ASSIGNMENT 09

Dataset 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

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.

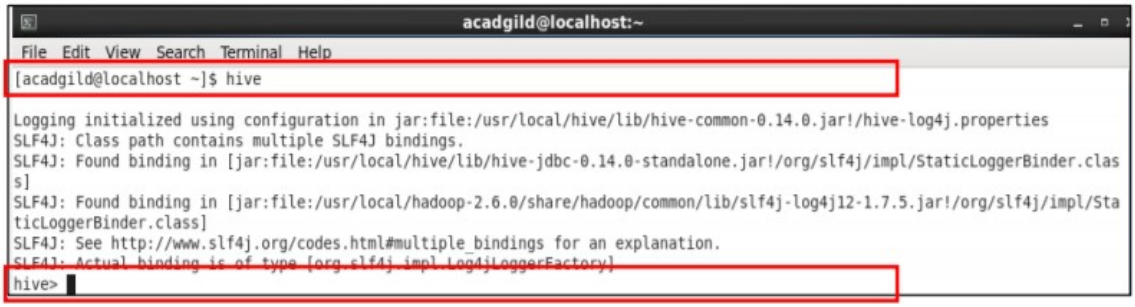
4. Write a Hive program to find the number of gold medals each country won.

Solution:

Input File The input file is present on the local file system as follows:

```
[acadgild@localhost hive]$ ls -l
total 552
-rw-rw-r--. 1 acadgild acadgild   3058 Sep 19 08:57 commands
-rw-rw-r--. 1 acadgild acadgild   2805 Sep 18 22:16 commands~
-rw-rw-r--. 1 acadgild acadgild    170 Sep 17 14:17 complexData
-rw-rw-r--. 1 acadgild acadgild    437 Sep 16 19:29 dataset_Session14.txt
-rw-rw-r--. 1 acadgild acadgild    159 Sep 19 08:49 emp_Details
-rw-rw-r--. 1 acadgild acadgild    159 Sep 19 08:24 emp_Details~
-rw-rw-r--. 1 acadgild acadgild     84 Sep 17 13:43 empDetails~
-rw-rw-r--. 1 acadgild acadgild    107 Sep 18 22:00 employee.csv
-rw-rw-r--. 1 acadgild acadgild    107 Sep 18 21:51 employee.csv~
-rw-rw-r--. 1 acadgild acadgild 518669 Sep 19 22:14 olympix data.csv
drwxrwxr-x. 2 acadgild acadgild   4096 Sep 19 08:53 output
-rw-rw-r--. 1 acadgild acadgild    170 Sep 17 14:17 Unsaved Document 1~
[acadgild@localhost hive]$
```

Start hive: We start the hive command line by executing the command hive as shown below:



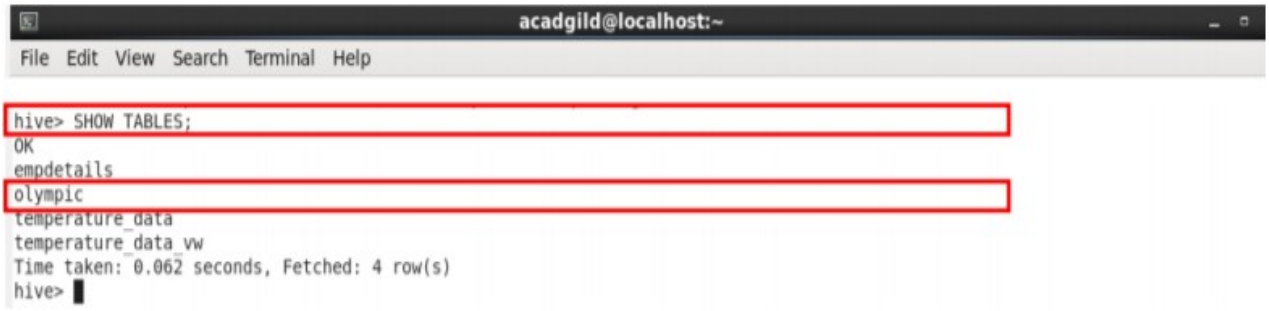
The above snapshot also shows that hive prompt has started. A pre-requisite to use hive is to start mysql server. This was done using the command sudo service mysqld start.

Step 1: We use **SHOW DATABASES** command to list the databases present. The database we will be using is custom as shown below:

Step 2: We use **USE** custom command to make use of custom database, as shown below:

Step 3: We create the table using **CREATE TABLE** command. The fields of the table are as mentioned above in the description of dataset. The fields are separated by the delimiter ‘ ’.

Step 4: **SHOW TABLES** command will help us verify that the table is created.



Step 5: Next is to load the data from input file, which is located at /home/acadgild/suman/hive as follows. We use the LOAD command and use the keyword LOCAL to specify that the file is present in the local file system and not HDFS.

Problem Statement 1:

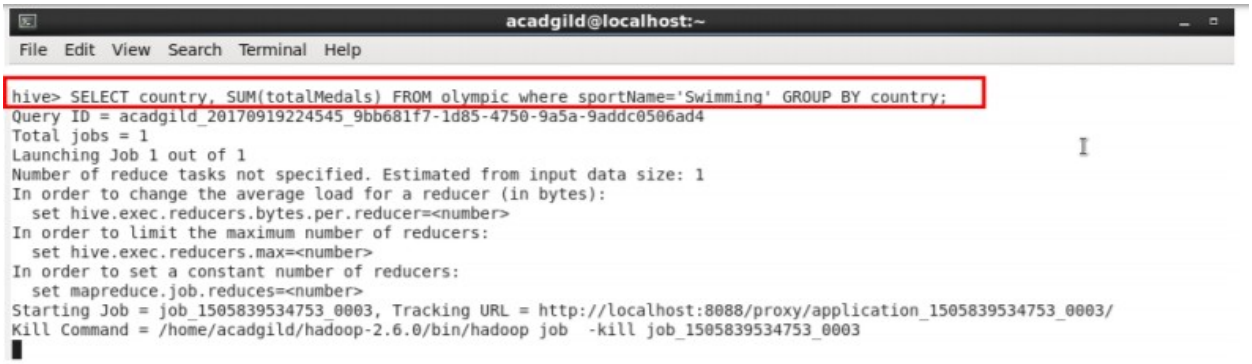
Write a Hive program to find the number of medals won by each country in swimming.

Solution :

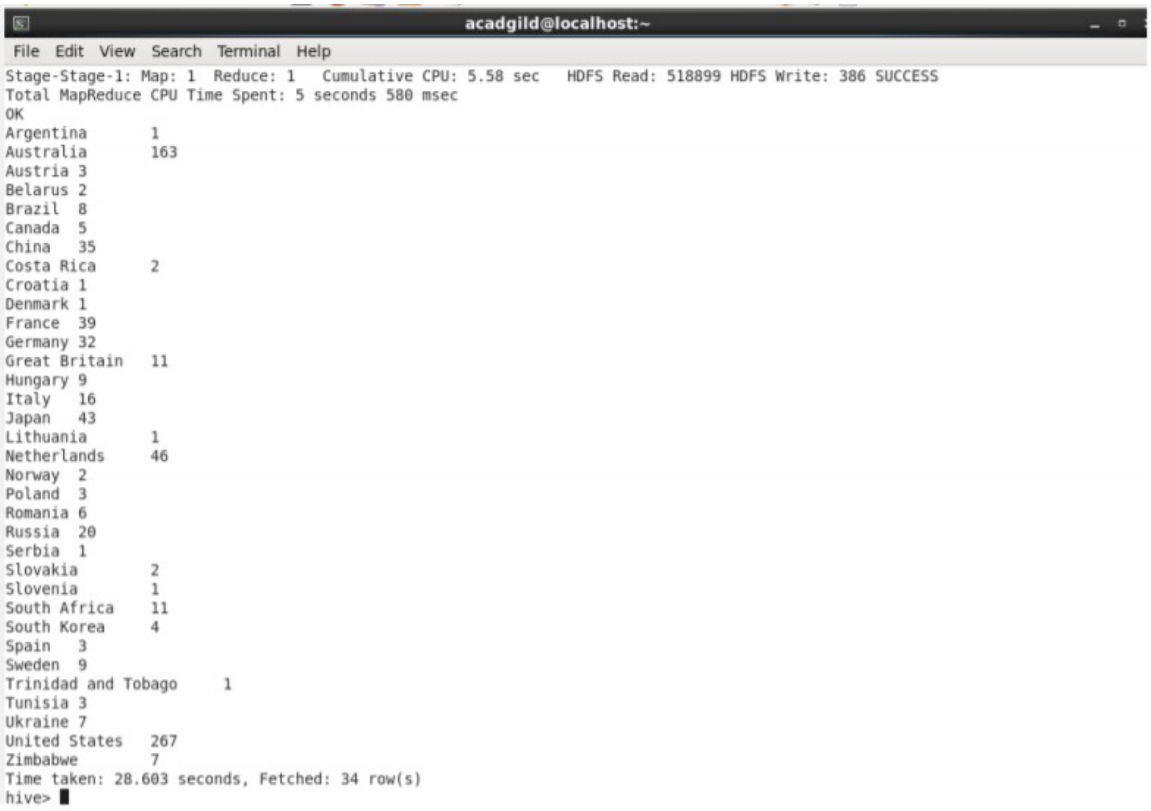
The query used to solve the above problem statement is

SELECT country, SUM(totalMedals) FROM olympic where sportName='Swimming' GROUP BY country;

We have used the predicate sportName='Swimming' to get details only for Swimming and used GROUP BY to get details per country as follows:



The output of the query is as follows:



Problem Statement 2:

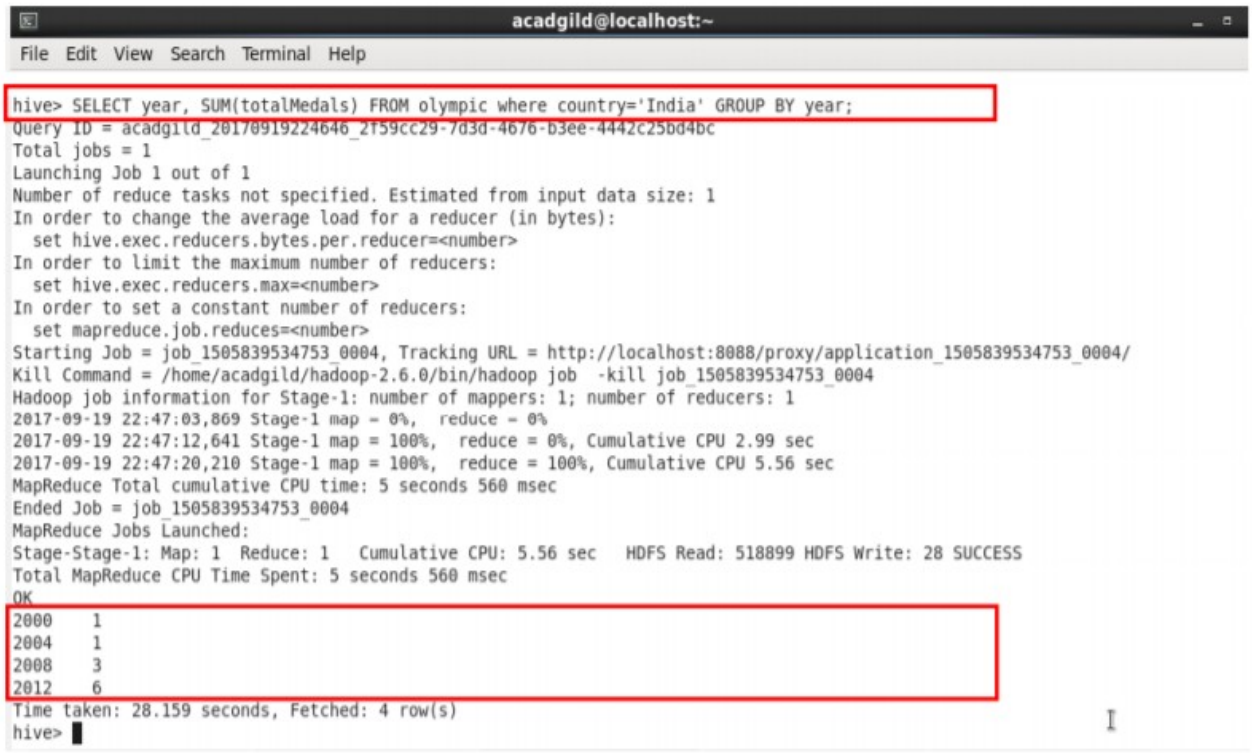
Write a Hive program to find the number of medals that India won year wise.

Solution:

The query used to solve the above problem statement is

SELECT year, SUM(totalMedals) FROM olympic where country='India' GROUP BY year;

We have used the predicate country='India' to get details only for India and used GROUP BY to get details per year as follows:



The output is shown in the above snapshot.

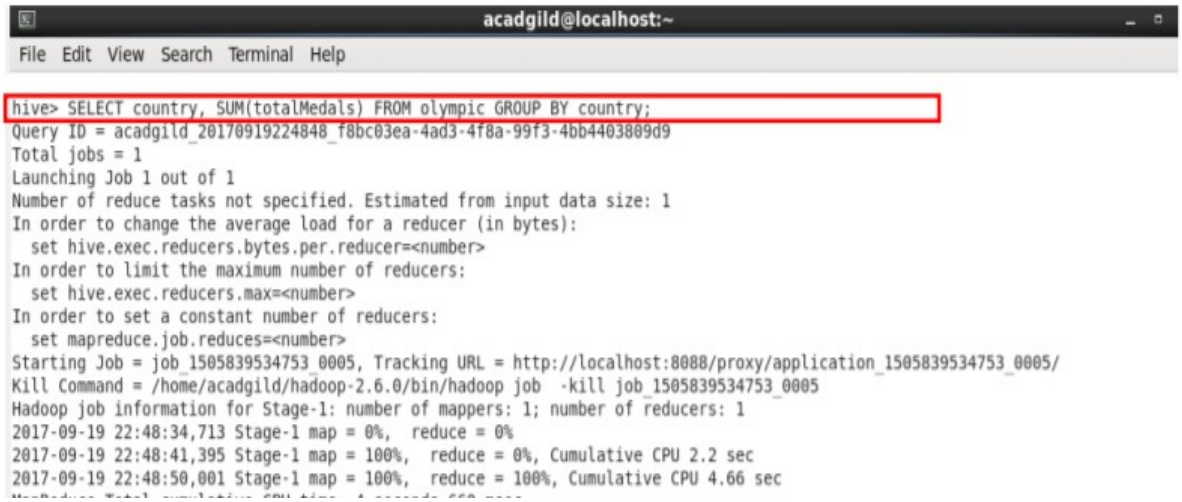
Problem Statement 3:

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

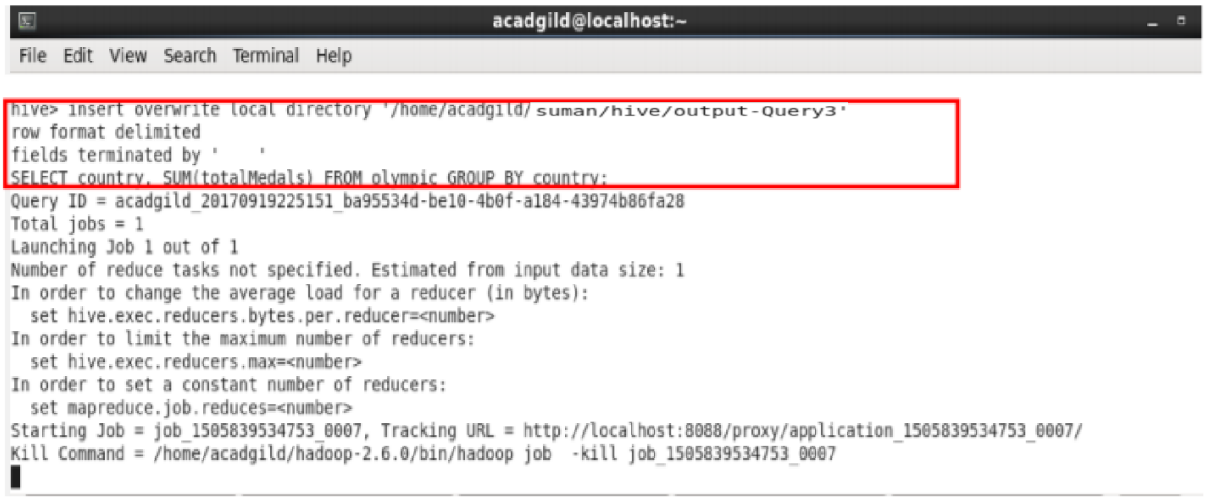
Solution:

The query used to solve the above problem statement is

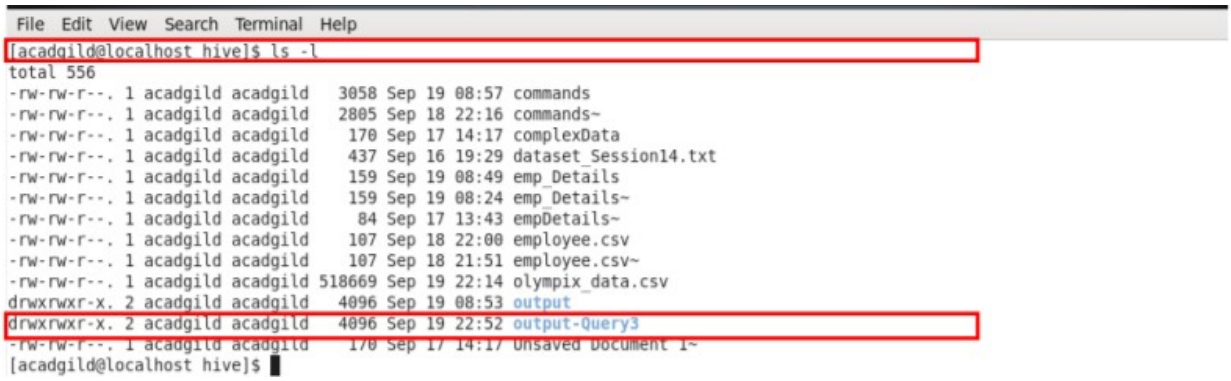
SELECT country, SUM(goldMedal) FROM olympic GROUP BY country; We have used GROUP BY to get details per country as follows:



We can also store this result into a file using INSERT command as follows. The delimiter used to separate the fields in the file is ‘



We can see the new directory created named output as follows:



The output directory has the following output file.

Problem Statement 4:

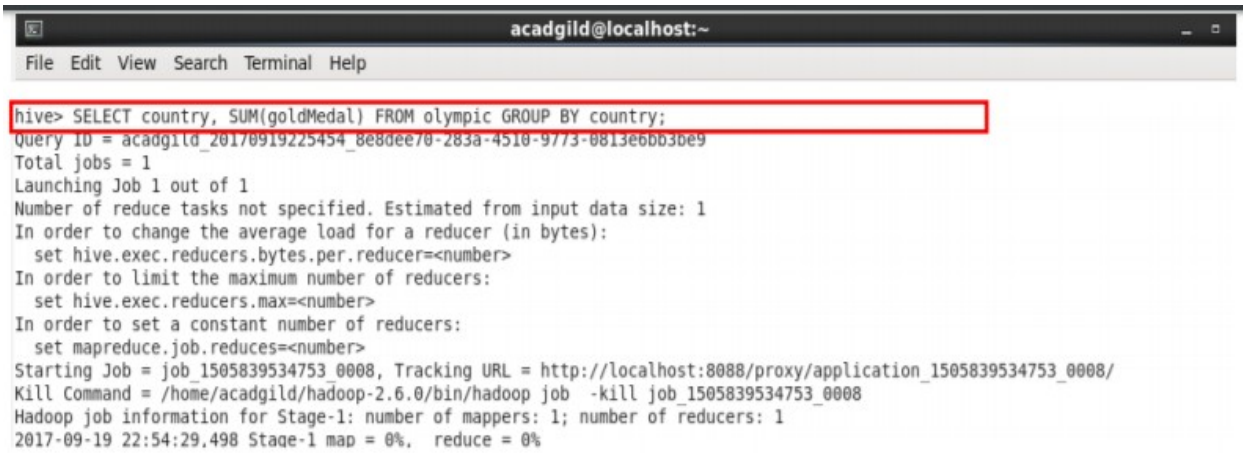
Write a Hive Program to find the total number of medals each country won.

Solution:

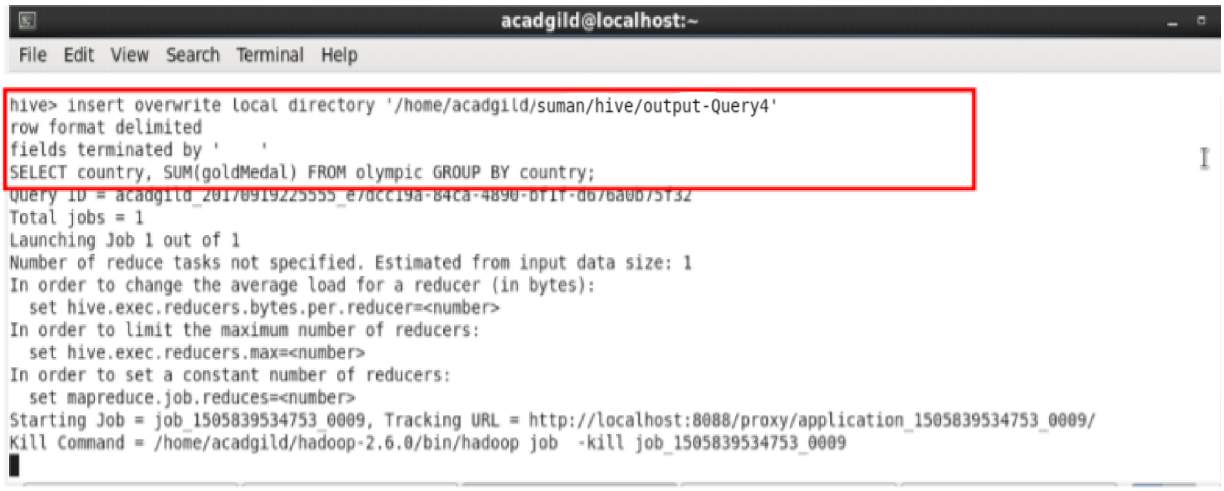
The query used to solve the above problem statement is

SELECT country, SUM(goldMedal) FROM olympic GROUP BY country;

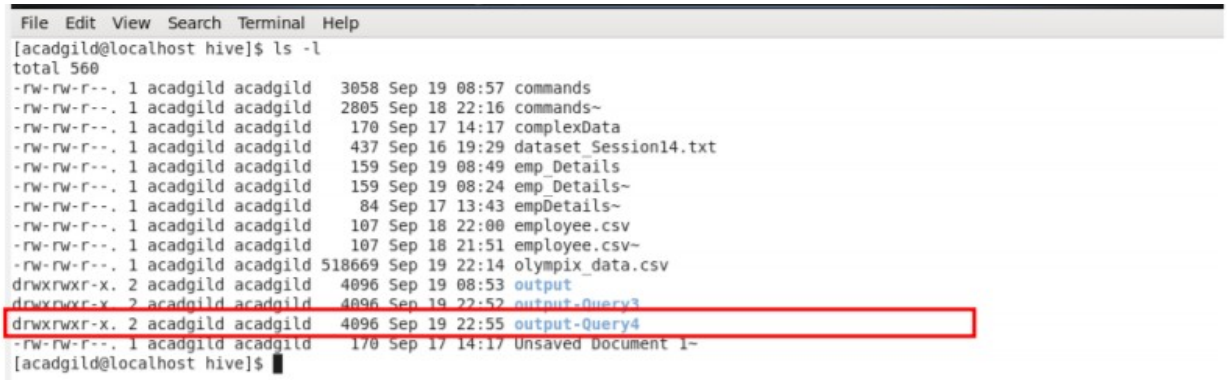
We have used GROUP BY to get details per country and used SUM function to get the total as follows:



We can also store this result into a file using INSERT command as follows. The output directory is /home/acadgild/suman/hive/output-Query4. The delimiter used to separate the fields in the file is ‘



We can see the new directory created named output as follows:



The output directory has the following output file.