BIG DATA
DEVELOPMENT

ACADGILD

Session 07: Hive Operations

Assignment 2

Downloaded dataset from given link and implemented task using acadgild blog

https://acadgild.com/blog/hive-real-life-use-cases/

https://drive.google.com/open?id=0ByJLBTmJojjzV1czX3Nha0R3bTQ

DATE SET DESCRIPTION

The data set consists of the following fields.

Athlete: This field consists of the athlete name

Age: This field consists of athlete ages

Country: This fields consists of the country names which participated in Olympics

Year: This field consists of the year

Closing Date: This field consists of the closing date of ceremony

Sport: Consists of the sports name

Gold Medals: No. of Gold medals

Silver Medals: No. of Silver medals

Bronze Medals: No. of Bronze medals

Total Medals: Consists of total no. of medals

```
hive> show databases;
OK
custom
default
Time taken: 0.485 seconds, Fetched: 2 row(s)
hive> use custom;
OK
Time taken: 0.037 seconds
hive> set hive.cli.print.current.db;
hive.cli.print.current.db=false
hive> set hive.cli.print.current.db=true;
```

hive (custom)>create table olympic (athelete STRING,age INT,country STRING,year STRING,closing STRING,sport STRING,gold INT,silver INT,bronze INT,total INT) row format delimited fields terminated by '\t' stored as textfile;

```
hive (custom)> create table olympic(athelete STRING,age INT,country STRING,year STRING,closing STRING,sport STRING,gold INT,silver INT,
bronze INT,total INT) row format delimited fields terminated by '\t' stored as textfile;
OK
Time taken: 0.467 seconds
```

```
hive (custom)> load data local inpath '/home/acadgild/Downloads/olympic_data.csv' into table olympic;
Loading data to table custom.olympic
Table custom.olympic stats: [numFiles=1, totalSize=518669]
OK
Time taken: 1.724 seconds
hive (custom)> select * from olympic LIMIT 3;
OK
Michael Phelps  23      United States   2008    08-24-08        Swimming        8       0       0       8
Michael Phelps  19      United States   2004    08-29-04        Swimming        6       0       2       8
Michael Phelps  27      United States   2012    08-12-12        Swimming        4       2       0       6
Time taken: 0.408 seconds, Fetched: 3 row(s)
hive (custom)>
```

hive (custom)> load data local inpath '/home/acadgild/Downloads/olympic_data.csv' into table olympic;

Loading data to table custom.olympic

Table custom.olympic stats: [numFiles=1, totalSize=518669]

OK

Time taken: 1.724 seconds

hive (custom)> select * from olympic LIMIT 3;

OK

Michael Phelps	23	United States	2008	08-24-08	Swimming	8	0	0	8
Michael Phelps	19	United States	2004	08-29-04	Swimming	6	0	2	8
Michael Phelps	27	United States	2012	08-12-12	Swimming	4	2	0	6

Time taken: 0.408 seconds, Fetched: 3 row(s)

1. Problem Statement

1. *Write a Hive program to find the number of medals won by each country in swimming.*

hive (custom)> select country,SUM(total) from olympic where sport = 'Swimming' GROUP BY country;

```
hive (custom)> select country,SUM(total) from olympic where sport = 'Swimming' GROUP BY country;
Query ID = acadgild_20171015101212_3b2cfad7-fc99-4947-a7fc-6d1da84446be
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_1508040195270_0001, Tracking URL = http://localhost:8088/proxy/application_1508040195270_0001/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1508040195270_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-15 10:12:42,385 Stage-1 map = 0%,  reduce = 0%
2017-10-15 10:12:48,997 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.16 sec
2017-10-15 10:12:56,615 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.27 sec
MapReduce Total cumulative CPU time: 2 seconds 270 msec
Ended Job = job_1508040195270_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.27 sec   HDFS Read: 518899 HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 270 msec
OK
```

Output :-

Argentina 1

Australia 163

Austria 3

Belarus 2

Brazil 8

Canada 5

China 35

Costa Rica 2

Croatia 1

Denmark 1

France 39

Germany 32

Great Britain 11

Hungary 9

Italy 16

Japan 43

Lithuania 1

Netherlands 46

Norway 2

Poland 3

Romania 6

Russia 20

Serbia 1

Slovakia 2

Slovenia 1

South Africa 11

South Korea 4

Spain 3

Sweden 9

Trinidad and Tobago 1

Tunisia 3

Ukraine 7

United States 267

Zimbabwe 7

Time taken: 28.097 seconds, Fetched: 34 row(s)

```
OK
Argentina       1
Australia       163
Austria 3
Belarus 2
Brazil  8
Canada  5
China   35
Costa Rica      2
Croatia 1
Denmark 1
France  39
Germany 32
Great Britain   11
Hungary 9
Italy   16
Japan   43
Lithuania       1
Netherlands     46
Norway  2
Poland  3
Romania 6
Russia  20
Serbia  1
Slovakia        2
Slovenia        1
South Africa    11
South Korea     4
Spain   3
Sweden  9
Trinidad and Tobago     1
Tunisia 3
Ukraine 7
United States   267
Zimbabwe        7
Time taken: 28.097 seconds, Fetched: 34 row(s)
hive (custom)>
```

2. *Write a Hive program to find the number of medals that India won year wise.*

```
hive (custom)> select year,SUM(total) from olympic where country = 'India' GROUP BY year;
```

```
hive (custom)> select year,SUM(total) from olympic where country = 'India' GROUP BY year;
Query ID = acadgild_20171015101515_8d29f7d8-f12b-444d-abad-74c4c020f179
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_1508040195270_0002, Tracking URL = http://localhost:8088/proxy/application_1508040195270_0002/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1508040195270_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-15 10:16:05,452 Stage-1 map = 0%,  reduce = 0%
2017-10-15 10:16:13,176 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.83 sec
2017-10-15 10:16:20,720 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.83 sec
MapReduce Total cumulative CPU time: 2 seconds 830 msec
Ended Job = job_1508040195270_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.83 sec   HDFS Read: 518899 HDFS Write: 28 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 830 msec
OK
2000    1
2004    1
2008    3
2012    6
Time taken: 23.529 seconds, Fetched: 4 row(s)
hive (custom)>
```

Output:-

2000 1

2004 1

2008 3

2012 6

3. *Write a Hive Program to find the total number of medals each country won.*

hive (custom)> select country,SUM(total) from olympic GROUP BY country;

```
hive (custom)> select country,SUM(total) from olympic GROUP BY country;
Query ID = acadgild_20171015101717_fc38c67a-6449-4ce9-849d-8370107e8f72
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>
```

```
Total MapReduce CPU Time Spent: 2 seconds 40 msec
OK
Afghanistan     2
Algeria 8
Argentina       141
Armenia 10
Australia       609
Austria 91
Azerbaijan      25
Bahamas 24
Bahrain 1
Barbados        1
Belarus 97
Belgium 18
Botswana        1
Brazil  221
Bulgaria        41
Cameroon        20
Canada  370
Chile   22
China   530
Chinese Taipei  20
Colombia        13
Costa Rica      2
```

Output:-

Afghanistan 2

Algeria 8

Argentina 141

Armenia 10

Australia 609

Austria 91

Azerbaijan 25

Bahamas 24

Bahrain 1

Barbados 1

Belarus 97

Belgium 18

Botswana 1

Brazil 221

Bulgaria 41

Cameroon 20

Canada 370

Chile 22

China 530

Chinese Taipei 20

Colombia 13

Costa Rica 2

Croatia 81

Cuba 188

Cyprus 1

Czech Republic 81

Denmark 89

Dominican Republic 5

Ecuador 1

Egypt 8

Eritrea 1

Estonia 18

Ethiopia 29

Finland 118

France 318

Gabon 1

Georgia 23

Germany 629

Great Britain 322

Greece 59

Grenada 1

Guatemala 1

Hong Kong 3

Hungary 145

Iceland 15

India 11

Indonesia 22

Iran 24

Ireland 9

Israel 4

Italy 331

Jamaica 80

Japan 282

Kazakhstan 42

Kenya 39

Kuwait 2

Kyrgyzstan 3

Latvia 17

Lithuania 30

Macedonia 1

Malaysia 3

Mauritius 1

Mexico 38

Moldova 5

Mongolia 10

Montenegro 14

Morocco 11

Mozambique 1

Netherlands 318

New Zealand 52

Nigeria 39

North Korea 21

Norway 192

Panama 1

Paraguay 17

Poland 80

Portugal 9

Puerto Rico 2

Qatar 3

Romania 123

Russia 768

Saudi Arabia 6

Serbia 31

Serbia and Montenegro 38

Singapore 7

Slovakia 35

Slovenia 25

South Africa 25

South Korea 308

Spain 205

Sri Lanka 1

Sudan 1

Sweden 181

Switzerland 93

Syria 1

Tajikistan 3

Thailand 18

Togo 1

Trinidad and Tobago 19

Tunisia 4

Turkey 28

Uganda 1

Ukraine 143

United Arab Emirates 1

United States 1312

Uruguay 1

Uzbekistan 19

Venezuela 4

Vietnam 2

Zimbabwe 7

Time taken: 21.1 seconds, Fetched: 110 row(s)

4. *Write a Hive program to find the number of gold medals each country won.*

```
hive (custom)> select country,SUM(gold) from olympic GROUP BY country;
```

```
hive (custom)> select country,SUM(gold) from olympic GROUP BY country;
Query ID = acadgild_20171015102020_11401cee-e97c-4582-9b6b-a27b1b6e3951
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_1508040195270_0004, Tracking URL = http://localhost:8088/proxy/application_1508040195270_0004/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1508040195270_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-15 10:20:51,897 Stage-1 map = 0%,  reduce = 0%
2017-10-15 10:20:58,292 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 0.91 sec
2017-10-15 10:21:04,840 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.11 sec
MapReduce Total cumulative CPU time: 2 seconds 110 msec
Ended Job = job_1508040195270_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.11 sec   HDFS Read: 518899 HDFS Write: 1276 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 110 msec
OK
Afghanistan     0
Algeria 2
Argentina       49
Armenia 0
```

Output:-

Afghanistan 0

Algeria 2

Argentina 49

Armenia 0

Australia 163

Austria 36

Azerbaijan 6

Bahamas 11

Bahrain 0

Barbados 0

Belarus 17

Belgium 2

Botswana 0

Brazil 46

Bulgaria 8

Cameroon 20

Canada 168

Chile 3

China 234

Chinese Taipei 2

Colombia 2

Costa Rica 0

Croatia 35

Cuba 57

Cyprus 0

Czech Republic 14

Denmark 46

Dominican Republic 3

Ecuador 0

Egypt 1

Eritrea 0

Estonia 6

Ethiopia 13

Finland 11

France 108

Gabon 0

Georgia 6

Germany 223

Great Britain 124

Greece 12

Grenada 1

Guatemala 0

Hong Kong 0

Hungary 77

Iceland 0

India 1

Indonesia 5

Iran 10

Ireland 1

Israel 1

Italy 86

Jamaica 24

Japan 57

Kazakhstan 13

Kenya 11

Kuwait 0

Kyrgyzstan 0

Latvia 3

Lithuania 5

Macedonia 0

Malaysia 0

Mauritius 0

Mexico 19

Moldova 0

Mongolia 2

Montenegro 0

Morocco 2

Mozambique 1

Netherlands 101

New Zealand 18

Nigeria 6

North Korea 6

Norway 97

Panama 1

Paraguay 0

Poland 20

Portugal 1

Puerto Rico 0

Qatar 0

Romania 57

Russia 234

Saudi Arabia 0

Serbia 1

Serbia and Montenegro 11

Singapore 0

Slovakia 10

Slovenia 5

South Africa 10

South Korea 110

Spain 19

Sri Lanka 0

Sudan 0

Sweden 57

Switzerland 21

Syria 0

Tajikistan 0

Thailand 6

Togo 0

Trinidad and Tobago 1

Tunisia 2

Turkey 9

Uganda 1

Ukraine 31

United Arab Emirates 1

United States 552

Uruguay 0

Uzbekistan 5

Venezuela 1

Vietnam 0

Zimbabwe 2

Time taken: 21.874 seconds, Fetched: 110 row(s)