Hive Oprations Assignment 2 :

This Data set is about Olympics. This assignment uses olympics_data.csv file . The solutions for the each query are given below.First we have to create the table and load the data for Olympics.

Creating the table and loading the data

1) First, we have to create a table named olympics_data, with the fields corresponding to the data in the olympics_data.CSV data file. The command used is as below.

```
CREATE TABLE olympics_data
(
name string,
age int,
country string,
year int,
closing_date string,
sport string,
gold int,
silver int,
bronze int,
total_medals int
)
ROW FORMAT DELIMITED FIELDS TERMINATED by '\t';
```

Screenshot of Mobaxterm for creating the table Olympics_data:

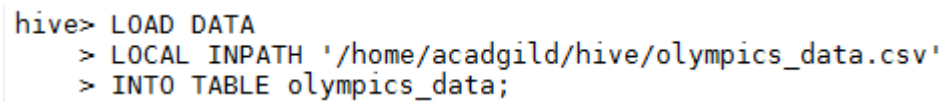
```
hive> CREATE TABLE olympics_data
    > (
    > name string,
    > age int,
    > country string,
    > year int,
    > closing_date string,
    > sport string,
    > gold int,
    > silver int,
    > bronze int,
    > total_medals int
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED by '\t';
OK
```

2) Next we have to load the contents of olympics_data.csv. I stored the file in the path '/home/acadgild/hive/olympics_data.csv'

   So we are loading the contents of the olympics_data.csv into the table olympics_data using the below commands.

```
LOAD DATA
LOCAL INPATH '/home/acadgild/hive/olympics_data.csv'
INTO TABLE olympics_data;
```

Screenshot of Mobaxterm for loading Olympics_data.csv into olympics_data:



Using this table we are going to provide solution for all the queries in this assignment.

Q1) Write a Hive program to find the number of medals won by each country in swimming.

Steps:

For this from the Olympics_data table , we have to select country, sum(total_medals) that is for the sport swimming by each country so we are grouping interms of country. The query is as below.

Query:

```
SELECT country, SUM(total_medals) from olympics_data where sport ='Swimming' GROUP BY country;
```

Output : The output will return country and sum(total_medals) won in swmimming:

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

Screenshot of Mobaxterm for the query with sample output

```
hive> SELECT country, SUM(total_medals) from olympics_data where sport ='Swimming' GROUP BY country;
Query ID = acadgild_20171101210404_06524272-38e0-4050-baa7-489e2b290819
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_1509548254184_0004, Tracking URL = http://localhost:8088/proxy/application_1509548254184_0004/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509548254184_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-01 21:04:47,613 Stage-1 map = 0%,  reduce = 0%
2017-11-01 21:04:57,998 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.63 sec
2017-11-01 21:05:09,246 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.78 sec
MapReduce Total cumulative CPU time: 4 seconds 780 msec
Ended Job = job_1509548254184_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.78 sec   HDFS Read: 518906 HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 780 msec
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
```

Q2) Write a Hive program to find the number of medals that India won year wise..

Steps:

For this from the Olympics_data table , we have to select year, sum(total_medals) that is won by 'India' for each year, so we are grouping interms of year. The query is as below.

Query:

SELECT year, SUM(total_medals) from olympics_data where country ='India' GROUP BY year;

Output : : The output will return year and sum(total_medals) won by country='India'

2000 1
2004 1
2008 3
2012 6

Screenshot of Mobaxterm for the query and output

```
hive> SELECT year, SUM(total_medals) from olympics_data where country ='India' GROUP BY year;
Query ID = acadgild_20171101211010_5ea85527-7b99-485d-bf4d-4b3586a9496c
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_1509548254184_0005, Tracking URL = http://localhost:8088/proxy/application_1509548254184_0005/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509548254184_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-01 21:10:43,657 Stage-1 map = 0%,  reduce = 0%
2017-11-01 21:10:55,366 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.82 sec
2017-11-01 21:11:08,249 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.07 sec
MapReduce Total cumulative CPU time: 6 seconds 70 msec
Ended Job = job_1509548254184_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.07 sec   HDFS Read: 518906 HDFS Write: 28 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 70 msec
OK
2000    1
2004    1
2008    3
2012    6
Time taken: 38.338 seconds, Fetched: 4 row(s)
```

Q3) Write a Hive Program to find the total number of medals each country won

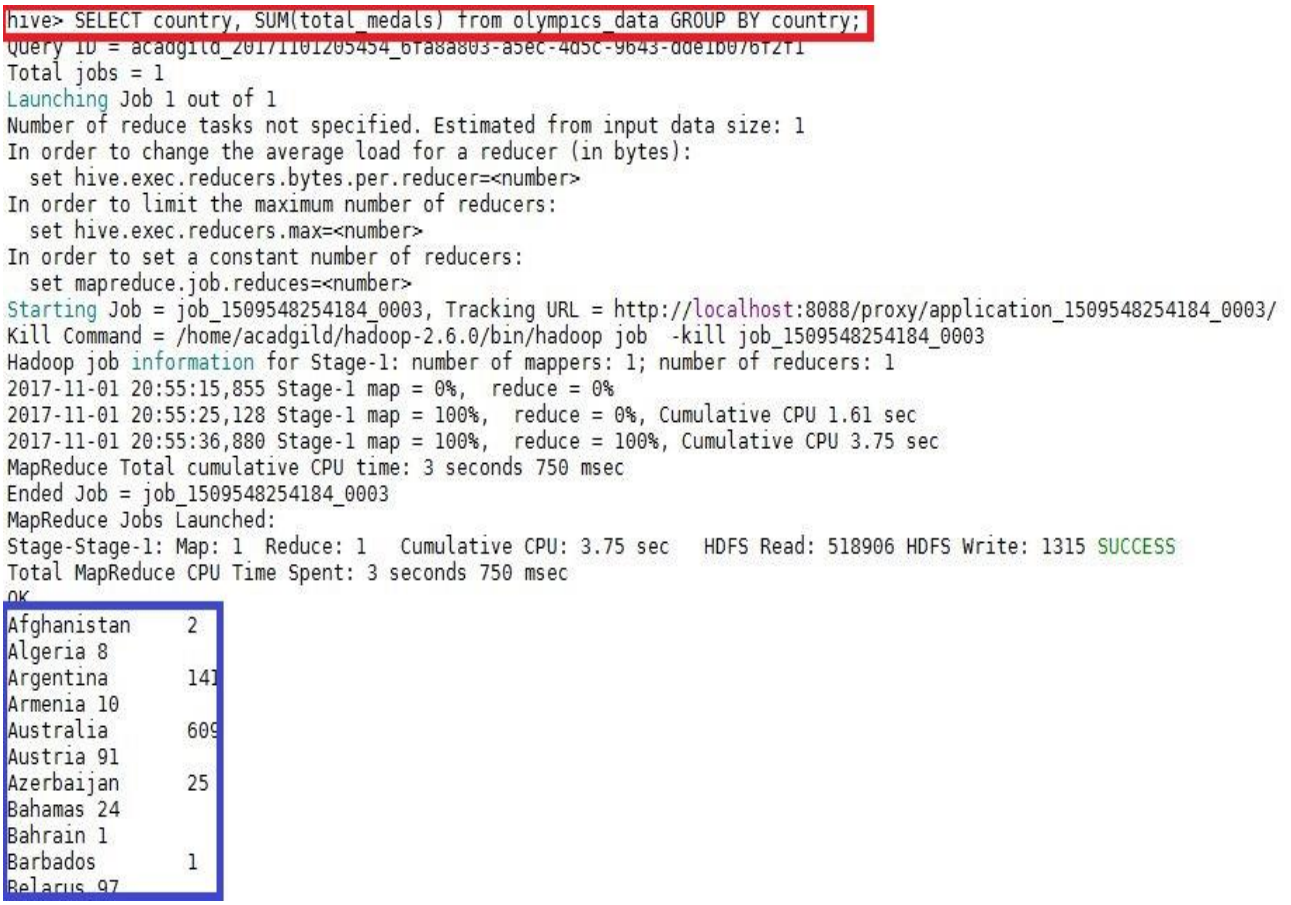
Steps:

For this from the Olympics_data table , we have to select country, sum(total_medals) won by each country so we are grouping interms of country. The query for this is as below.

Query:

SELECT country, SUM(total_medals) from olympics_data GROUP BY country;

Screenshot of Mobaxterm for the query with sample output



Output : The output will return country and sum(total_medals)

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

Q4) Write a Hive program to find the number of gold medals each country won…

Steps:

For this from the Olympics_data table , we have to select country, sum(gold) won by each country so we are grouping interms of country. The query for this is as below.

Query :

SELECT country, SUM(gold) from olympics_data GROUP BY country;

Screenshot of Mobaxterm for the query with sample output

```
hive> SELECT country, SUM(gold) from olympics_data GROUP BY country;
Query ID = acadgild_20171101211818_4cb6f509-6d2a-44b9-8b2d-52f3ad8b395f
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_1509548254184_0006, Tracking URL = http://localhost:8088/proxy/application_1509548254184_0006/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509548254184_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-11-01 21:18:35,107 Stage-1 map = 0%,  reduce = 0%
2017-11-01 21:18:42,819 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.46 sec
2017-11-01 21:18:51,633 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.21 sec
MapReduce Total cumulative CPU time: 3 seconds 210 msec
Ended Job = job_1509548254184_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.21 sec   HDFS Read: 518906 HDFS Write: 1276 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 210 msec
OK
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
```

Output : The output will return country and sum(Gold)

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