# Big Data Hadoop Training

Session 15 Assignment 2 Solution:

**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

1. I have placed the dataset in HDFS @ /tmp/hive/BDDS15A2 directory.

Now, create a table olympix\_data inside a database say, BDDS15A2.

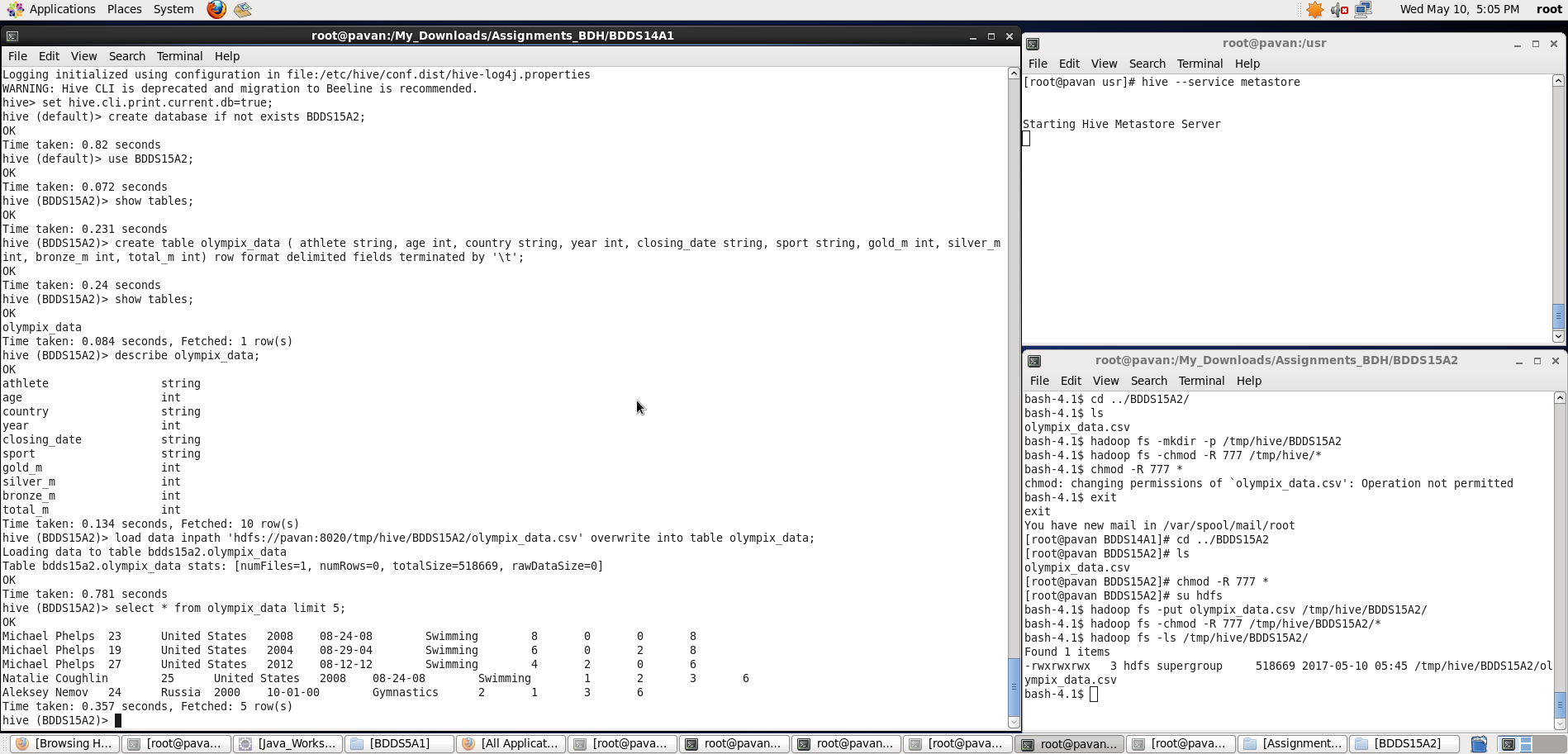
**hive> create table olympix\_data ( athlete string, age int, country string, year int, closing\_date string, sport string, gold\_m int, silver\_m int, bronze\_m int, total\_m int) row format delimited fields terminated by ',';**

We can see a table being created. Use **show** and **describe** commands to see the table.

Now, load the dataset into the table:

**hive> load data inpath 'hdfs://pavan:8020/tmp/hive/BDDS15A2/olympix\_data.csv' overwrite into table olympix\_data;**

**hive> select \* from olympix\_data limit 5; -- Used to view the first 5 contents of the table**

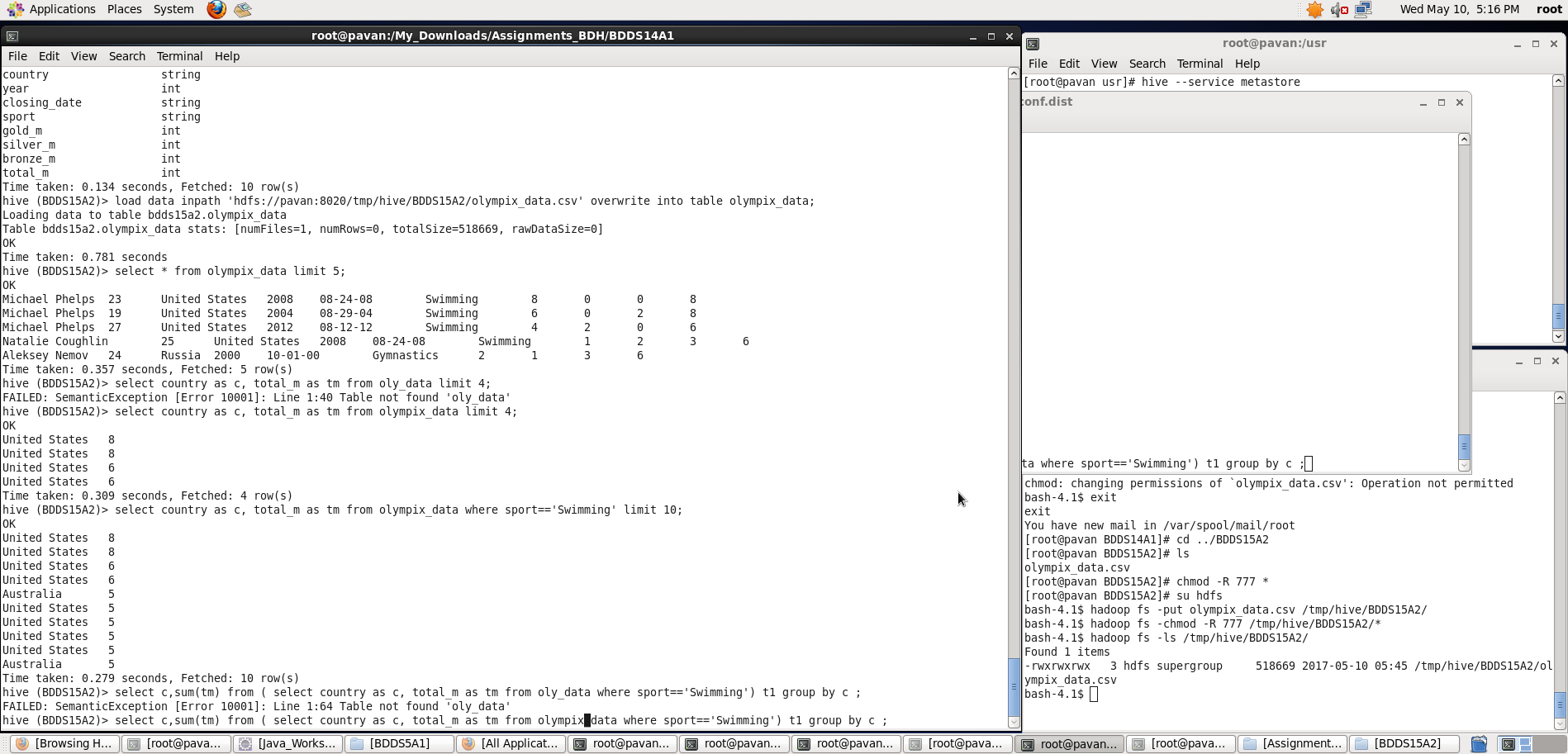


Once, we have successfully loaded the data into table, we will perform the below specified tasks

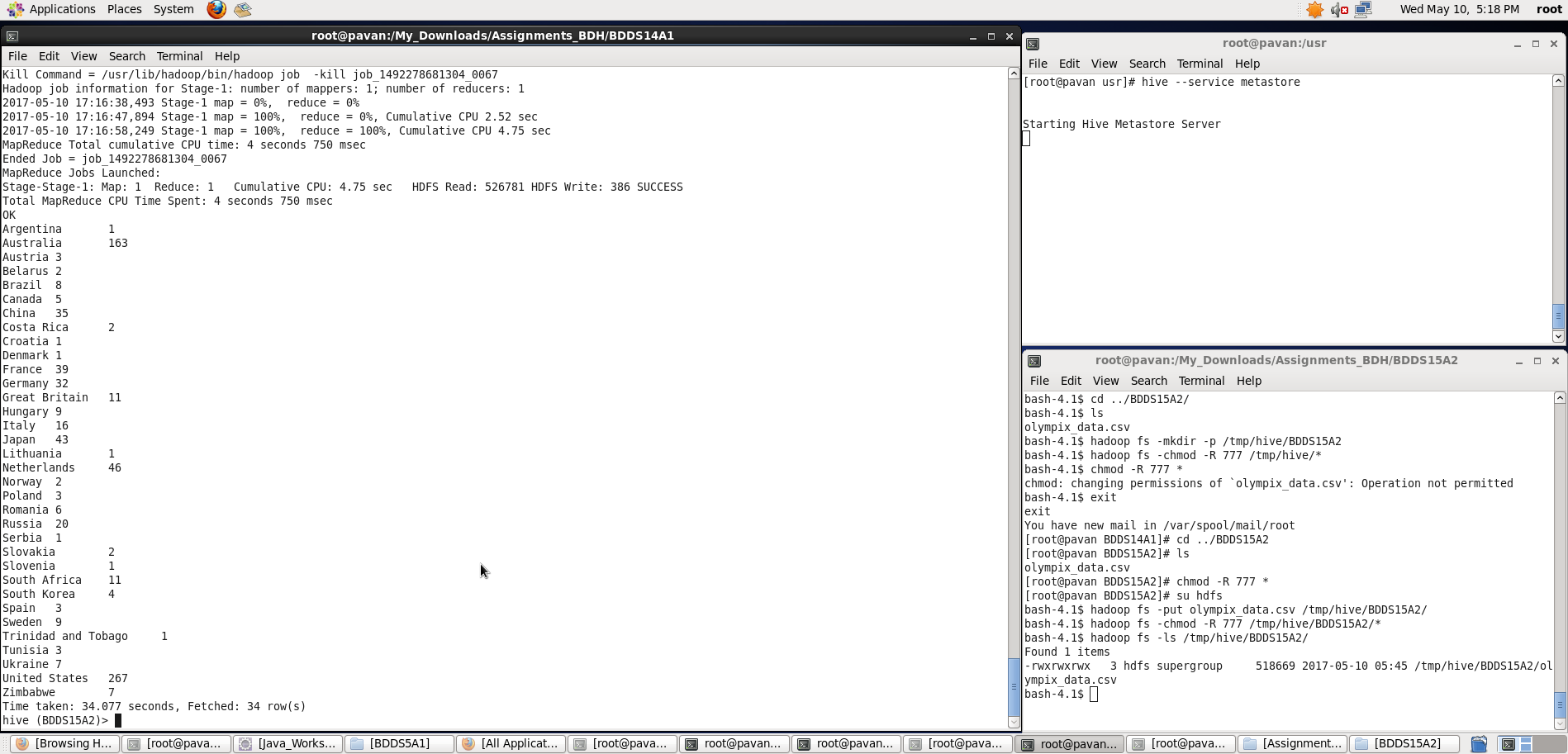
**1. Write a Hive program to find the number of medals won by each country in swimming.**

**A.** First, we filter out the records based on sport==’Swimming’ then group by country and then find sum of all total\_medals column corresponding to that country using **sum()** function.

**hive>** **select c,sum(tm) from ( select country as c, total\_m as tm from olympix\_data where sport=='Swimming') t1 group by c ;**



Output can be seen as below:

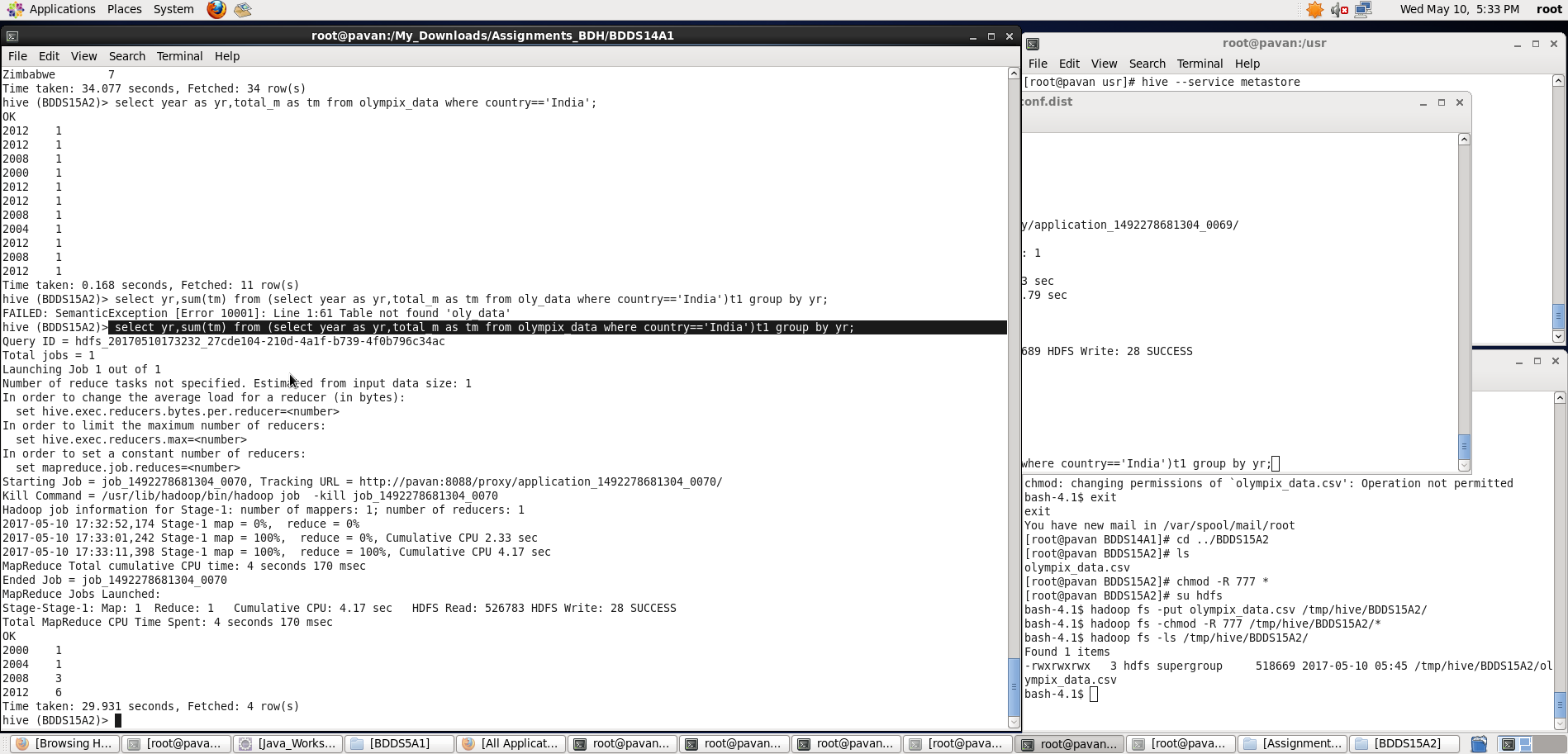


**2. Write a Hive program to find the number of medals that India won year wise**.

**A.** First, we filter out the records based on country==’India’ then group by year and then find sum of all total\_medals column corresponding to that year using **sum()** function

**hive>** **select yr,sum(tm) from (select year as yr,total\_m as tm from olympix\_data where country=='India')t1 group by yr;**

**Output can be seen as below:**

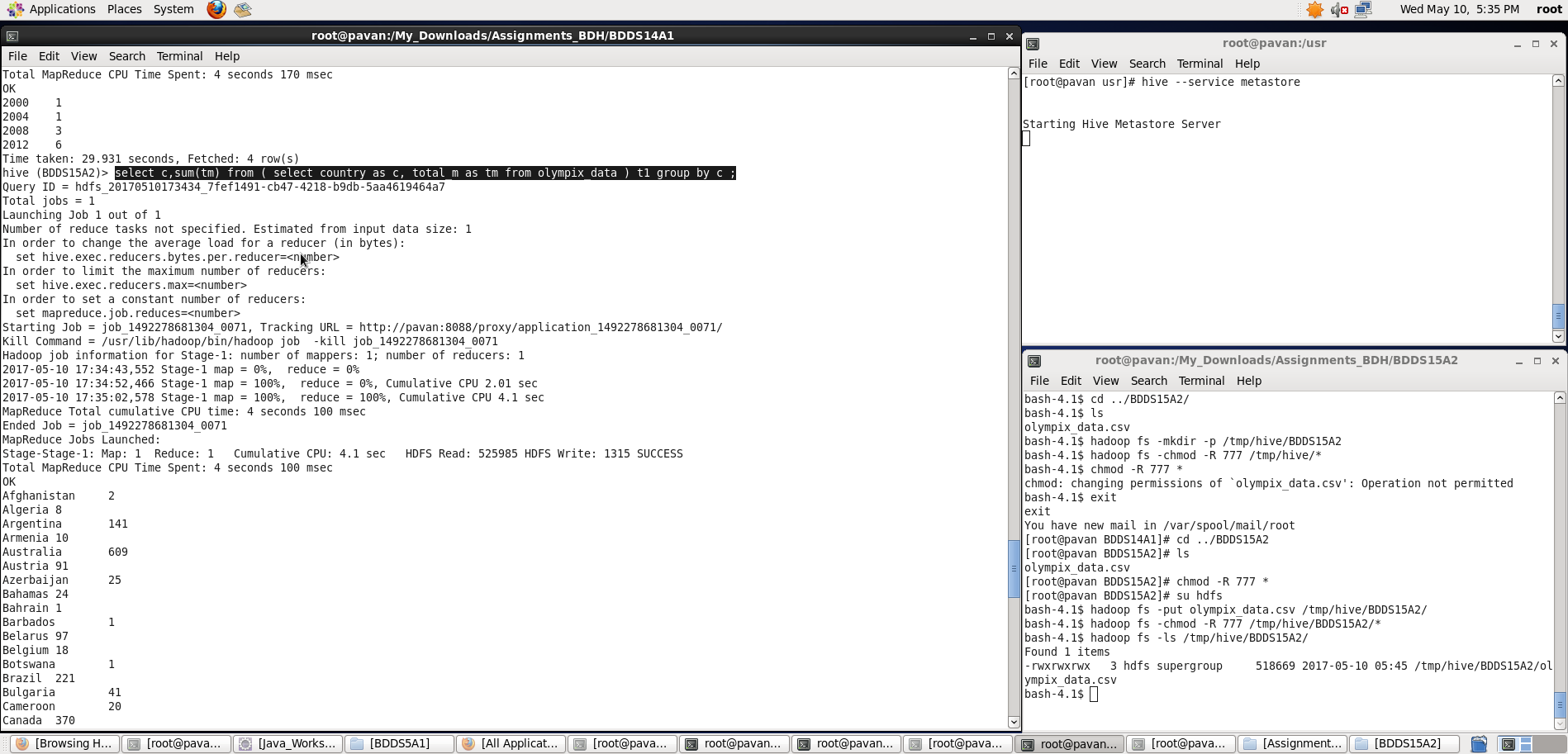


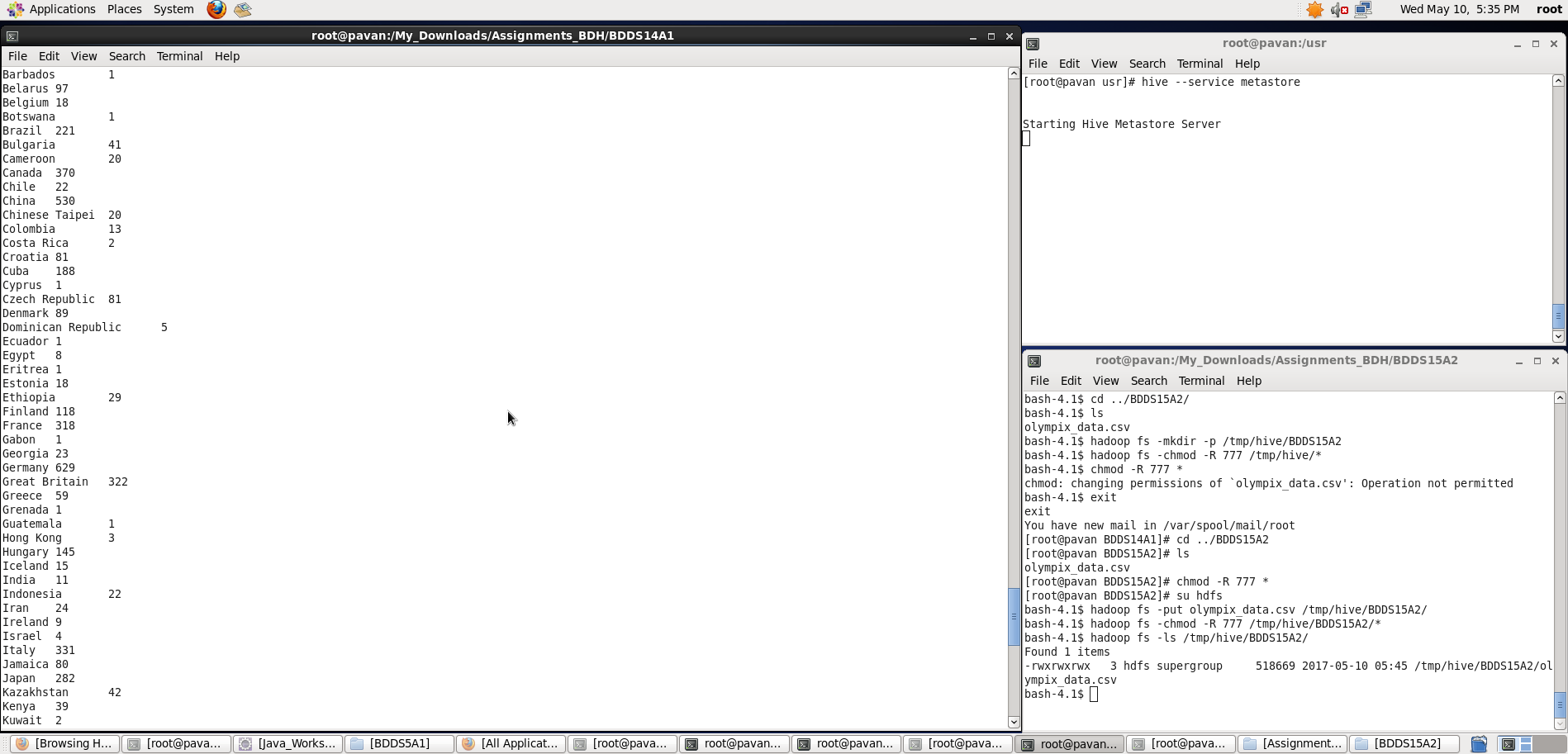
**3. Write a Hive Program to find the total number of medals each country won.**

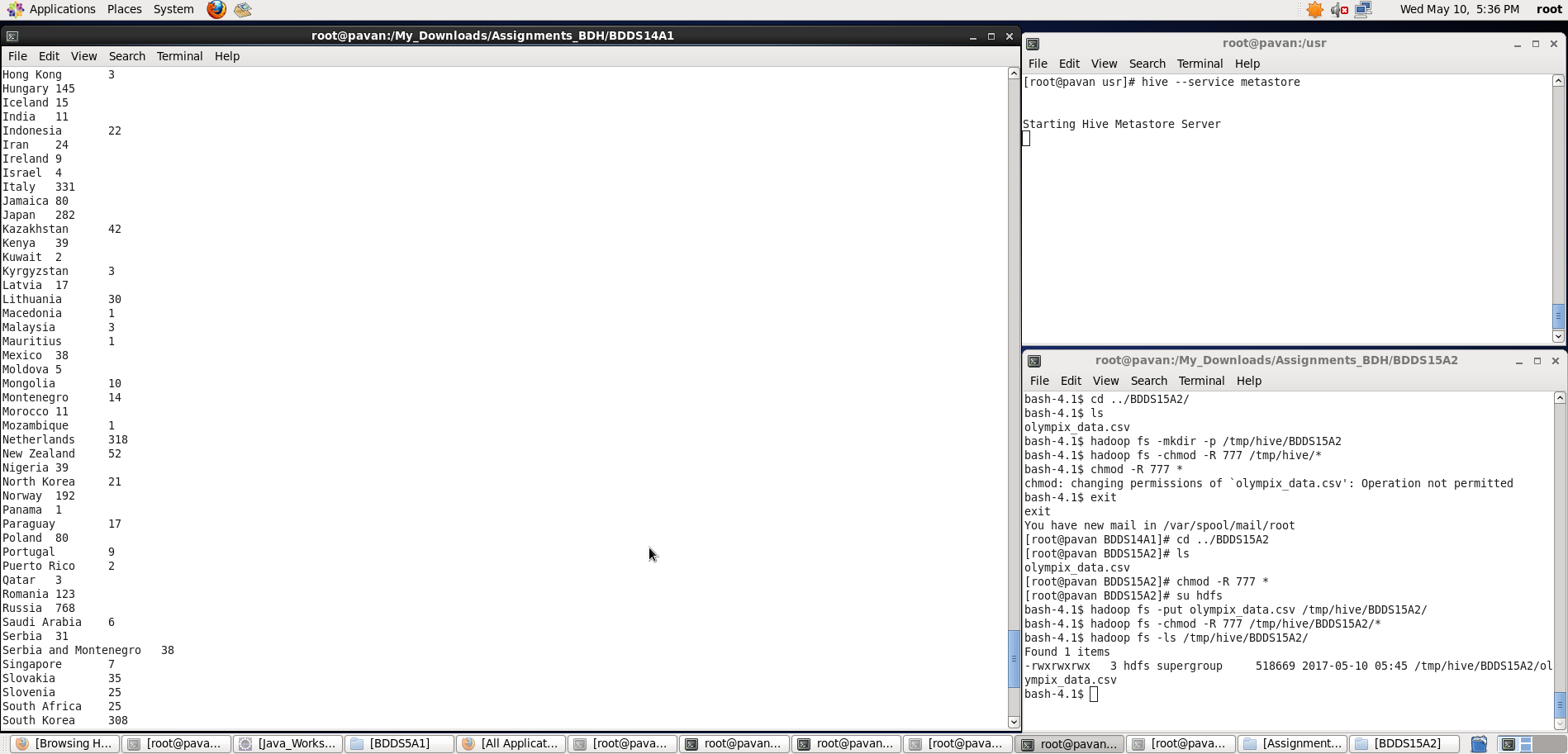
**A.** First, we will group the records by country and then find sum of all total\_medals column corresponding to that country using **sum()** function

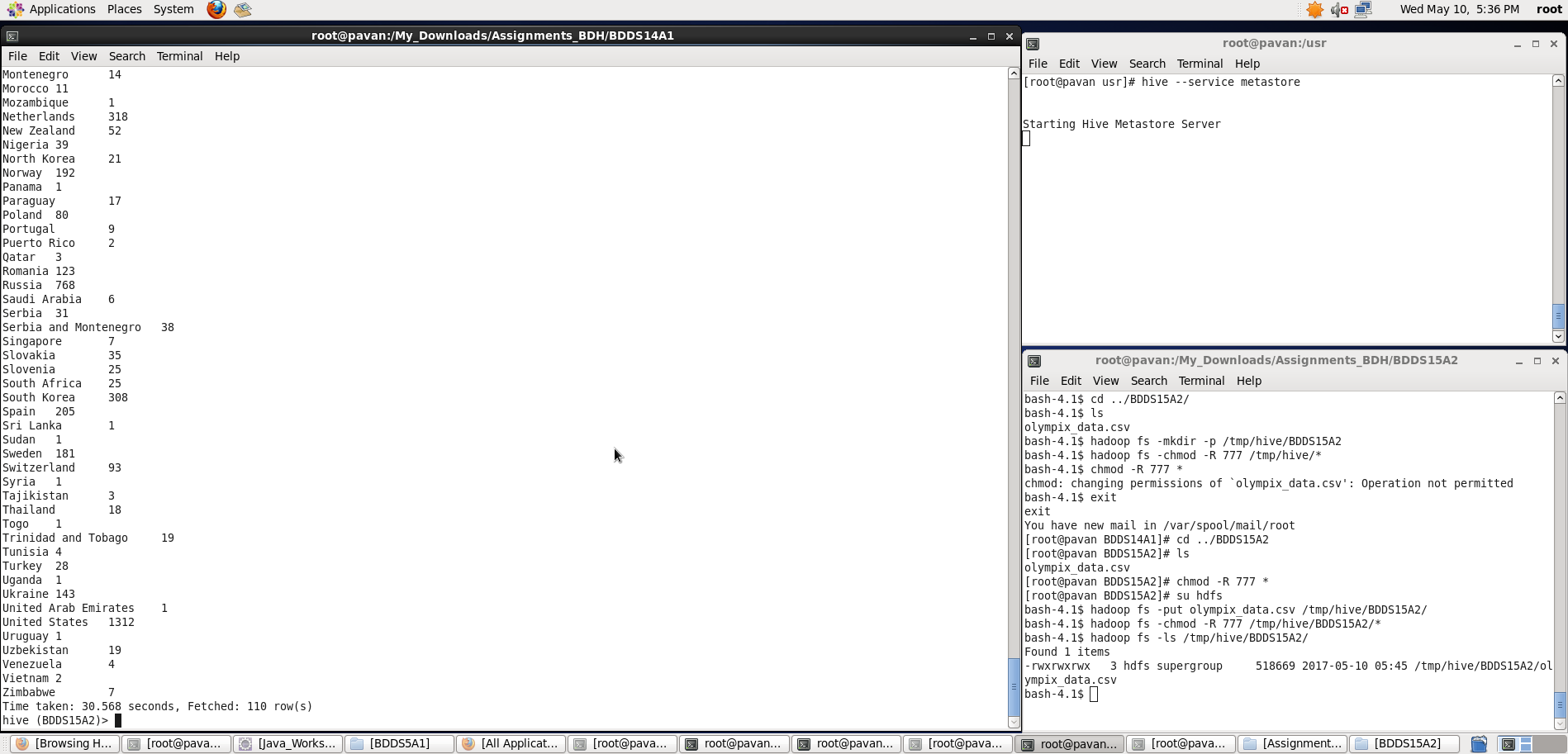
**hive>** **select c,sum(tm) from ( select country as c, total\_m as tm from olympix\_data ) t1 group by c ;**

**Output can be seen as below:**







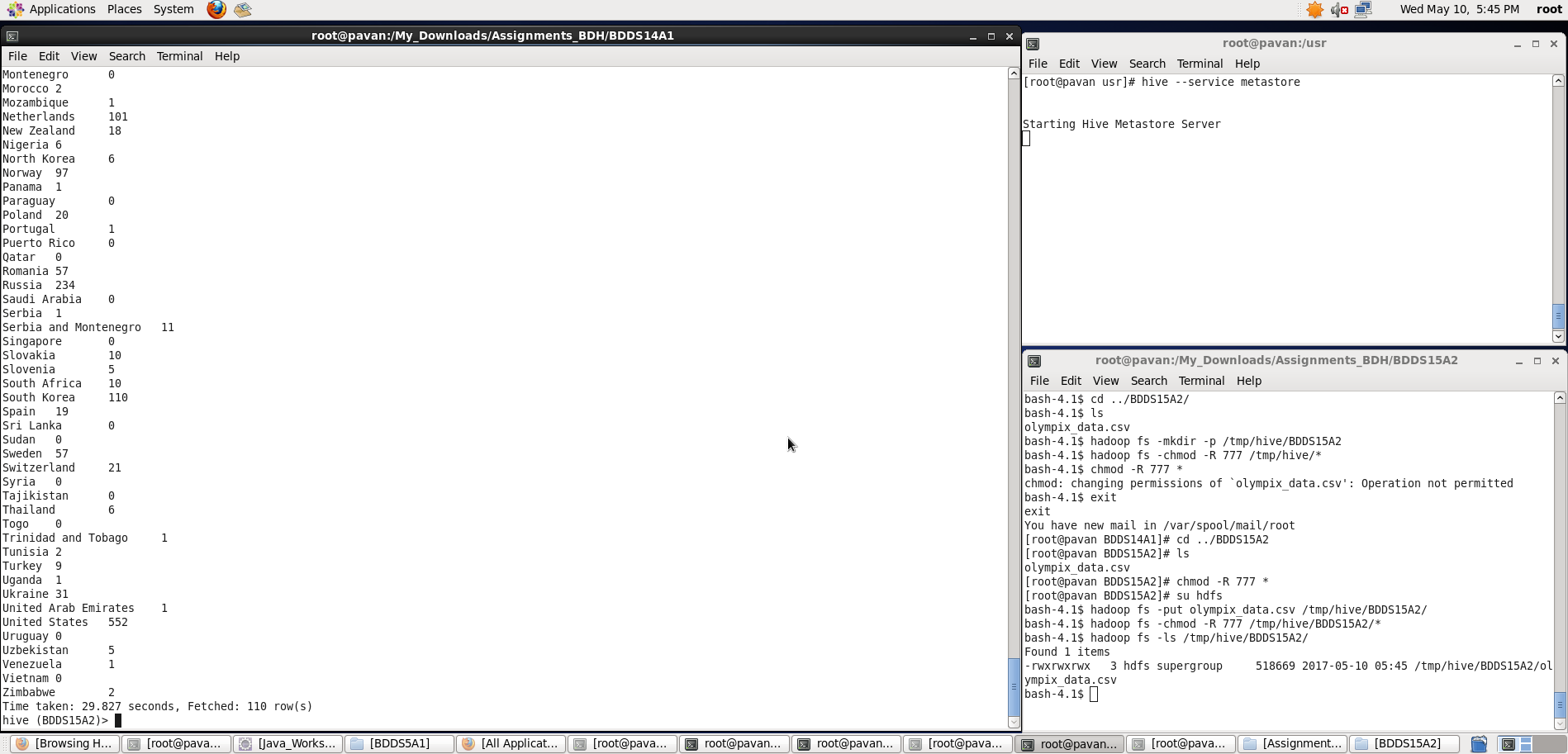
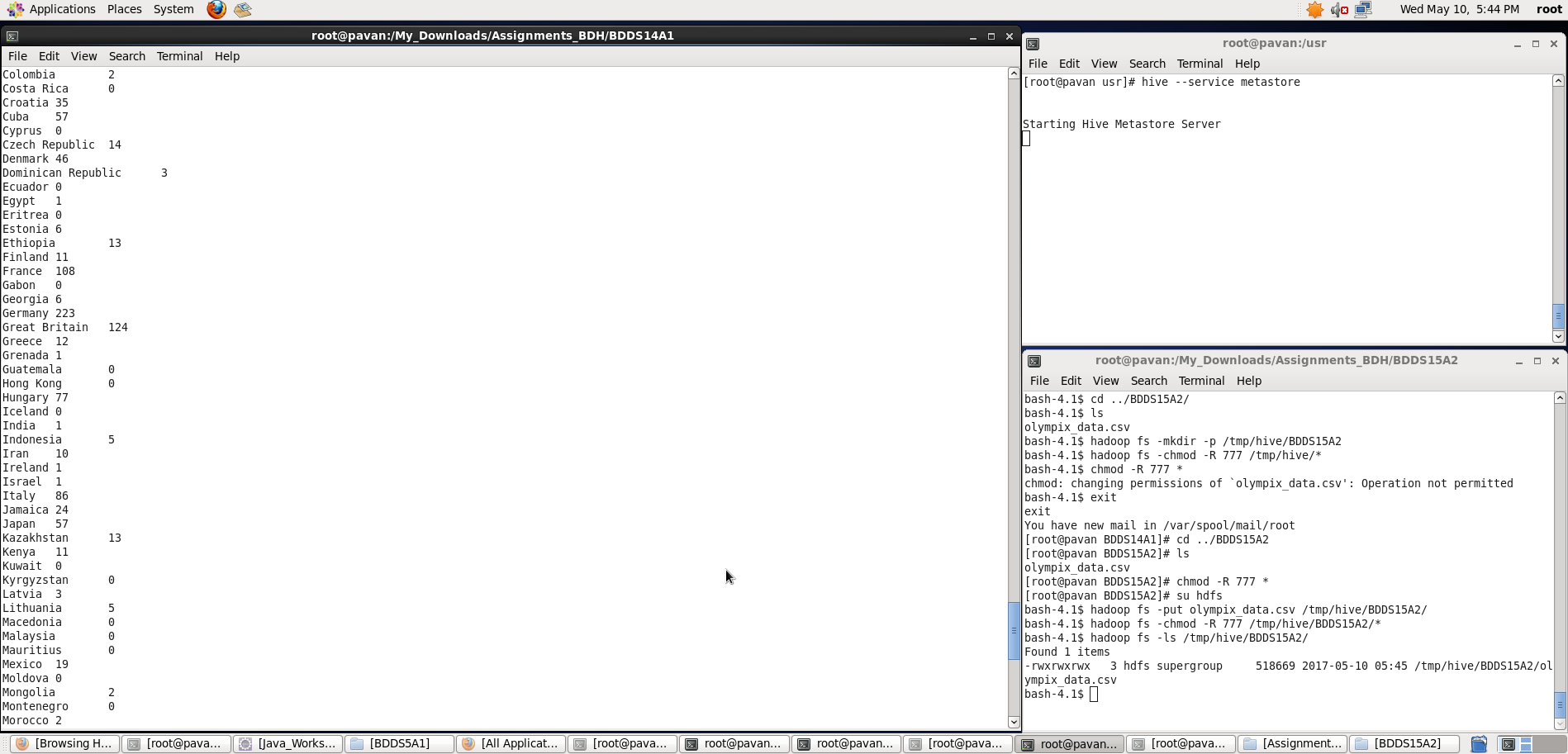
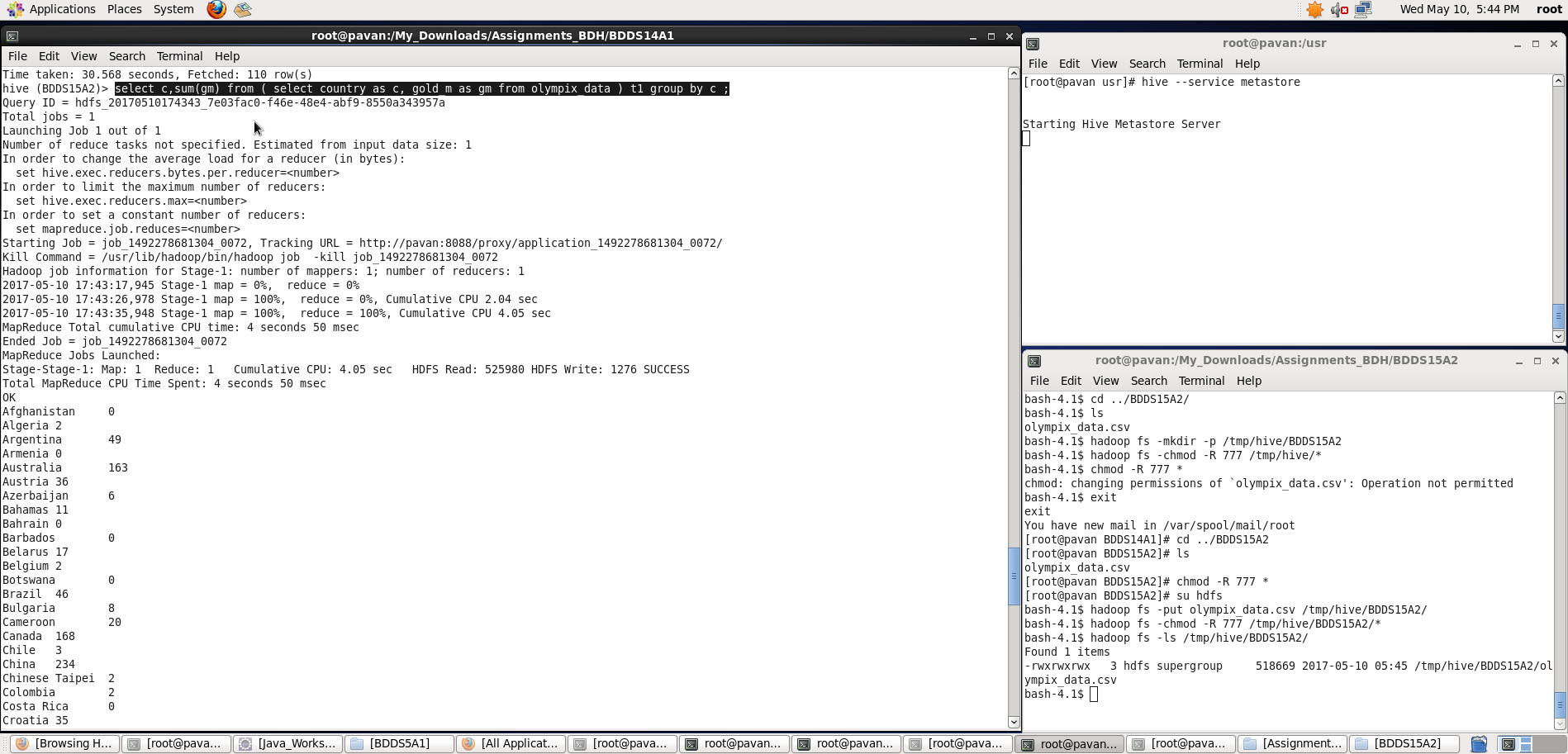


**4. Write a Hive program to find the number of gold medals each country won.**

1. First, we will group the records by country and then find sum of all gold\_medals column corresponding to that country using **sum()** function

**hive>** **select c,sum(gm) from ( select country as c, gold\_m as gm from olympix\_data ) t1 group by c ;**

**Output can be seen as below:**



Thus, we have performed required tasks on the dataset which is loaded into a hive table.