Session 8_HIVE BASICS_Assignment1

Task 1

Create a database named 'custom'.

Solution:

hive> create database custom;

OK

Time taken: 2.356 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.

Solution:

hive> use custom;

OK

Time taken: 0.114 seconds

hive> CREATE TABLE temperature_data (TempDate STRING, ZIP_Code INT, Temperature INT) row format delimited fields terminated by ',';

OK

Time taken: 4.001 seconds

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

Solution:

Dataset.txt created at below location and put downloaded data in it.

[acadgild@localhost ~]$ nano dataset.txt

/home/acadgild/ dataset.txt

Executed below command to create table:

```
hive> LOAD DATA LOCAL INPATH '/home/acadgild/dataset.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 12.012 seconds
hive>
```

Validated data loaded in table:

```
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: 12.446 seconds, Fetched: 20 row(s)
```

Task 2

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

Solution:

Below command executed:

```
hive> select * from temperature_data where ZIP_Code 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: 5.022 seconds, Fetched: 12 row(s)
hive>
```

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

Solution:

Below command executed

```
hive> select  max(Temperature),substring(TempDate,7,4)as Year  from temperature_data group by substring(TempDate,7,4) ;
```

Output:

Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 23.02 sec   HDFS Read: 9400 HDFS Write: 167 SUCCESS

Total MapReduce CPU Time Spent: 23 seconds 20 msec

OK

Result:

23 1990

22 1991

16 1993

23 1994

Time taken: 361.349 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.

Solution:

Below command executed:

hive> select substring(TempDate,7,4)as Year,max(Temperature),count(substring(TempDate,7,4)) from temperature_data group by substring(TempDate,7,4) having count(substring(TempDate,7,4))>1;

Output:

Starting Job = job_1528961787366_0004, Tracking URL = http://localhost:8088/proxy/application_1528961787366_0004/

Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1528961787366_0004

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2018-06-14 17:04:34,003 Stage-1 map = 0%, reduce = 0%

2018-06-14 17:05:34,707 Stage-1 map = 0%, reduce = 0%

2018-06-14 17:05:52,721 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.23 sec

2018-06-14 17:06:53,934 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.23 sec

2018-06-14 17:07:00,428 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 14.67 sec

2018-06-14 17:07:28,309 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 26.35 sec

MapReduce Total cumulative CPU time: 26 seconds 350 msec

Ended Job = job_1528961787366_0004

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 26.35 sec   HDFS Read: 10294 HDFS Write: 175 SUCCESS

Total MapReduce CPU Time Spent: 26 seconds 350 msec

OK

Result:

1990 23 7

1991 22 9

1993 16 2

1994 23 2

Time taken: 276.146 seconds, Fetched: 4 row(s)

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

Solution:

Executed below command:

hive> Create VIEW temperature_data_vw as select substring(TempDate,7,4)as Year,max(Temperature),count(substring(TempDate,7,4)) from temperature_data group by substring(TempDate,7,4) having count(substring(TempDate,7,4))>1;

Result:

OK

Time taken: 3.685 seconds

Validated View

hive> select * from temperature_data_vw;

Total MapReduce CPU Time Spent: 32 seconds 800 msec

OK

1990 23 7

1991 22 9

1993 16 2

1994 23 2

Time taken: 433.622 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.

Solution:

Executed below command

hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/output' ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' SELECT * FROM temperature_data_vw;

Output:

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_20180614174514_8b378b88-852d-416c-aff3-8eab6e86de34

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_1528961787366_0006, Tracking URL = http://localhost:8088/proxy/application_1528961787366_0006/

Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1528961787366_0006

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2018-06-14 17:46:47,181 Stage-1 map = 0%, reduce = 0%

2018-06-14 17:47:47,972 Stage-1 map = 0%, reduce = 0%

2018-06-14 17:48:09,877 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 12.06 sec

2018-06-14 17:49:07,567 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 14.45 sec

2018-06-14 17:49:40,854 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 26.72 sec

MapReduce Total cumulative CPU time: 26 seconds 720 msec

Ended Job = job_1528961787366_0006

Moving data to local directory /home/acadgild/output

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 26.72 sec   HDFS Read: 10262 HDFS Write: 40 SUCCESS

Total MapReduce CPU Time Spent: 26 seconds 720 msec

OK

Time taken: 271.4 seconds

hive>

Validated by going to the output path

[acadgild@localhost output]$ ls

000000_0

[acadgild@localhost output]$ cat 000000_0

1990|23|7

1991|22|9

1993|16|2

1994|23|2

[acadgild@localhost output]$