY),
   zip_code string,
   temperature int)
row format delimited
field teminated by ',' ;
```

Terminal Execution:

```
hive> create table if not exists temperature_data(
    > s_date string,
    > zip_code string,
    > temperature int)
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 0.281 seconds
hive> show create table temperature_data;
OK
CREATE TABLE `temperature_data`(
  `s_date` string,
  `zip_code` string,
  `temperature` int)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES (
  'field.delim'=',',
  'serialization.format'=',')
STORED AS INPUTFORMAT
  'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION
  'hdfs://localhost:8020/user/hive/warehouse/custom.db/temperature_data'
TBLPROPERTIES (
  'transient_lastDdlTime'='1531691365')
Time taken: 0.192 seconds, Fetched: 17 row(s)
hive>
```

Problem 3 : Load the dataset.txt (which is ',' delimited) in the table.

Commands:

load data local inpath '/home/acadgild/Desktop/dataset_Session 14.txt' into table temperature_data;

Terminal Execution:

```
hive> load data local inpath '/home/acadgild/Desktop/dataset_Session 14.txt' into table
temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 1.257 seconds



hive> select
to_date(from_unixtime(UNIX_TIMESTAMP(s_date,"MM-DD-YYYY"))),zip_code,temperature from
temperature_data;
OK
1989-12-31    123112        10
1990-12-30    283901        11
1989-12-31    381920        15
1990-12-30    302918        22
1989-12-31    384902        9
1990-12-30    123112        11
1989-12-31    283901        12
1990-12-30    381920        16
1989-12-31    302918        23
1990-12-30    384902        10
1992-12-27    123112        11
1993-12-26    283901        12
1992-12-27    381920        16
1993-12-26    302918        23
1990-12-30    384902        10
1990-12-30    123112        11
1989-12-31    283901        12
1990-12-30    381920        16
1989-12-31    302918        23
1990-12-30    384902        10
Time taken: 0.314 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.

Solution :

```
Select s_date, temperature
from  temperature_data
where (cast(zip_code as int)>300000) and (cast(zip_code as int) <399999);
```

Terminal Execution

```
hive> Select s_date, temperature
    > from  temperature_data
    > where (cast(zip_code as int)>300000) and (cast(zip_code as int) <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: 0.432 seconds, Fetched: 12 row(s)
```

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

Command:

```
select year, MAX(t1.temperature) as temperature
from (select substring(s_date,7,4) year, temperature from temperature_data)t1
group by year ;
```

Terminal Execution

```
hive> select year, MAX(t1.temperature) as temperature
    > from (select substring(s_date,7,4) year, temperature from temperature_data)t1
    > 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 execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180716065008_a4ea7d8c-d1a3-460d-9b6f-fd3789b7ca02
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_1531673901515_0004, Tracking URL =
http://localhost:8088/proxy/application_1531673901515_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill
job_1531673901515_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-07-16 06:50:30,932 Stage-1 map = 0%,  reduce = 0%
2018-07-16 06:50:46,854 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.08 sec
2018-07-16 06:51:05,607 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.76 sec
MapReduce Total cumulative CPU time: 5 seconds 760 msec
Ended Job = job_1531673901515_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.76 sec   HDFS Read: 9233 HDFS Write: 167
SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 760 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 57.991 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.

Command:

```
select year, MAX(t1.temperature) as temperature
from (select substring(s_date,7,4) year, temperature from temperature_data)t1
group by year
having count(t1.year)>2;
```

Terminal Execution

```
hive> select year, MAX(t1.temperature) as temperature
    > from (select substring(s_date,7,4) year, temperature from temperature_data)t1
    > group by year
    > having count(t1.year)>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 releases.
Query ID = acadgild_20180716064421_83a6ec05-9c67-41e2-9bce-eacdb66e1a8a
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_1531673901515_0003, Tracking URL =
http://localhost:8088/proxy/application_1531673901515_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill
job_1531673901515_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-07-16 06:44:54,603 Stage-1 map = 0%,  reduce = 0%
2018-07-16 06:45:19,755 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.74 sec
2018-07-16 06:45:41,383 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 6.69 sec
2018-07-16 06:45:44,344 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.93 sec
MapReduce Total cumulative CPU time: 7 seconds 930 msec
Ended Job = job_1531673901515_0003
```

MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 7.93 sec   HDFS Read: 10238 HDFS Write: 127 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 930 msec
OK
1990   23
1991   22
Time taken: 84.264 seconds, Fetched: 2 row(s)

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

Solution :

Create View temperature_data_vw **AS** select year, MAX(t1.temperature) as temperature
from (select substring(s_date,7,4) year, temperature from temperature_data)t1
group by year
having count(t1.year)>2;

Terminal Execution

hive> Create View temperature_data_vw AS select year, MAX(t1.temperature) as temperature
    > from (select substring(s_date,7,4) year, temperature from temperature_data)t1
    > group by year
    > having count(t1.year)>2;
OK
Time taken: 0.93 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 releases.
Query ID = acadgild_20180716070525_f60591bd-22dc-4c27-bf8b-9526c6119501
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_1531673901515_0005, Tracking URL =
http://localhost:8088/proxy/application_1531673901515_0005/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill
job_1531673901515_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-07-16 07:05:42,254 Stage-1 map = 0%, reduce = 0%

```
2018-07-16 07:05:56,082 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.1 sec
2018-07-16 07:06:13,536 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.92 sec
MapReduce Total cumulative CPU time: 6 seconds 920 msec
Ended Job = job_1531673901515_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.92 sec   HDFS Read: 10311 HDFS Write: 127
SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 920 msec
OK
1990    23
1991    22
Time taken: 50.704 seconds, Fetched: 2 row(s)
```

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

Solution :

```
[acadgild@localhost ~]$ pwd
/home/acadgild
[acadgild@localhost ~]$ ls /home/acadgild/Desktop
Assignment_7.11199.odt  dataset_Session 14.txt  problem3.pig~  query4.pig~
Assignment_8.11199.odt  PIG                     problem4.pig~  query5.pig~
Assignment Done         problem1.pig            query1.pig~    README
Assignment_Jars         problem1.pig~           query2.pig~    sample.txt~
Datasets                problem2.pig~           query3.pig~    word_count.pig~
[acadgild@localhost ~]$ mkdir /home/acadgild/Desktop/hive_local
[acadgild@localhost ~]$ ls /home/acadgild/Desktop/
Assignment_7.11199.odt  hive_local     problem4.pig~  README
Assignment_8.11199.odt  PIG            query1.pig~    sample.txt~
Assignment Done         problem1.pig   query2.pig~    word_count.pig~
Assignment_Jars         problem1.pig~  query3.pig~
Datasets                problem2.pig~  query4.pig~
dataset_Session 14.txt  problem3.pig~  query5.pig~

hive> insert overwrite local directory '/home/acadgild/Desktop/hive_local'
    > 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_20180716075622_a922221f-8df7-42b9-b5e4-1b39a911c4ad
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_1531673901515_0007, Tracking URL =
http://localhost:8088/proxy/application_1531673901515_0007/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill
job_1531673901515_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-07-16 07:56:40,711 Stage-1 map = 0%,  reduce = 0%
2018-07-16 07:56:54,263 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.98 sec
2018-07-16 07:57:10,299 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.48 sec
MapReduce Total cumulative CPU time: 6 seconds 480 msec
Ended Job = job_1531673901515_0007
Moving data to local directory /home/acadgild/Desktop/hive_local
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.48 sec   HDFS Read: 9948 HDFS Write: 16
SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 480 msec
OK
Time taken: 48.759 seconds
hive> exit;
[acadgild@localhost ~]$ ls /home/acadgild/Desktop/hive_local
000000_0
[acadgild@localhost ~]$ cat /home/acadgild/Desktop/hive_local/000000_0
1990|23
1991|22
[acadgild@localhost ~]$
```