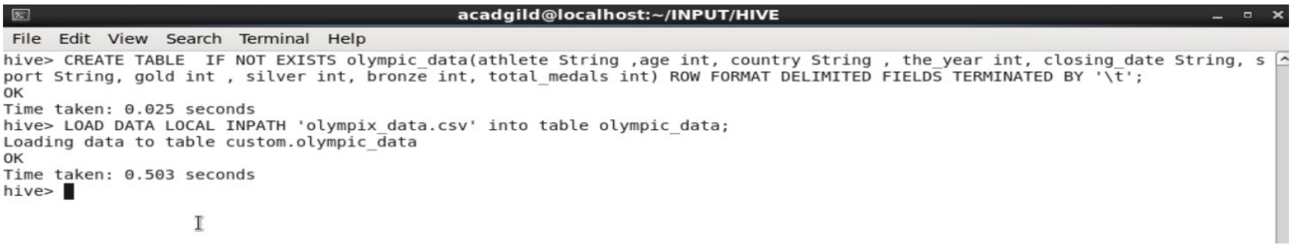
Bigdata Assignment 3.2

This 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

Table Creation:- Dataset olympix.csv was loaded into a table called olmpic_data with the given columns.

CREATE TABLE IF NOT EXISTS olympic_data(athlete String , age int , country String, the_year int , closing_date String , sport String , gold int , silver int , bronze int , total_medals int) ROW FORAT DELIMITED FIELDS TERMINATED BY '\t' ;
LOAD DATA LOCAL INPATH 'olmpix.csv' into table olympic_data ;

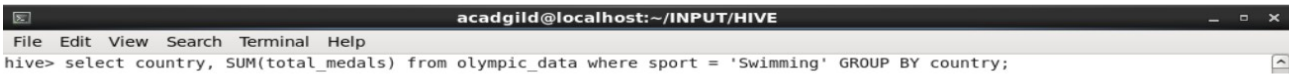

acadgild@localhost:~/INPUT/HIVE
File Edit View Search Terminal Help

```
hive> CREATE TABLE  IF NOT EXISTS olympic_data(athlete String ,age int, country String , the_year int, closing_date String, s
port String, gold int , silver int, bronze int, total_medals int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t';
OK
Time taken: 0.025 seconds
hive> LOAD DATA LOCAL INPATH 'olympix_data.csv' into table olympic_data;
Loading data to table custom.olympic_data
OK
Time taken: 0.503 seconds
hive> 
```

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

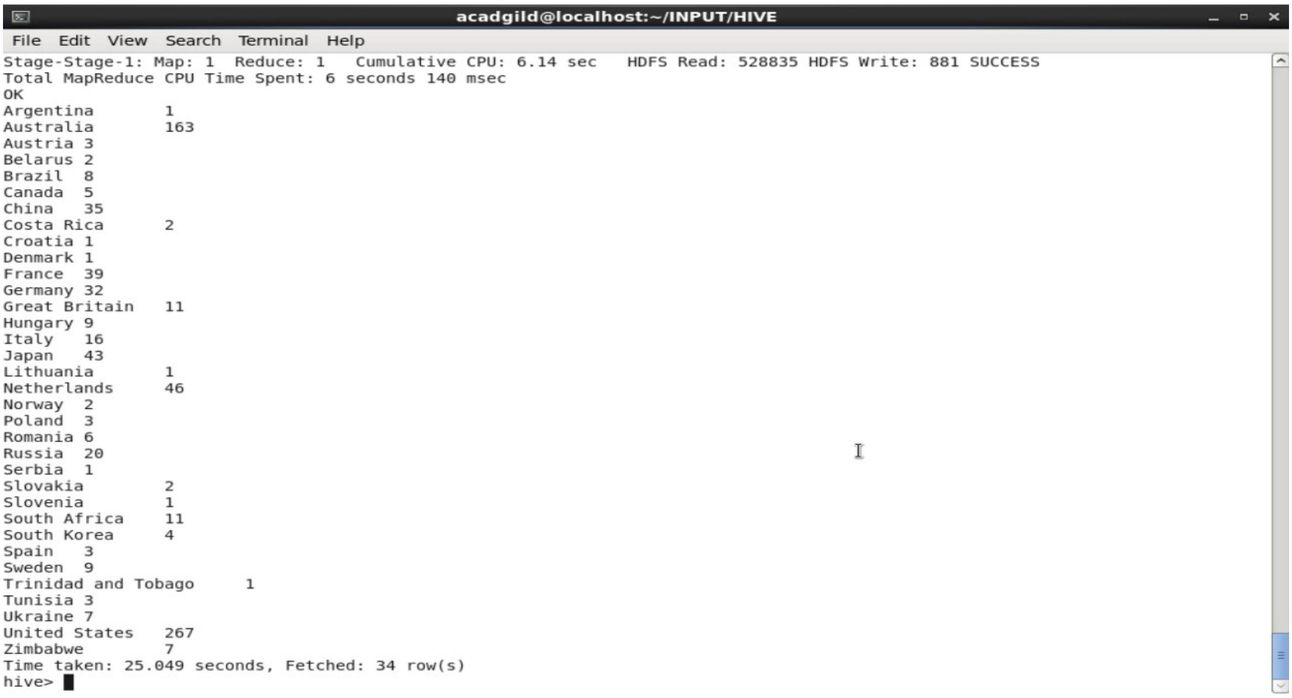
Ans – We selected each country and sum of total medals in Swimming using GROUP BY.
select country , SUM(total_medals) from olmpic_data where sport =

'Swimming' GROUP BY country;



```
hive> select country, SUM(total_medals) from olympic_data where sport = 'Swimming' GROUP BY country;
```

Output- We got the list of each country and its corresponding no of total medals in Swimming.

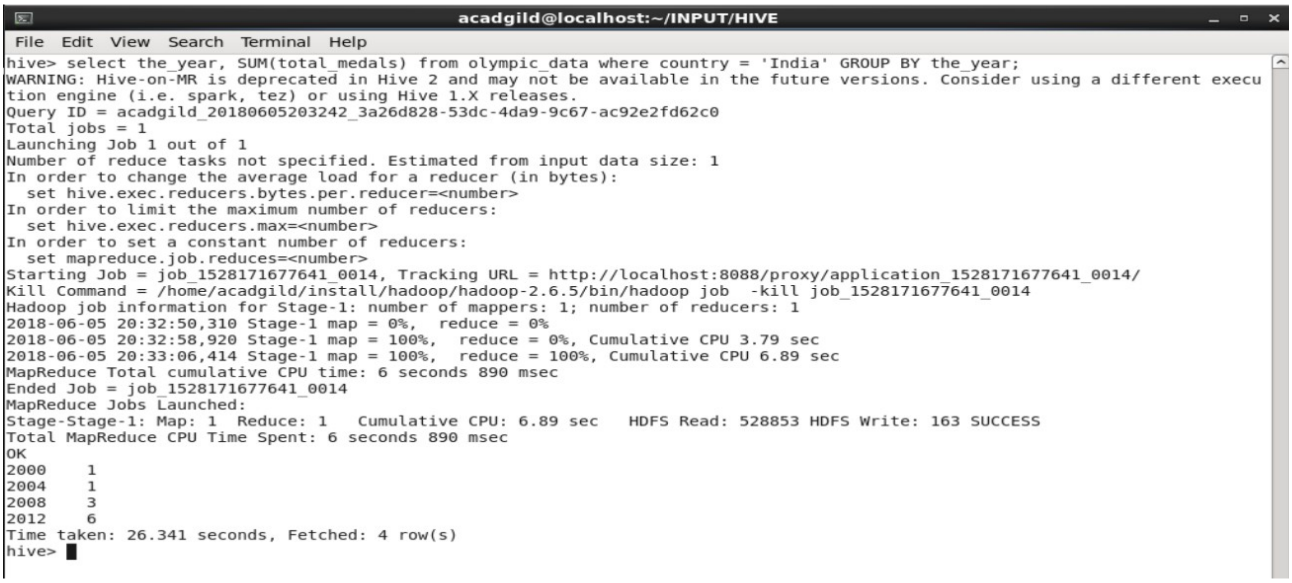


```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.14 sec   HDFS Read: 528835 HDFS Write: 881 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 140 msec
OK
Argentina       1
Australia       163
Austria 3
Belarus 2
Brazil  8
Canada  5
China   35
Costa Rica      2
Croatia 1
Denmark 1
France  39
Germany 32
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: 25.049 seconds, Fetched: 34 row(s)
hive> █
```

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

Ans – We calculated in each year no of medals won by India using GROUP BY.

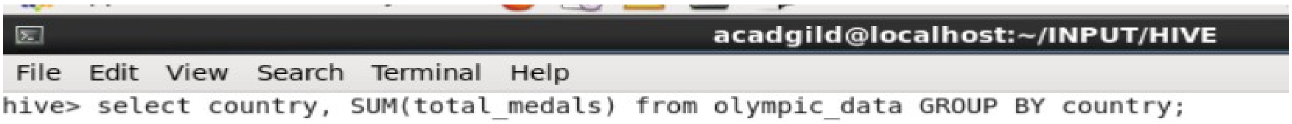
select the_year , SUM(total_medals) from olympic_data where country = 'India' GROUP BY the_year;


acadgild@localhost:~/INPUT/HIVE
File Edit View Search Terminal Help

```
hive> select the_year, SUM(total_medals) from olympic_data where country = 'India' GROUP BY the_year;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180605203242_3a26d828-53dc-4da9-9c67-ac92e2fd62c0
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_1528171677641_0014, Tracking URL = http://localhost:8088/proxy/application_1528171677641_0014/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1528171677641_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-05 20:32:50,310 Stage-1 map = 0%,  reduce = 0%
2018-06-05 20:32:58,920 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.79 sec
2018-06-05 20:33:06,414 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.89 sec
MapReduce Total cumulative CPU time: 6 seconds 890 msec
Ended Job = job_1528171677641_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.89 sec   HDFS Read: 528853 HDFS Write: 163 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 890 msec
OK
2000    1
2004    1
2008    3
2012    6
Time taken: 26.341 seconds, Fetched: 4 row(s)
hive> 
```

In the above screenshot we got the output in which year India got medals.

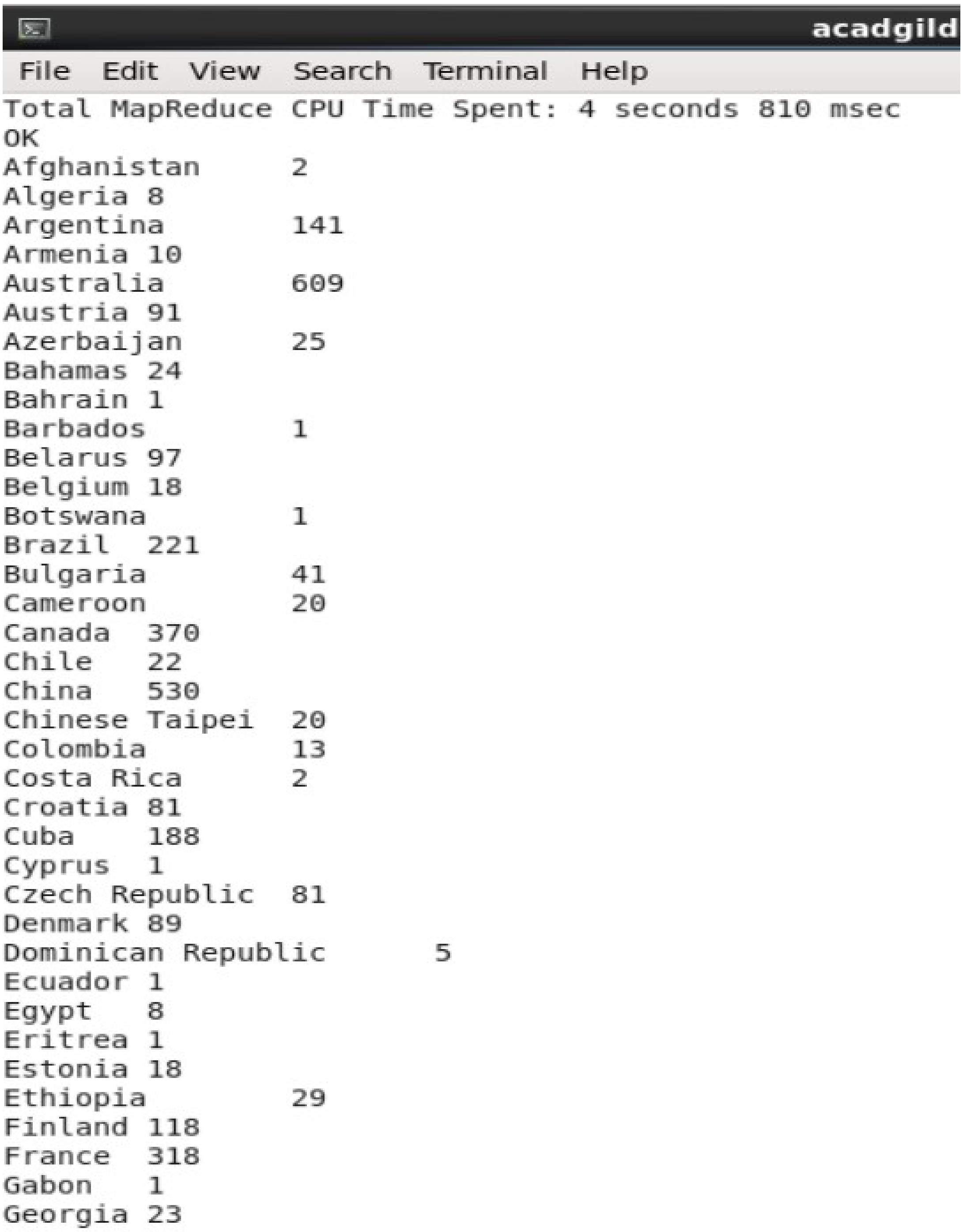
3. Write a Hive Program to find the total number of medals each country won.


acadgild@localhost:~/INPUT/HIVE
File Edit View Search Terminal Help

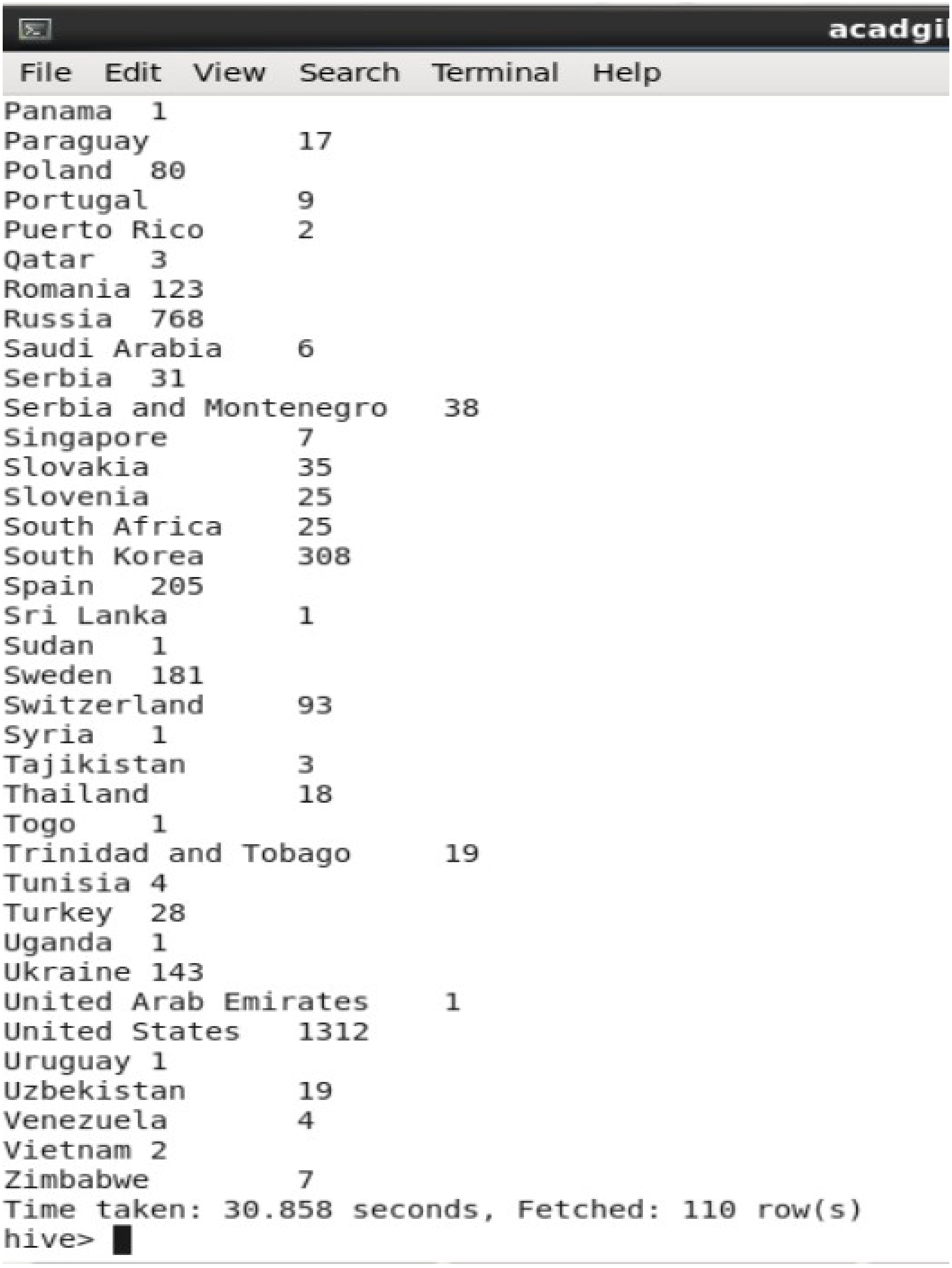
```
hive> select country, SUM(total_medals) from olympic_data GROUP BY country;
```

Ans – We calculated no of medals won by each country using GROUP By.
select country , SUM(total_medals) from olmpic_data GROUP BY country;

OUTPUT -
We got the list of country and its corresponding no of total medals won.



```
Total MapReduce CPU Time Spent: 4 seconds 810 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
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
```

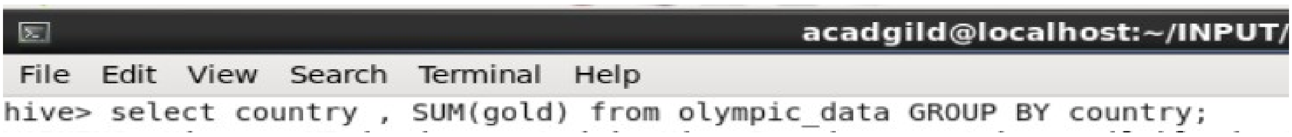


```
Panama  1
Paraguay        17
Poland  80
Portugal        9
Puerto Rico     2
Qatar   3
Romania 123
Russia  768
Saudi Arabia    6
Serbia  31
Serbia and Montenegro   38
Singapore       7
Slovakia        35
Slovenia        25
South Africa    25
South Korea     308
Spain   205
Sri Lanka       1
Sudan   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: 30.858 seconds, Fetched: 110 row(s)
hive> █
```

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

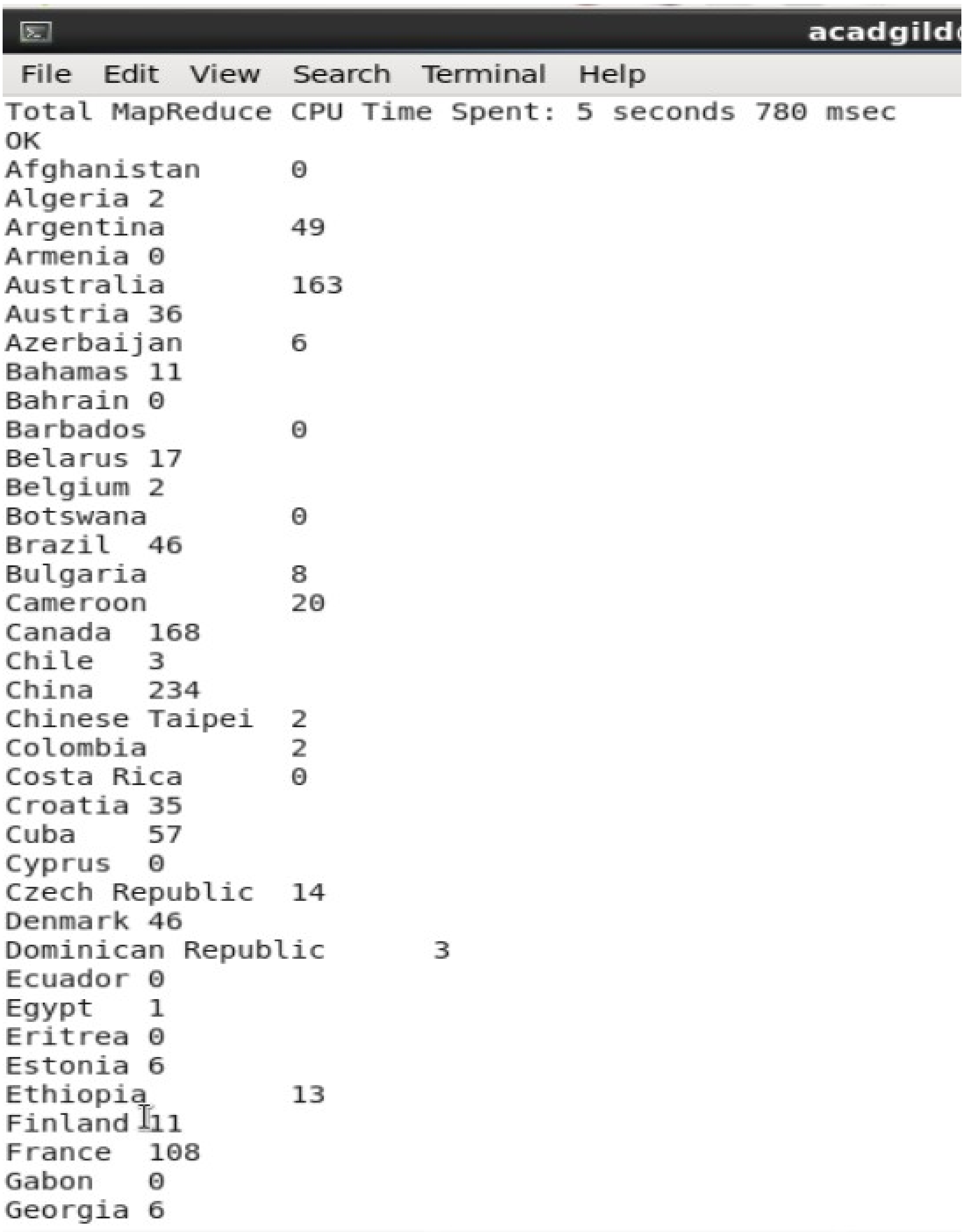
Ans -   We calculated no of gold medals won by each country using GROUP By.

select country , SUM(gold) from olmpic_data GROUP BY country;

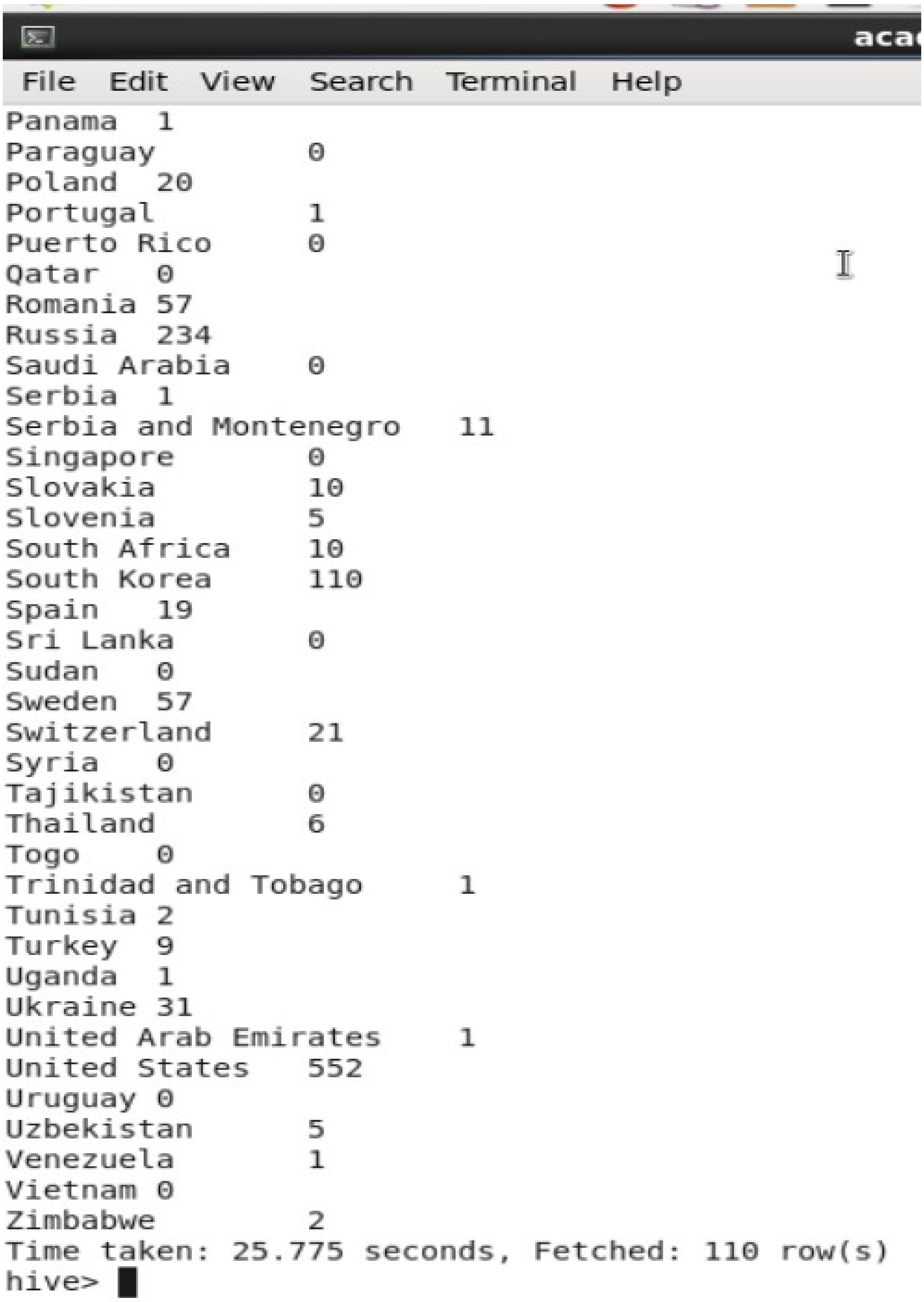


```
hive> select country , SUM(gold) from olympic_data GROUP BY country;
```

OUTPUT - We got the list of country and its corresponding no of total gold medals won.



```
Total MapReduce CPU Time Spent: 5 seconds 780 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
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
```



```
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: 25.775 seconds, Fetched: 110 row(s)
hive> 
```