BIG DATA HADOOP & SPARK TRAINING

ACADGILD

ASSIGNMENT 7.2

BY :-

SAHIL KHURANA

PROBLEM STATEMENT

1. Write a Hive program to find the number of medals won by each country in swimming.
2. Write a Hive program to find the number of medals that India won year wise.
3. Write a Hive Program to find the total number of medals each country won.
4. Write a Hive program to find the number of gold medals each country won.

Associated Data Files

olympix_data.csv

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

DATE SET DESCRIPTION

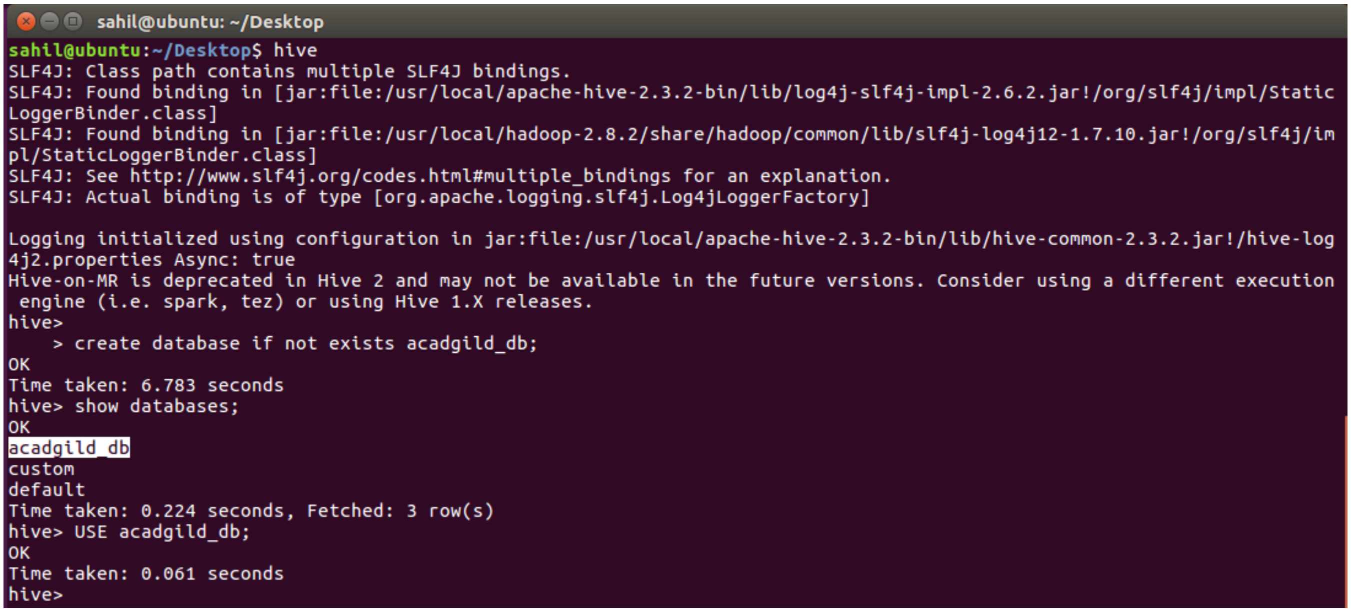
The data set consists of the following fields.

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

Note: - To solve the Assignment, I have created a VM with Ubuntu 16.04 OS and configured Hadoop 2.8.2 and hive-2.3.2 on the same.

Open the Hive Shell and CREATE the DATABASE.

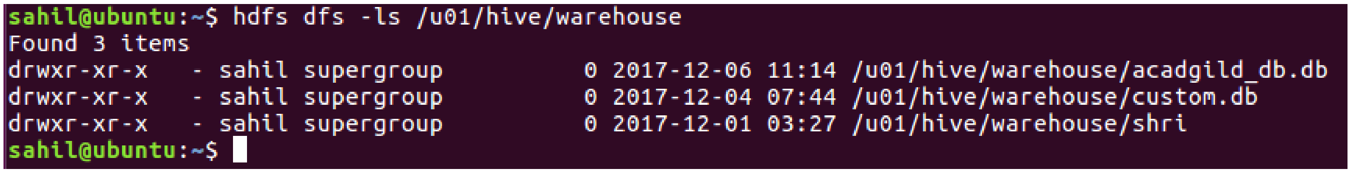
```
hive                                              -- open the hive shell
hive> create database if not exists acadgild_db;  -- create database
hive> show databases;                             -- check whether database is created or not
hive> USE acadgild_db;                            -- use database is USE
```



```
sahil@ubuntu:~/Desktop$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.8.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/im
pl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log
4j2.properties Async: true
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.
hive>
    > create database if not exists acadgild_db;
OK
Time taken: 6.783 seconds
hive> show databases;
OK
acadgild_db
custom
default
Time taken: 0.224 seconds, Fetched: 3 row(s)
hive> USE acadgild_db;
OK
Time taken: 0.061 seconds
hive>
```

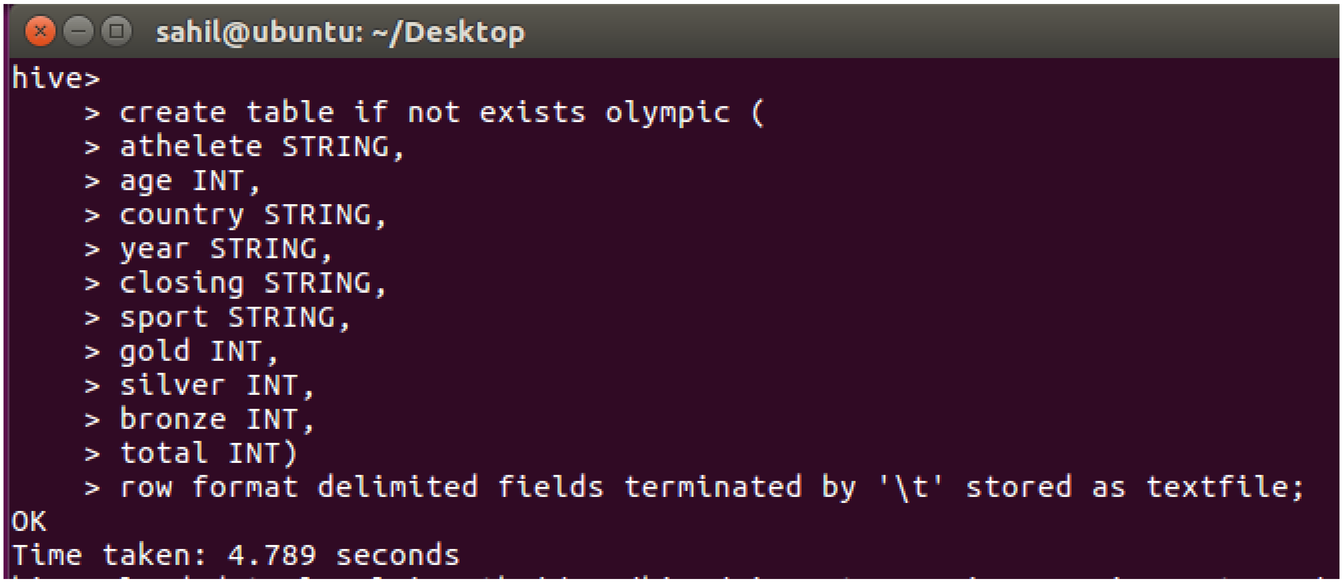
Custom database created in default directory in hive



```
sahil@ubuntu:~$ hdfs dfs -ls /u01/hive/warehouse
Found 3 items
drwxr-xr-x   - sahil supergroup          0 2017-12-06 11:14 /u01/hive/warehouse/acadgild_db.db
drwxr-xr-x   - sahil supergroup          0 2017-12-04 07:44 /u01/hive/warehouse/custom.db
drwxr-xr-x   - sahil supergroup          0 2017-12-01 03:27 /u01/hive/warehouse/shri
sahil@ubuntu:~$
```

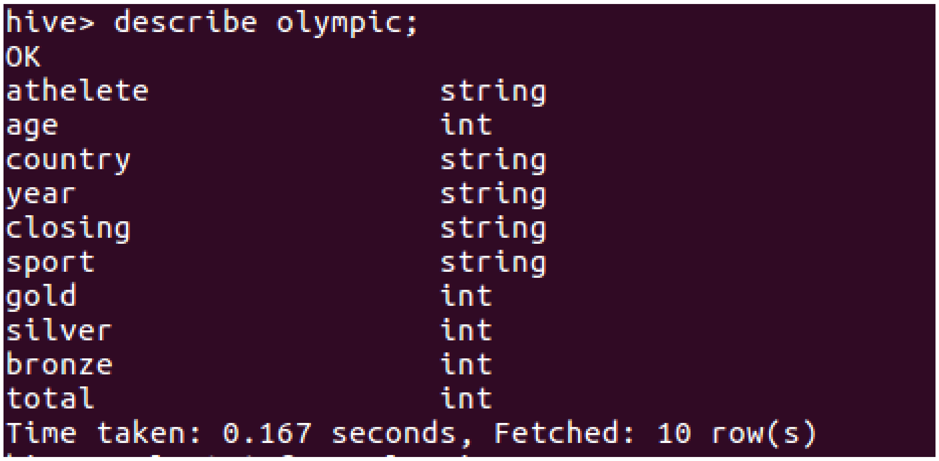
CREATE TABLE

```
hive>
    > create table if not exists 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;
```



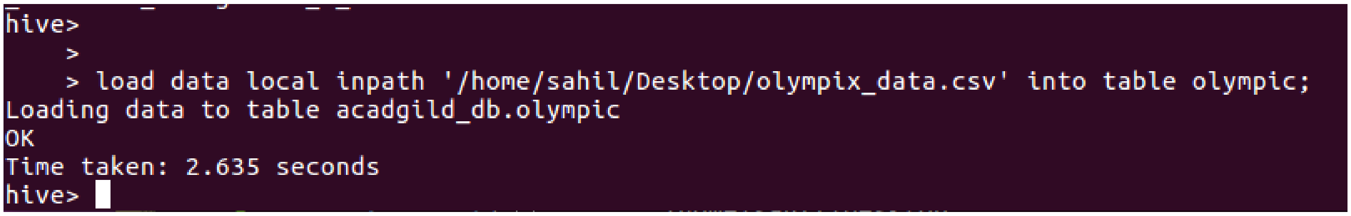
```
hive>
    > create table if not exists olympic (
    > athelete STRING,
    > age INT,
    > country STRING,
    > year STRING,
    > closing STRING,
    > sport STRING,
    > gold INT,
    > silver INT,
    > bronze INT,
    > total INT)
    > row format delimited fields terminated by '\t' stored as textfile;
OK
Time taken: 4.789 seconds
```

hive> describe olympic;



```
hive> describe olympic;
OK
athelete                string
age                     int
country                 string
year                    string
closing                 string
sport                   string
gold                    int
silver                  int
bronze                  int
total                   int
Time taken: 0.167 seconds, Fetched: 10 row(s)
```

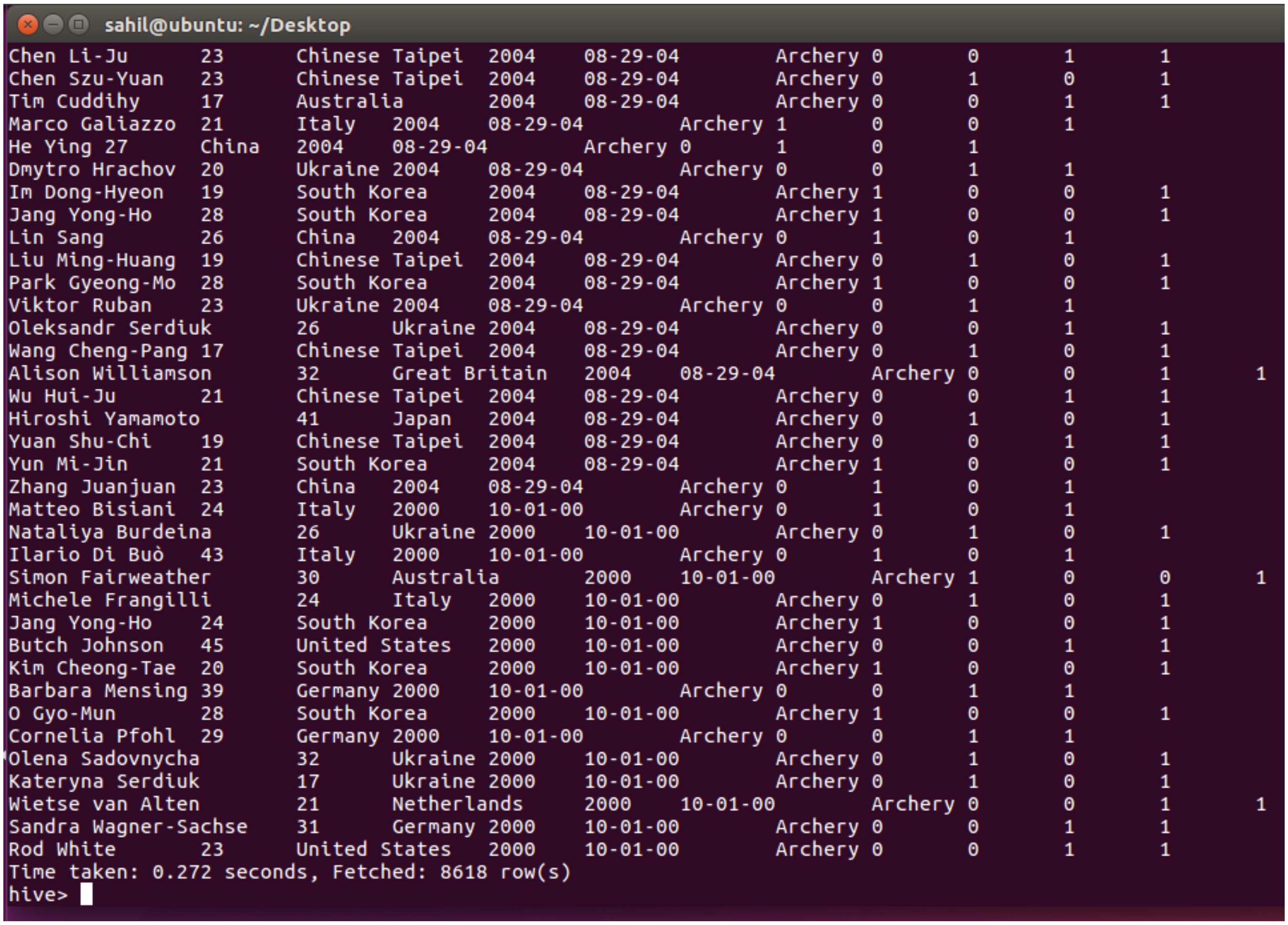
hive>
    > load data local inpath '/home/sahil/Desktop/olympix_data.csv' into table olympic;



```
hive>
    >
    > load data local inpath '/home/sahil/Desktop/olympix_data.csv' into table olympic;
Loading data to table acadgild_db.olympic
OK
Time taken: 2.635 seconds
hive> 
```

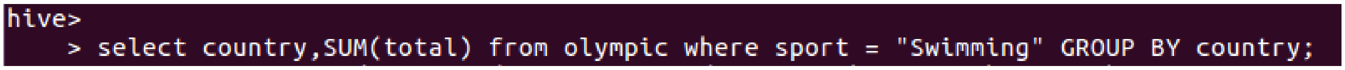
Check whether the dataset is imported in the hive table or not.

hive>
    > select * from olympic;



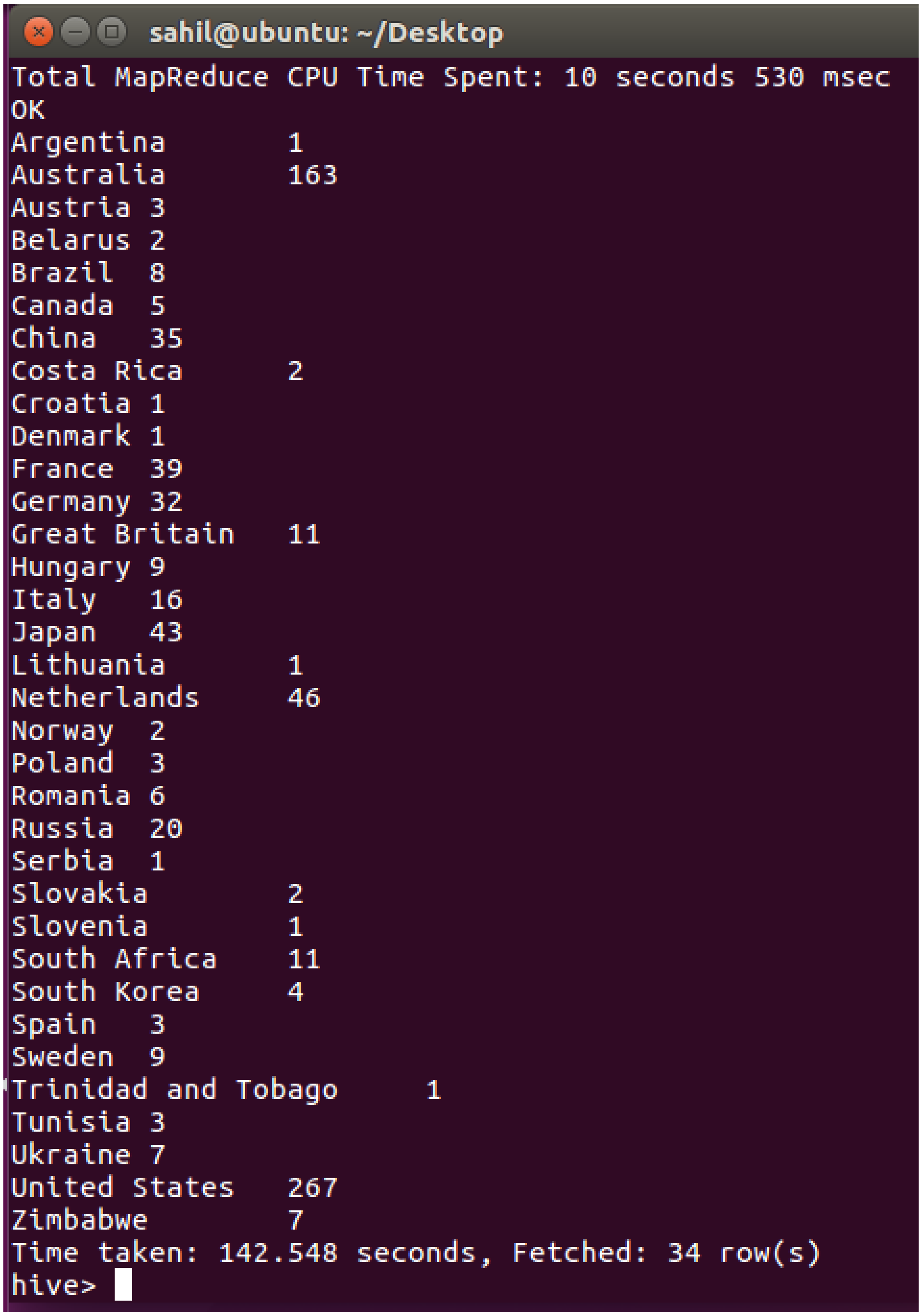
```
Chen Li-Ju      23      Chinese Taipei  2004    08-29-04        Archery 0       0       1       1
Chen Szu-Yuan   23      Chinese Taipei  2004    08-29-04        Archery 0       1       0       1
Tim Cuddihy     17      Australia       2004    08-29-04        Archery 0       0       1       1
Marco Galiazzo  21      Italy   2004    08-29-04        Archery 1       0       0       1
He Ying 27      China   2004    08-29-04        Archery 0       1       0       1
Dmytro Hrachov  20      Ukraine 2004    08-29-04        Archery 0       0       1       1
Im Dong-Hyeon   19      South Korea     2004    08-29-04        Archery 1       0       0       1
Jang Yong-Ho    28      South Korea     2004    08-29-04        Archery 1       0       0       1
Lin Sang        26      China   2004    08-29-04        Archery 0       1       0       1
Liu Ming-Huang  19      Chinese Taipei  2004    08-29-04        Archery 0       1       0       1
Park Gyeong-Mo  28      South Korea     2004    08-29-04        Archery 1       0       0       1
Viktor Ruban    23      Ukraine 2004    08-29-04        Archery 0       0       1       1
Oleksandr Serdiuk       26      Ukraine 2004    08-29-04        Archery 0       0       1       1
Wang Cheng-Pang 17      Chinese Taipei  2004    08-29-04        Archery 0       1       0       1
Alison Williamson       32      Great Britain   2004    08-29-04        Archery 0       0       1       1
Wu Hui-Ju       21      Chinese Taipei  2004    08-29-04        Archery 0       0       1       1
Hiroshi Yamamoto        41      Japan   2004    08-29-04        Archery 0       1       0       1
Yuan Shu-Chi    19      Chinese Taipei  2004    08-29-04        Archery 0       0       1       1
Yun Mi-Jin      21      South Korea     2004    08-29-04        Archery 1       0       0       1
Zhang Juanjuan  23      China   2004    08-29-04        Archery 0       1       0       1
Matteo Bisiani  24      Italy   2000    10-01-00        Archery 0       1       0       1
Nataliya Burdeina       26      Ukraine 2000    10-01-00        Archery 0       1       0       1
Ilario Di Buò   43      Italy   2000    10-01-00        Archery 0       1       0       1
Simon Fairweather       30      Australia       2000    10-01-00        Archery 1       0       0       1
Michele Frangilli       24      Italy   2000    10-01-00        Archery 0       1       0       1
Jang Yong-Ho    24      South Korea     2000    10-01-00        Archery 1       0       0       1
Butch Johnson   45      United States   2000    10-01-00        Archery 0       0       1       1
Kim Cheong-Tae  20      South Korea     2000    10-01-00        Archery 1       0       0       1
Barbara Mensing 39      Germany 2000    10-01-00        Archery 0       0       1       1
O Gyo-Mun       28      South Korea     2000    10-01-00        Archery 1       0       0       1
Cornelia Pfohl  29      Germany 2000    10-01-00        Archery 0       0       1       1
Olena Sadovnycha        32      Ukraine 2000    10-01-00        Archery 0       1       0       1
Kateryna Serdiuk        17      Ukraine 2000    10-01-00        Archery 0       1       0       1
Wietse van Alten        21      Netherlands     2000    10-01-00        Archery 0       0       1       1
Sandra Wagner-Sachse    31      Germany 2000    10-01-00        Archery 0       0       1       1
Rod White       23      United States   2000    10-01-00        Archery 0       0       1       1
Time taken: 0.272 seconds, Fetched: 8618 row(s)
hive> 
```

1. Write a Hive program to find the number of medals won by each country in swimming.
   Commands of Problem 1
   hive>
   > select country,SUM(total) from olympic where sport = "Swimming" GROUP BY country;



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

Result:-



```
Total MapReduce CPU Time Spent: 10 seconds 530 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: 142.548 seconds, Fetched: 34 row(s)
hive> 
```

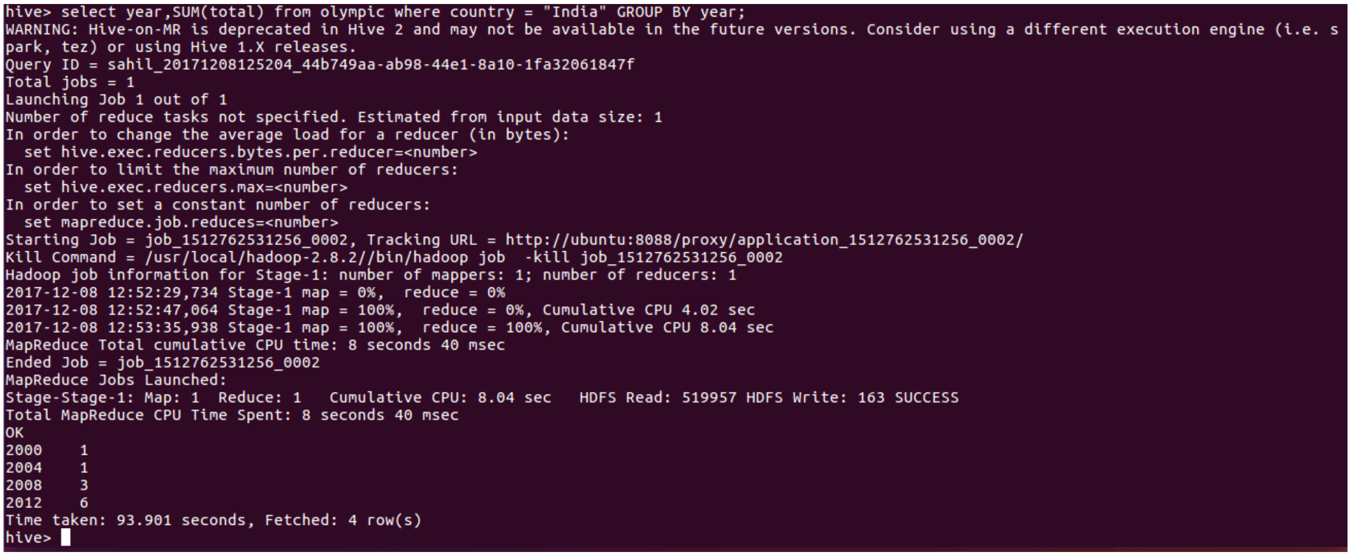
2. Write a Hive program to find the number of medals that India won year wise.

   Commands of Problem 2

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

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

Result:-



```
hive> select year,SUM(total) from olympic where country = "India" GROUP BY year;
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. s
park, tez) or using Hive 1.X releases.
Query ID = sahil_20171208125204_44b749aa-ab98-44e1-8a10-1fa32061847f
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_1512762531256_0002, Tracking URL = http://ubuntu:8088/proxy/application_1512762531256_0002/
Kill Command = /usr/local/hadoop-2.8.2//bin/hadoop job  -kill job_1512762531256_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-08 12:52:29,734 Stage-1 map = 0%,  reduce = 0%
2017-12-08 12:52:47,064 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.02 sec
2017-12-08 12:53:35,938 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.04 sec
MapReduce Total cumulative CPU time: 8 seconds 40 msec
Ended Job = job_1512762531256_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.04 sec   HDFS Read: 519957 HDFS Write: 163 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 40 msec
OK
2000    1
2004    1
2008    3
2012    6
Time taken: 93.901 seconds, Fetched: 4 row(s)
hive>
```

2000	1
2004	1
2008	3
2012	6

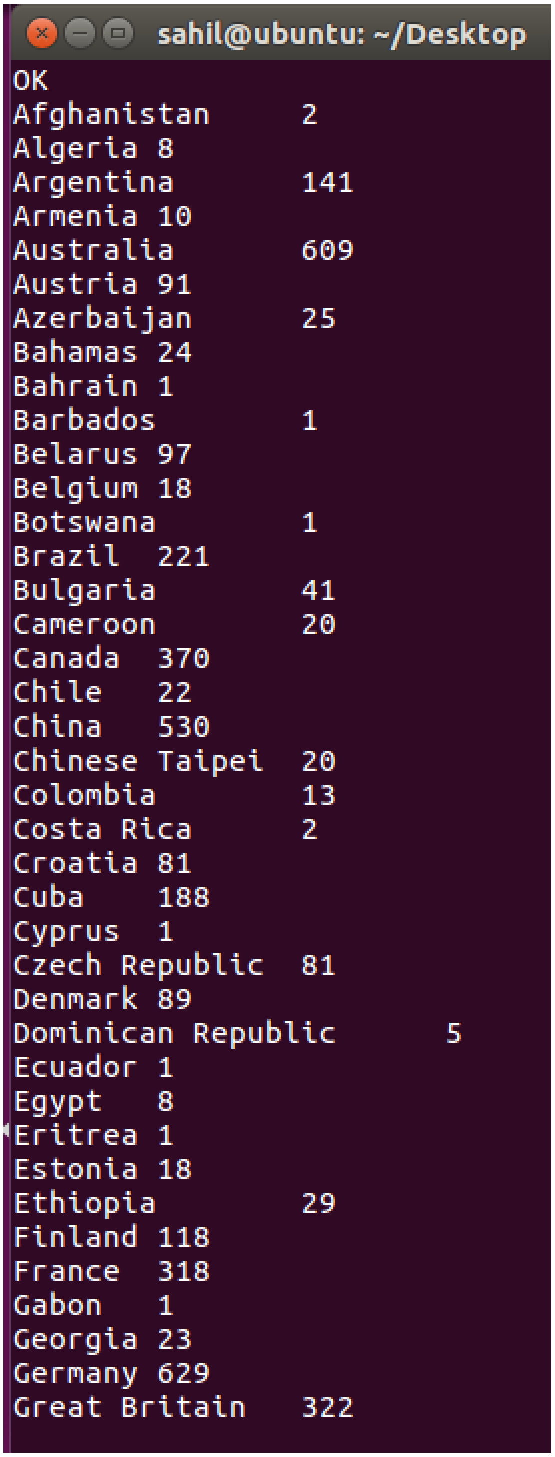
3. Write a Hive Program to find the total number of medals each country won.

   Commands of Problem 3

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

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

Result:-



```
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
Germany 629
Great Britain   322
```

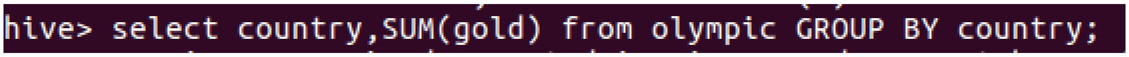
Afghanistan	2
Algeria	8
Argentina	141
Armenia	10
Australia	609
Austria	91
Azerbaijan	25
Bahamas	24
Bahrain	1

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

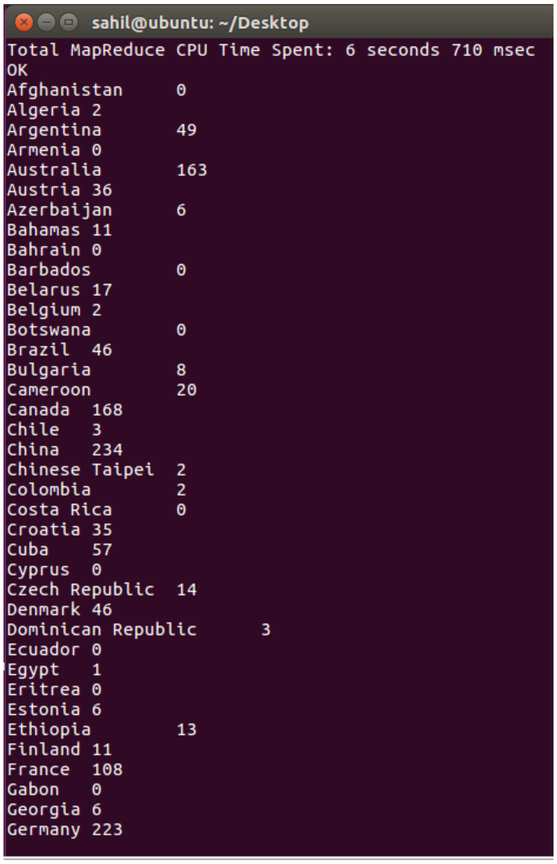
```
Switzerland	93
Syria	1
Tajikistan	3
Thailand	18
Togo	1
Trinidad and Tobago	19
Tunisia	4
Turkey	28
Uganda	1
Ukraine	143
United Arab Emirates	1
United States	1312
Uruguay	1
Uzbekistan	19
Venezuela	4
Vietnam	2
Zimbabwe	7
```

4. Write a Hive program to find the number of gold medals each country won.
   Commands of Problem 4
   hive> select country,SUM(gold) from olympic GROUP BY country;



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

Result:-



```
Total MapReduce CPU Time Spent: 6 seconds 710 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
Germany 223
```

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