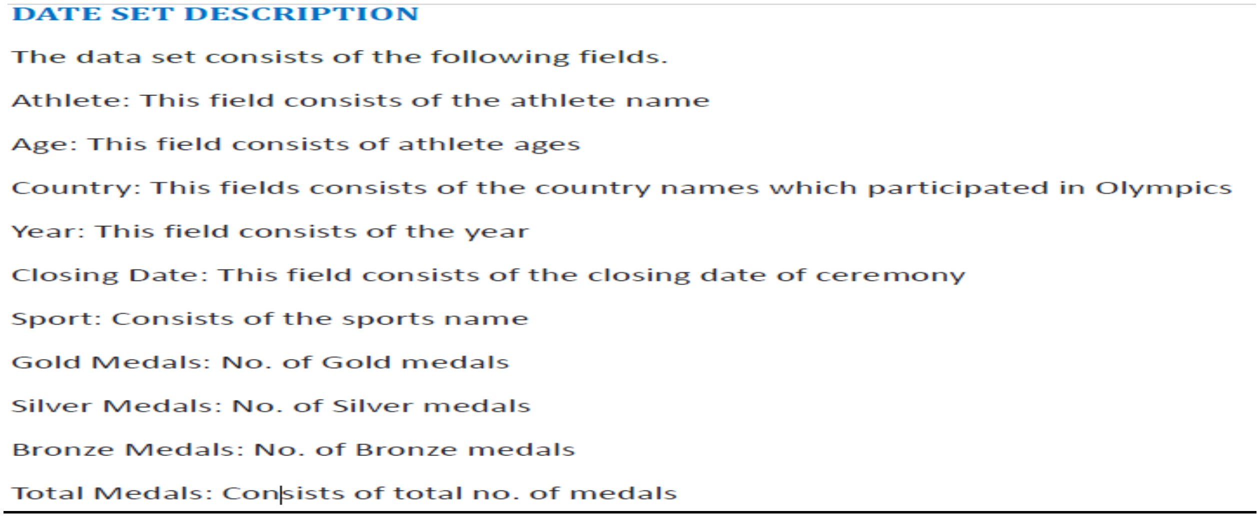
Problem Statement

Data set is about Olympics. You can download the data set from the below link:
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

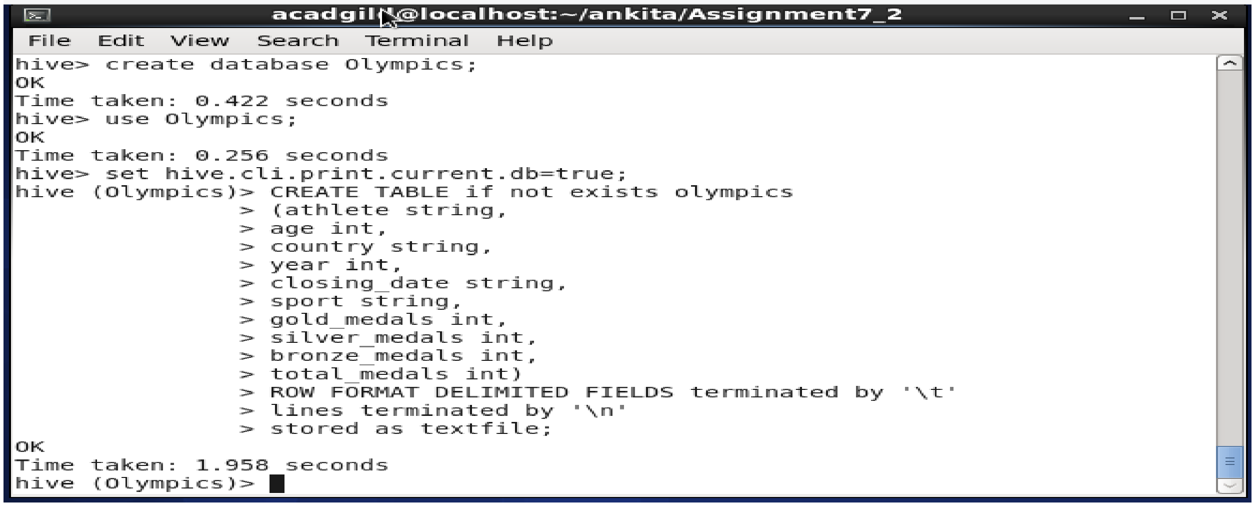
Gold Medals: No. of Gold medals

Silver Medals: No. of Silver medals

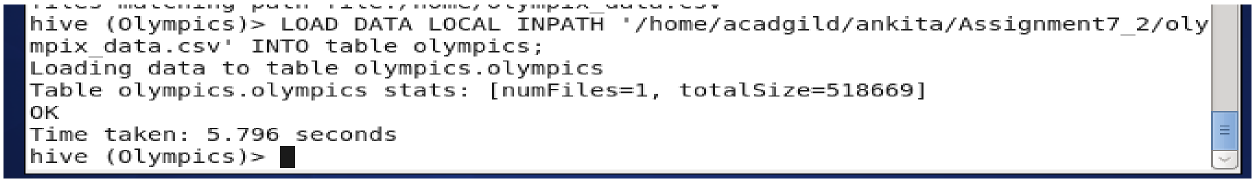
Bronze Medals: No. of Bronze medals

Total Medals: Consists of total no. of medals

Creating a database in Hive named Olympics and creating a table named Olmpics and loading olympix_data.csv dataset into the table


acadgild@localhost:~/ankita/Assignment7_2
File Edit View Search Terminal Help

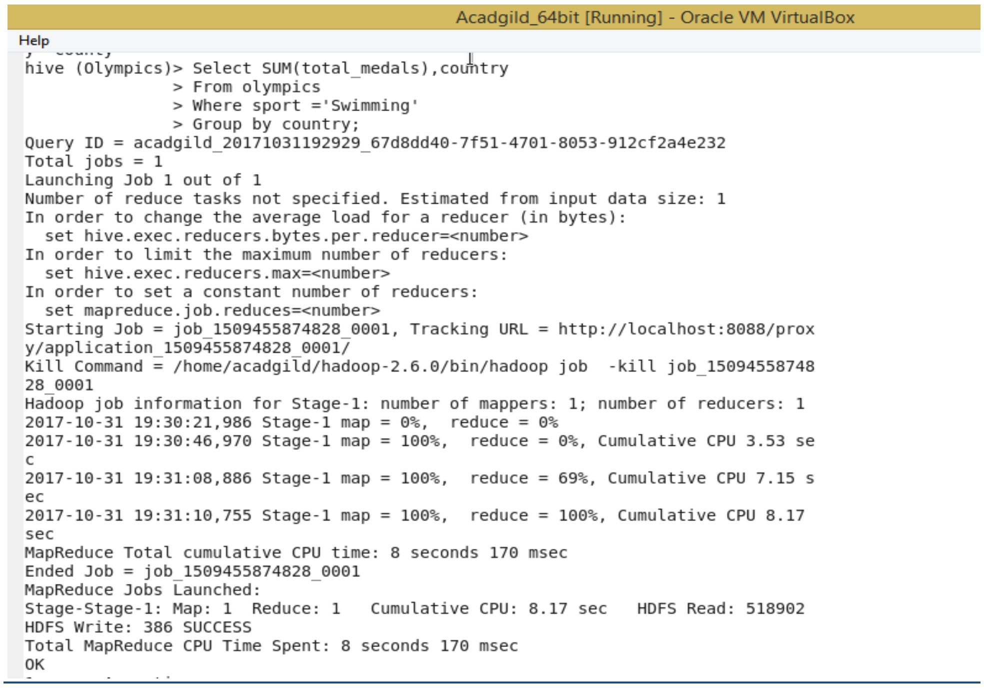
```
hive> create database Olympics;
OK
Time taken: 0.422 seconds
hive> use Olympics;
OK
Time taken: 0.256 seconds
hive> set hive.cli.print.current.db=true;
hive (Olympics)> CREATE TABLE if not exists olympics
               > (athlete string,
               > age int,
               > country string,
               > year int,
               > closing_date string,
               > sport string,
               > gold_medals int,
               > silver_medals int,
               > bronze_medals int,
               > total_medals int)
               > ROW FORMAT DELIMITED FIELDS terminated by '\t'
               > lines terminated by '\n'
               > stored as textfile;
OK
Time taken: 1.958 seconds
hive (Olympics)>
```



```
Files matching path file:/home/olympix_data.csv
hive (Olympics)> LOAD DATA LOCAL INPATH '/home/acadgild/ankita/Assignment7_2/oly
mpix_data.csv' INTO table olympics;
Loading data to table olympics.olympics
Table olympics.olympics stats: [numFiles=1, totalSize=518669]
OK
Time taken: 5.796 seconds
hive (Olympics)>
```

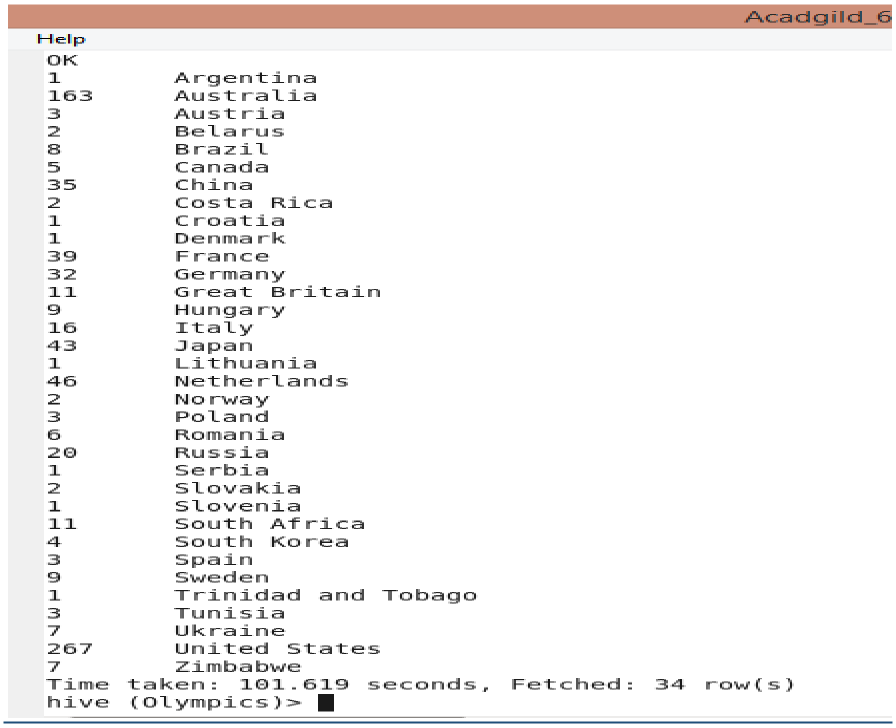
1. Write a Hive program to find the number of medals won by each country in swimming.

Solution:



```
hive (Olympics)> Select SUM(total_medals),country
               > From olympics
               > Where sport ='Swimming'
               > Group by country;
Query ID = acadgild_20171031192929_67d8dd40-7f51-4701-8053-912cf2a4e232
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_1509455874828_0001, Tracking URL = http://localhost:8088/prox
y/application_1509455874828_0001/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_15094558748
28_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-31 19:30:21,986 Stage-1 map = 0%,  reduce = 0%
2017-10-31 19:30:46,970 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.53 se
c
2017-10-31 19:31:08,886 Stage-1 map = 100%,  reduce = 69%, Cumulative CPU 7.15 s
ec
2017-10-31 19:31:10,755 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.17
sec
MapReduce Total cumulative CPU time: 8 seconds 170 msec
Ended Job = job_1509455874828_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.17 sec   HDFS Read: 518902
HDFS Write: 386 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 170 msec
OK
```

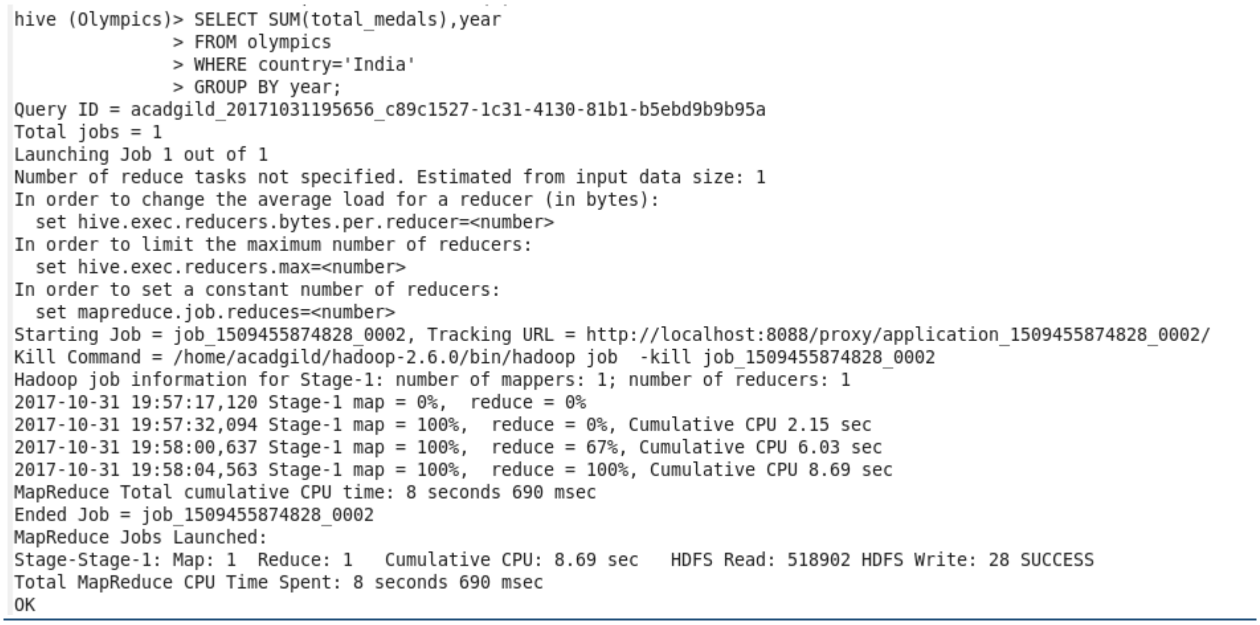
Output:



```
OK
1       Argentina
163     Australia
3       Austria
2       Belarus
8       Brazil
5       Canada
35      China
2       Costa Rica
1       Croatia
1       Denmark
39      France
32      Germany
11      Great Britain
9       Hungary
16      Italy
43      Japan
1       Lithuania
46      Netherlands
2       Norway
3       Poland
6       Romania
20      Russia
1       Serbia
2       Slovakia
1       Slovenia
11      South Africa
4       South Korea
3       Spain
9       Sweden
1       Trinidad and Tobago
3       Tunisia
7       Ukraine
267     United States
7       Zimbabwe
Time taken: 101.619 seconds, Fetched: 34 row(s)
hive (Olympics)> 
```

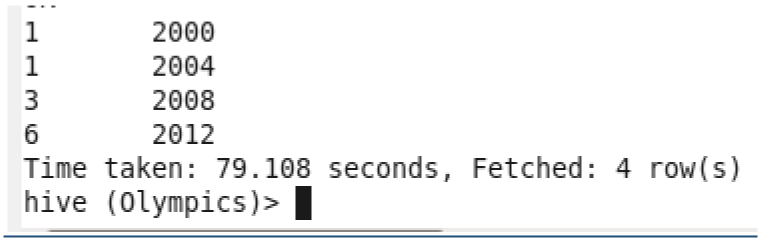
2. Write a Hive program to find the number of medals that India won year wise.

Solution:



```
hive (Olympics)> SELECT SUM(total_medals),year
               > FROM olympics
               > WHERE country='India'
               > GROUP BY year;
Query ID = acadgild_20171031195656_c89c1527-1c31-4130-81b1-b5ebd9b9b95a
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_1509455874828_0002, Tracking URL = http://localhost:8088/proxy/application_1509455874828_0002/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509455874828_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-31 19:57:17,120 Stage-1 map = 0%,  reduce = 0%
2017-10-31 19:57:32,094 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.15 sec
2017-10-31 19:58:00,637 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 6.03 sec
2017-10-31 19:58:04,563 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.69 sec
MapReduce Total cumulative CPU time: 8 seconds 690 msec
Ended Job = job_1509455874828_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.69 sec   HDFS Read: 518902 HDFS Write: 28 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 690 msec
OK
```

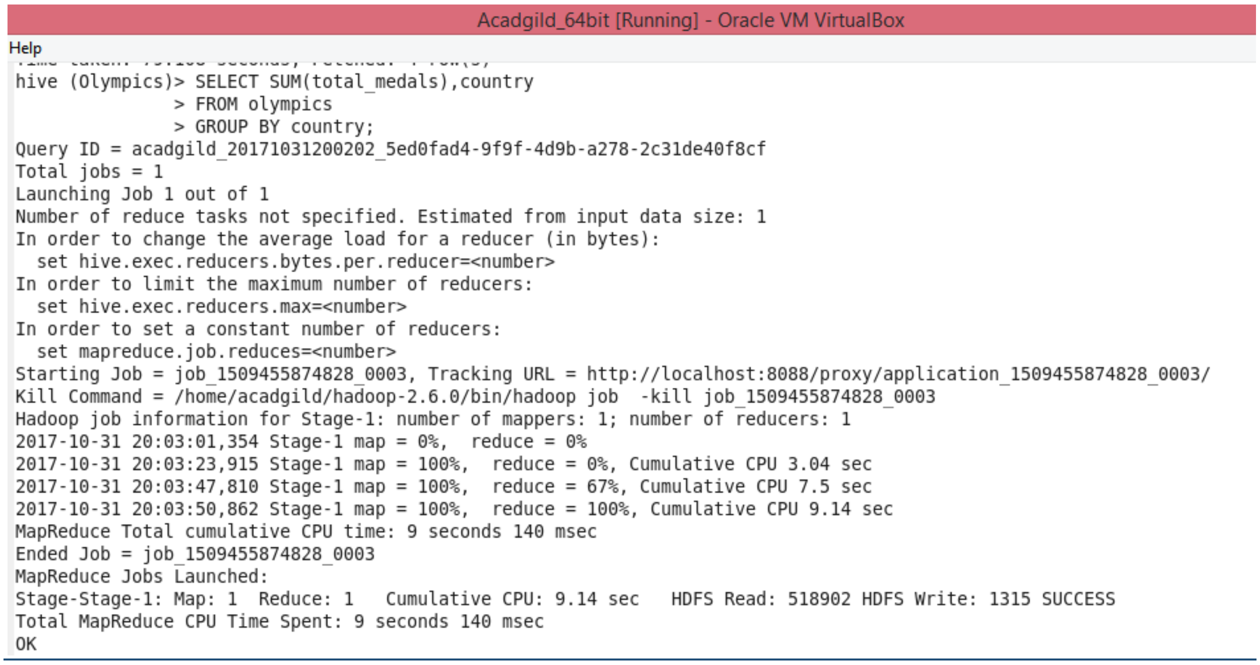
Output:



```
1       2000
1       2004
3       2008
6       2012
Time taken: 79.108 seconds, Fetched: 4 row(s)
hive (Olympics)> 
```

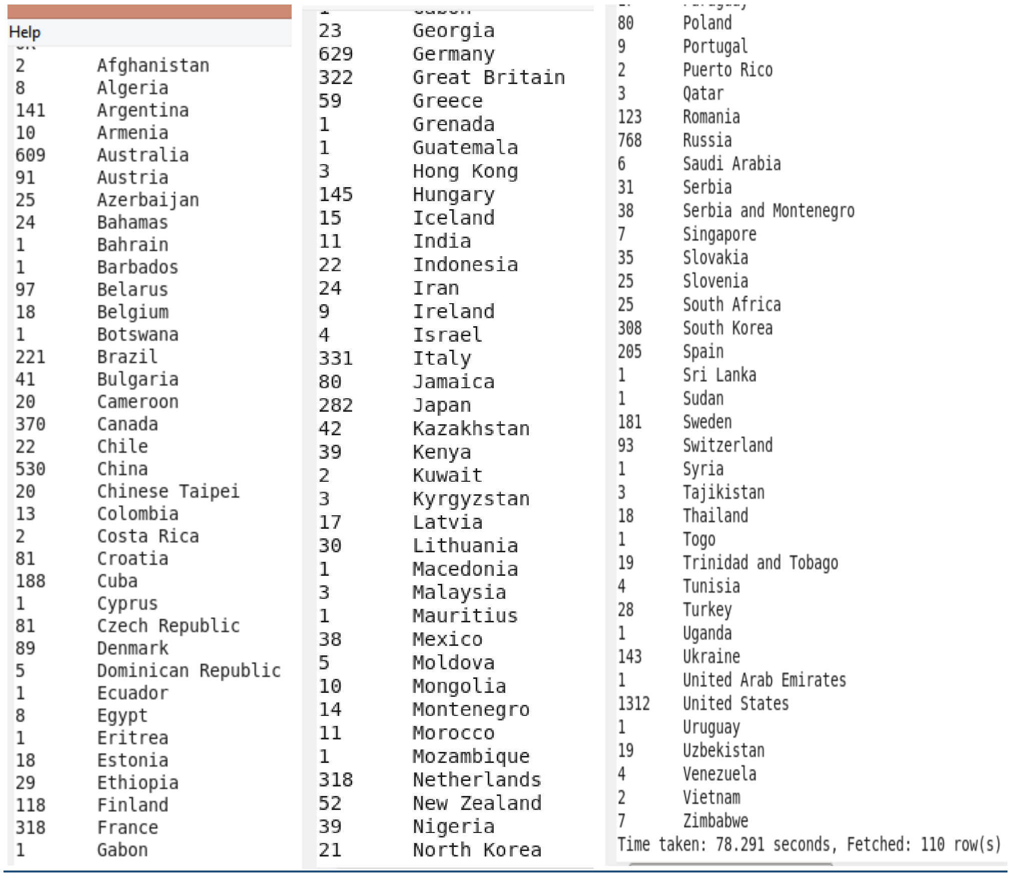
3. Write a Hive Program to find the total number of medals each country won.

Solution:



```
hive (Olympics)> SELECT SUM(total_medals),country
               > FROM olympics
               > GROUP BY country;
Query ID = acadgild_20171031200202_5ed0fad4-9f9f-4d9b-a278-2c31de40f8cf
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_1509455874828_0003, Tracking URL = http://localhost:8088/proxy/application_1509455874828_0003/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509455874828_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-31 20:03:01,354 Stage-1 map = 0%,  reduce = 0%
2017-10-31 20:03:23,915 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.04 sec
2017-10-31 20:03:47,810 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 7.5 sec
2017-10-31 20:03:50,862 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.14 sec
MapReduce Total cumulative CPU time: 9 seconds 140 msec
Ended Job = job_1509455874828_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.14 sec   HDFS Read: 518902 HDFS Write: 1315 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 140 msec
OK
```

Output:



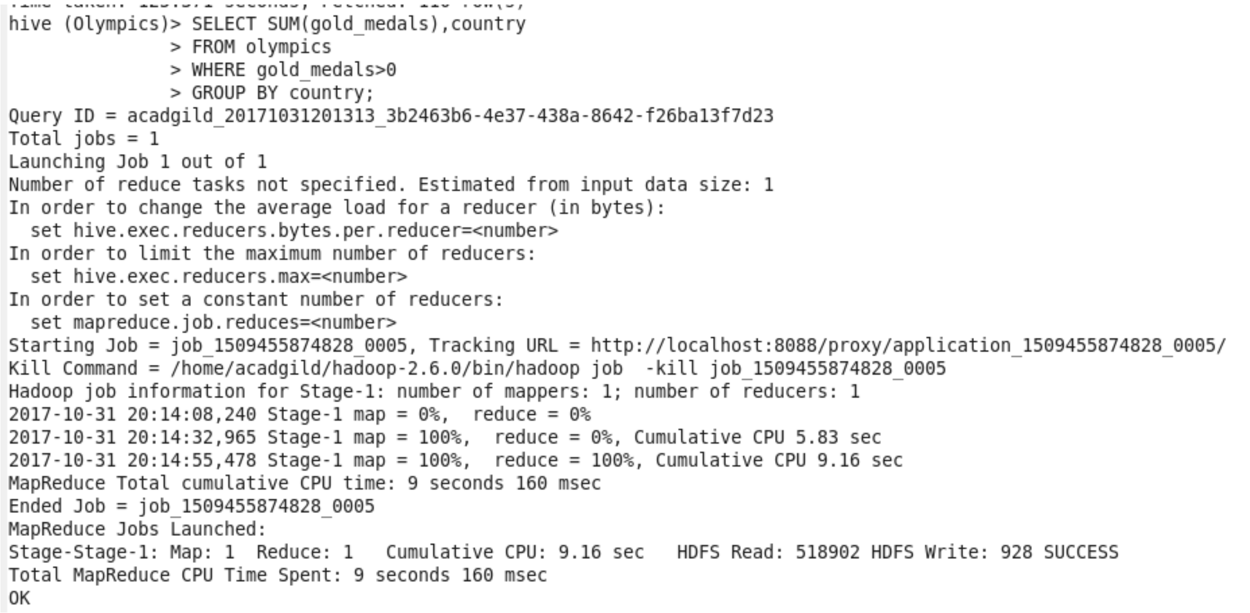
```
2       Afghanistan
8       Algeria
141     Argentina
10      Armenia
609     Australia
91      Austria
25      Azerbaijan
24      Bahamas
1       Bahrain
1       Barbados
97      Belarus
18      Belgium
1       Botswana
221     Brazil
41      Bulgaria
20      Cameroon
370     Canada
22      Chile
530     China
20      Chinese Taipei
13      Colombia
2       Costa Rica
81      Croatia
188     Cuba
1       Cyprus
81      Czech Republic
89      Denmark
5       Dominican Republic
1       Ecuador
8       Egypt
1       Eritrea
18      Estonia
29      Ethiopia
118     Finland
318     France
1       Gabon
```

```
23      Georgia
629     Germany
322     Great Britain
59      Greece
1       Grenada
1       Guatemala
3       Hong Kong
145     Hungary
15      Iceland
11      India
22      Indonesia
24      Iran
9       Ireland
4       Israel
331     Italy
80      Jamaica
282     Japan
42      Kazakhstan
39      Kenya
2       Kuwait
3       Kyrgyzstan
17      Latvia
30      Lithuania
1       Macedonia
3       Malaysia
1       Mauritius
38      Mexico
5       Moldova
10      Mongolia
14      Montenegro
11      Morocco
1       Mozambique
318     Netherlands
52      New Zealand
39      Nigeria
21      North Korea
```

```
80      Poland
9       Portugal
2       Puerto Rico
3       Qatar
123     Romania
768     Russia
6       Saudi Arabia
31      Serbia
38      Serbia and Montenegro
7       Singapore
35      Slovakia
25      Slovenia
25      South Africa
308     South Korea
205     Spain
1       Sri Lanka
1       Sudan
181     Sweden
93      Switzerland
1       Syria
3       Tajikistan
18      Thailand
1       Togo
19      Trinidad and Tobago
4       Tunisia
28      Turkey
1       Uganda
143     Ukraine
1       United Arab Emirates
1312    United States
1       Uruguay
19      Uzbekistan
4       Venezuela
2       Vietnam
7       Zimbabwe
Time taken: 78.291 seconds, Fetched: 110 row(s)
```

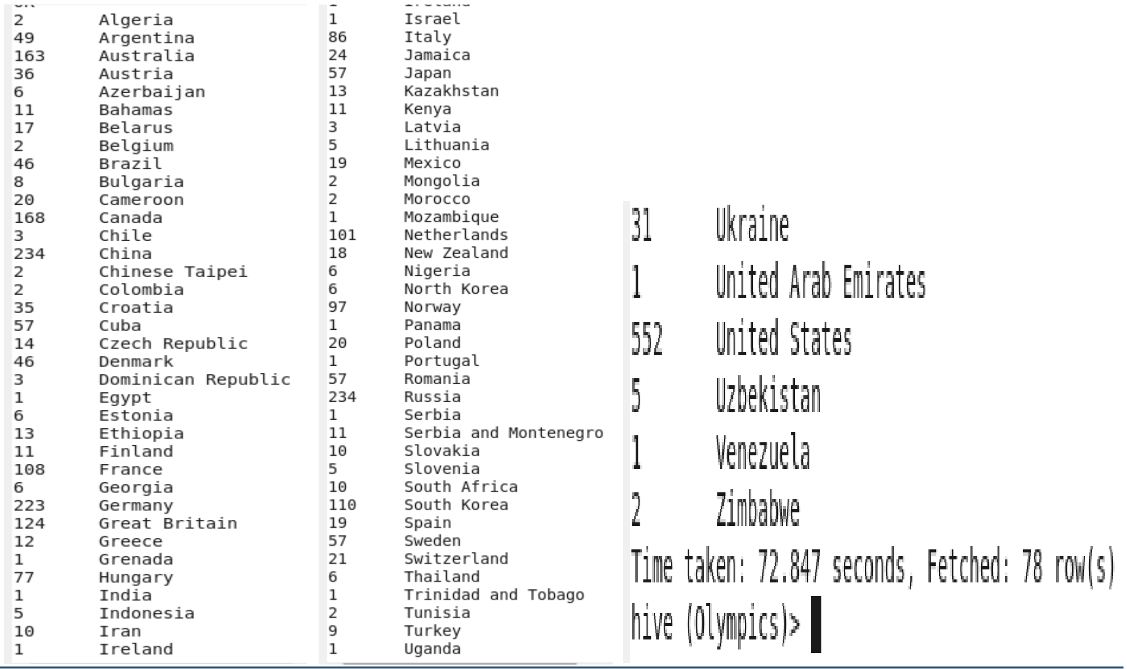
4. Write a Hive program to find the number of gold medals each country won.

Solution:



```
hive (Olympics)> SELECT SUM(gold_medals),country
               > FROM olympics
               > WHERE gold_medals>0
               > GROUP BY country;
Query ID = acadgild_20171031201313_3b2463b6-4e37-438a-8642-f26ba13f7d23
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_1509455874828_0005, Tracking URL = http://localhost:8088/proxy/application_1509455874828_0005/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509455874828_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-31 20:14:08,240 Stage-1 map = 0%,  reduce = 0%
2017-10-31 20:14:32,965 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.83 sec
2017-10-31 20:14:55,478 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.16 sec
MapReduce Total cumulative CPU time: 9 seconds 160 msec
Ended Job = job_1509455874828_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.16 sec   HDFS Read: 518902 HDFS Write: 928 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 160 msec
OK
```

Output:



```
2       Algeria
49      Argentina
163     Australia
36      Austria
6       Azerbaijan
11      Bahamas
17      Belarus
2       Belgium
46      Brazil
8       Bulgaria
20      Cameroon
168     Canada
3       Chile
234     China
2       Chinese Taipei
2       Colombia
35      Croatia
57      Cuba
14      Czech Republic
46      Denmark
3       Dominican Republic
1       Egypt
6       Estonia
13      Ethiopia
11      Finland
108     France
6       Georgia
223     Germany
124     Great Britain
12      Greece
1       Grenada
77      Hungary
1       India
5       Indonesia
10      Iran
1       Ireland
```

```
1       Israel
86      Italy
24      Jamaica
57      Japan
13      Kazakhstan
11      Kenya
3       Latvia
5       Lithuania
19      Mexico
2       Mongolia
2       Morocco
1       Mozambique
101     Netherlands
18      New Zealand
6       Nigeria
6       North Korea
97      Norway
1       Panama
20      Poland
1       Portugal
57      Romania
234     Russia
1       Serbia
11      Serbia and Montenegro
10      Slovakia
5       Slovenia
10      South Africa
110     South Korea
19      Spain
57      Sweden
21      Switzerland
6       Thailand
1       Trinidad and Tobago
2       Tunisia
9       Turkey
1       Uganda
```

```
31      Ukraine
1       United Arab Emirates
552     United States
5       Uzbekistan
1       Venezuela
2       Zimbabwe
Time taken: 72.847 seconds, Fetched: 78 row(s)
hive (Olympics)>
```