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

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

DATA 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

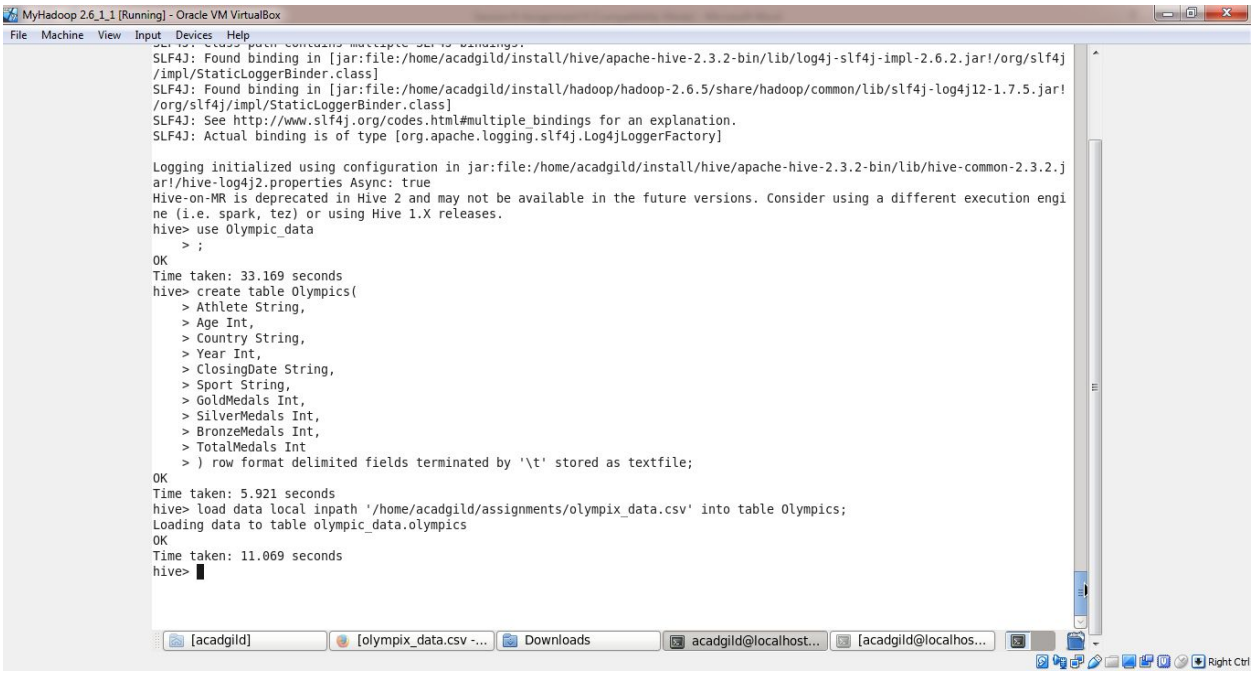
Gold Medals: No. of Gold medals

Silver Medals: No. of Silver medals

Bronze Medals: No. of Bronze medals

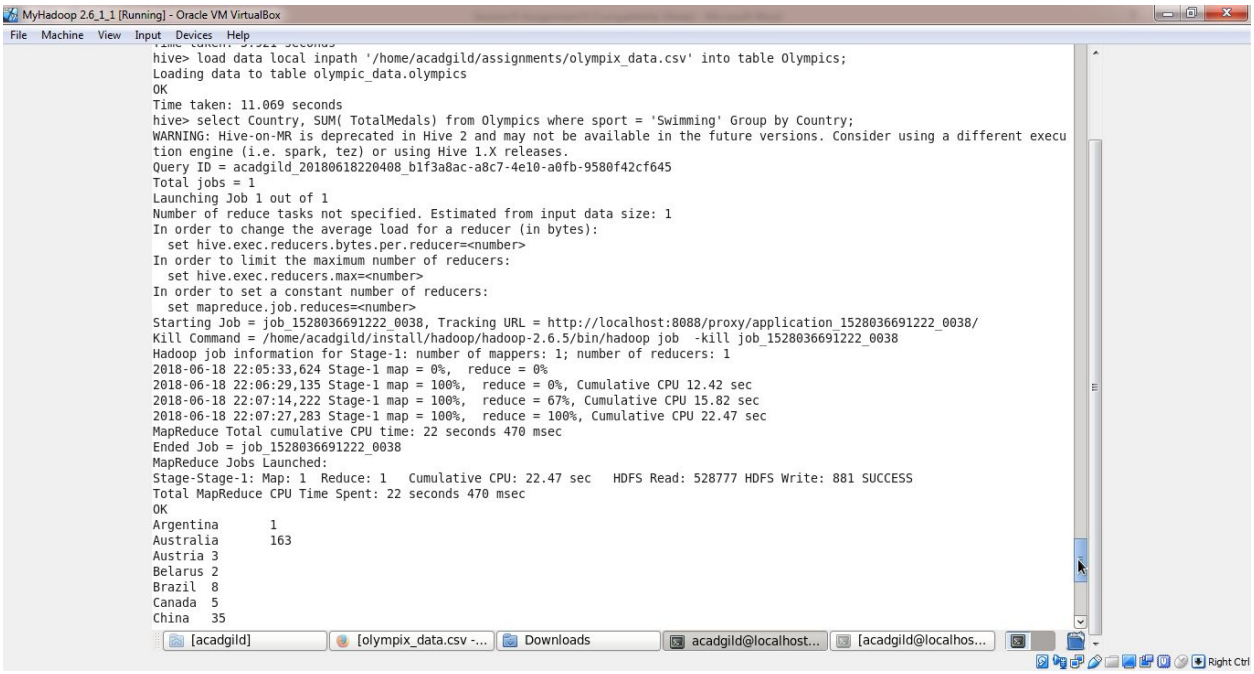
Total Medals: Consists of total no. of medals

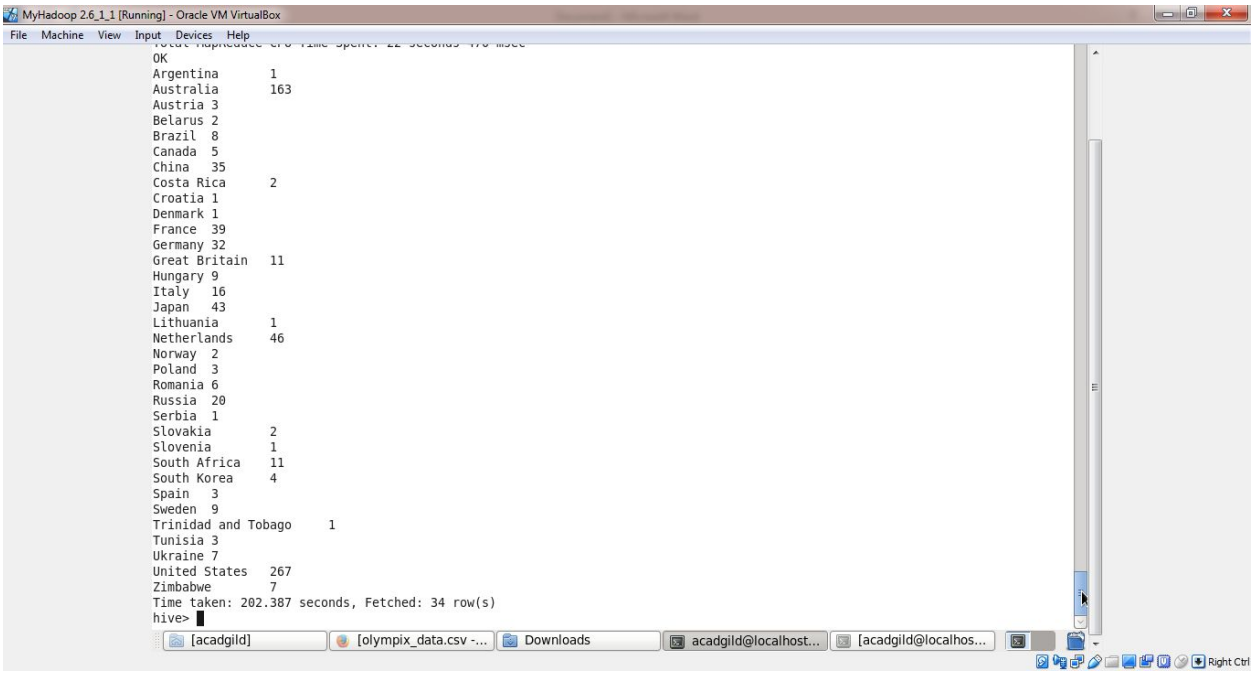
Load input data into db:



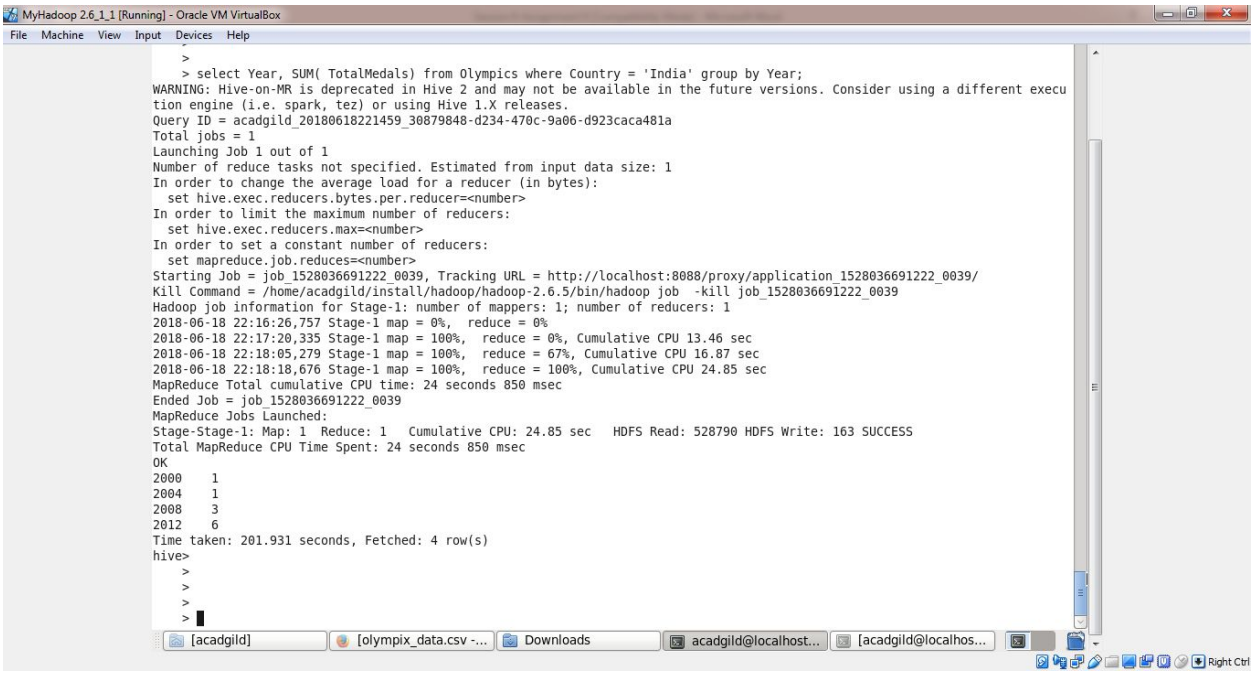
Task 1

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

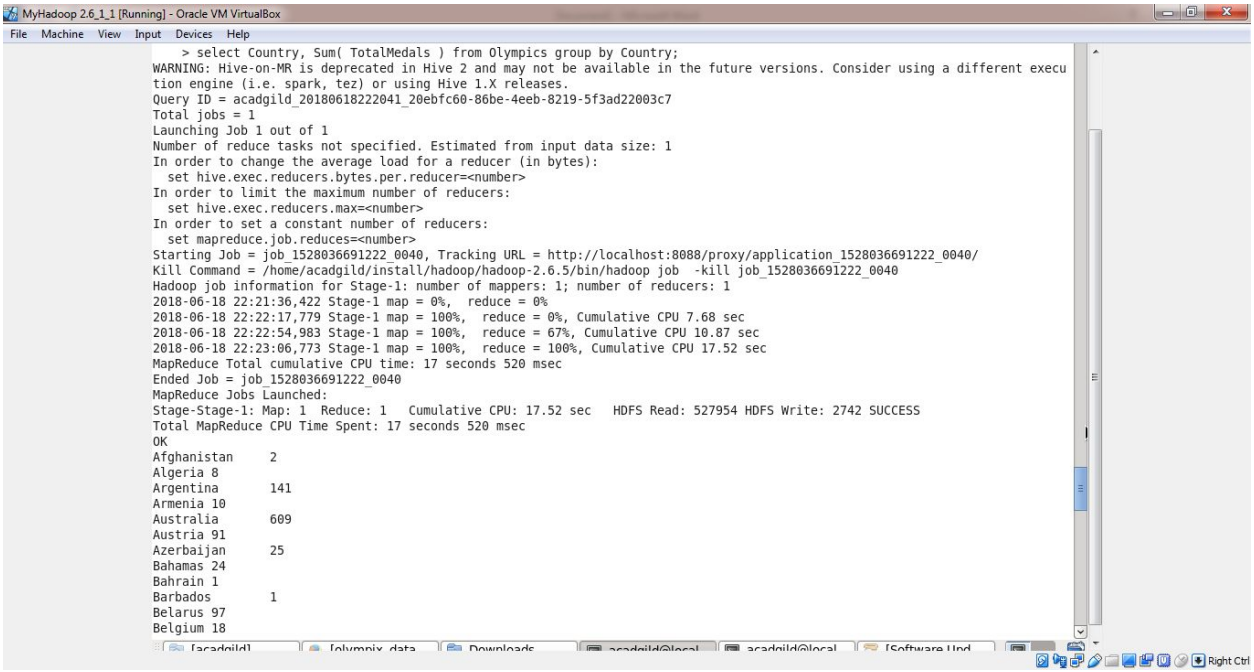


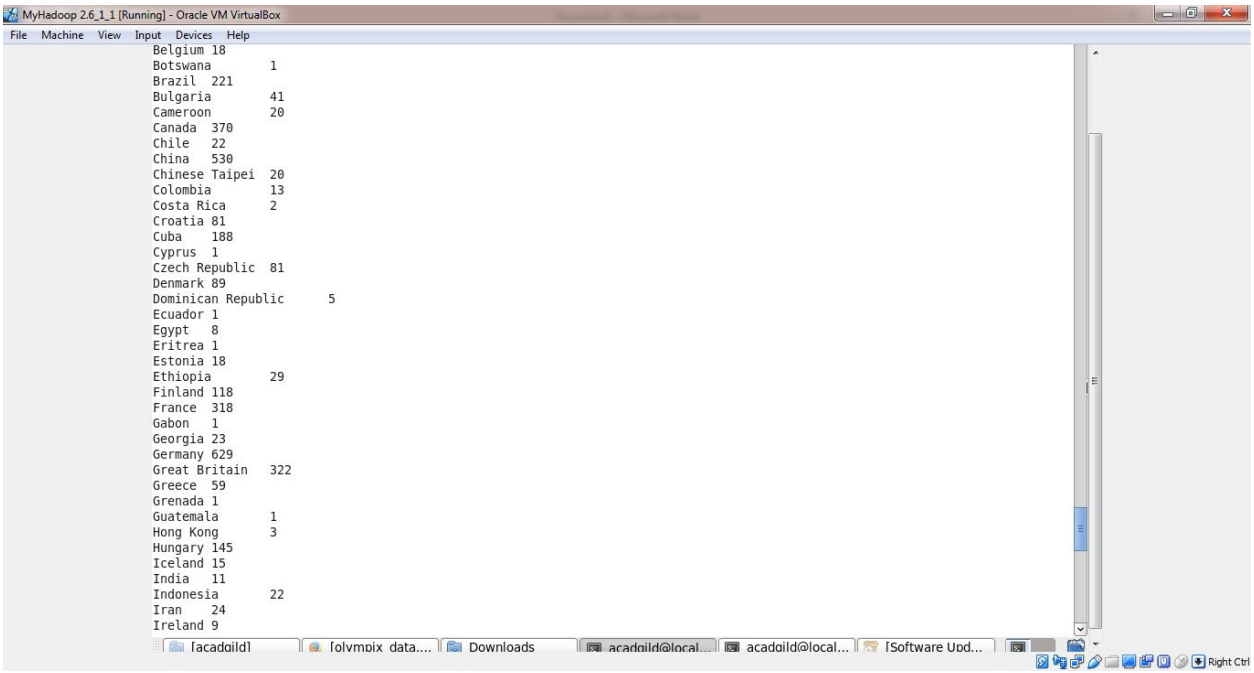


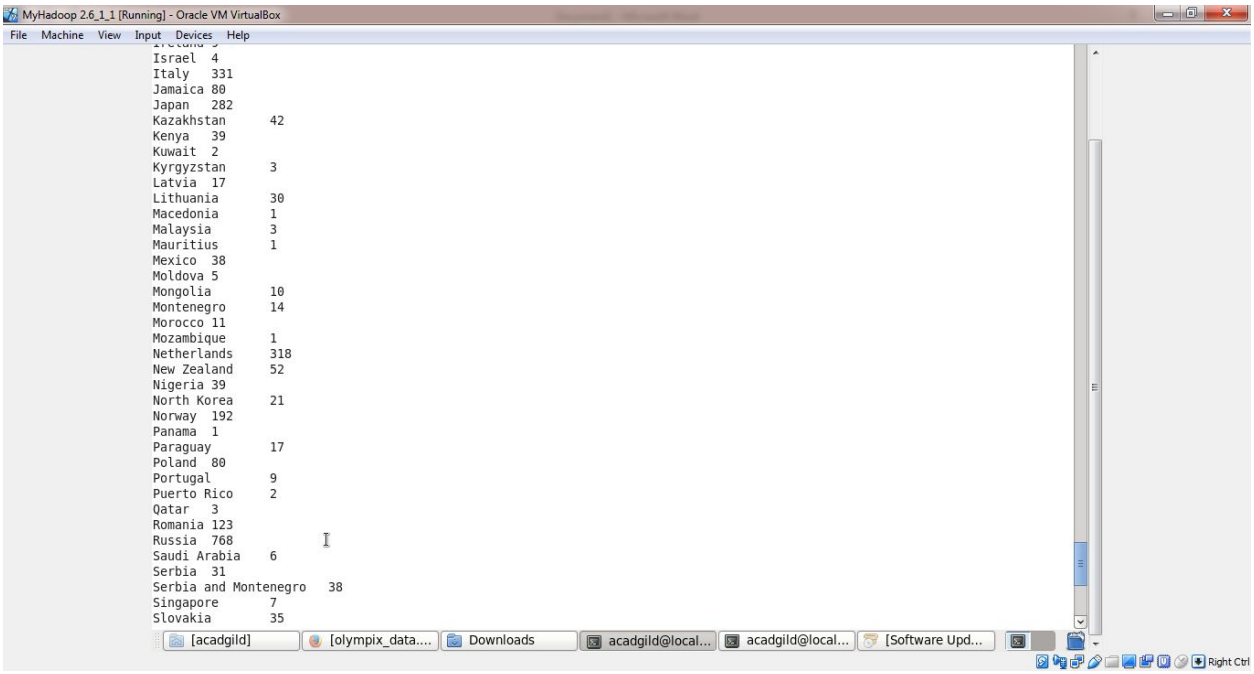
2. Write a Hive program to find the number of medals that India won year wise.

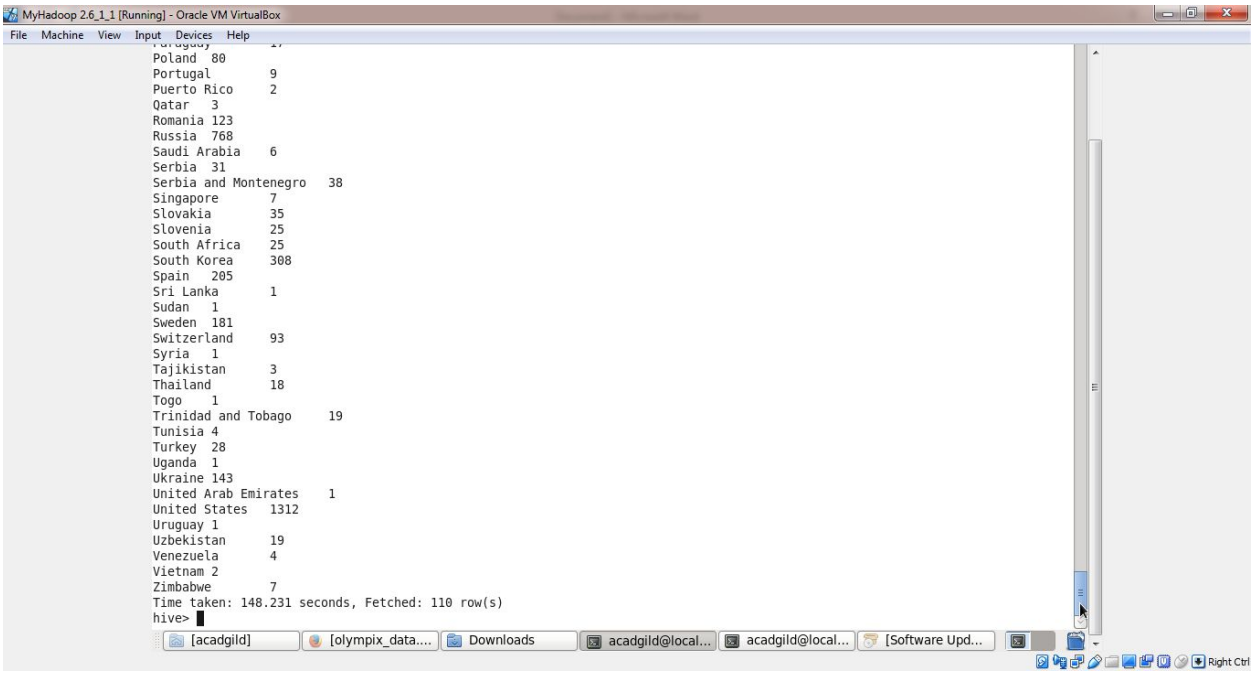


3. Write a Hive Program to find the total number of medals each country won.

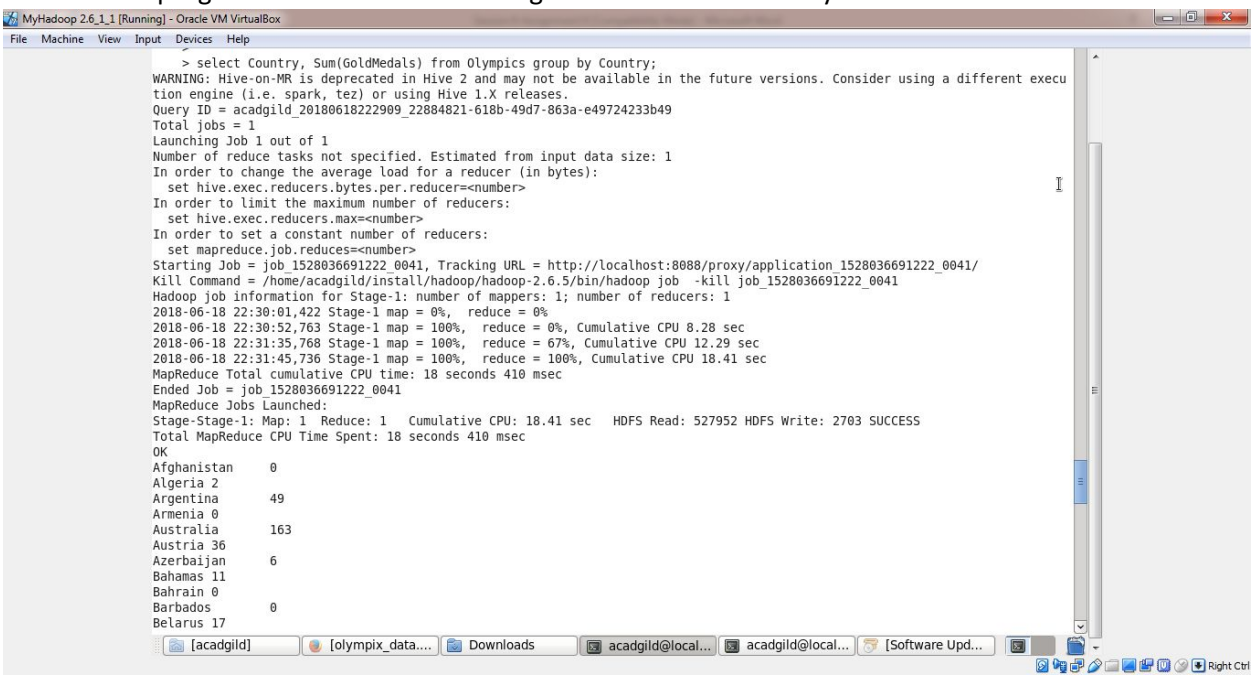


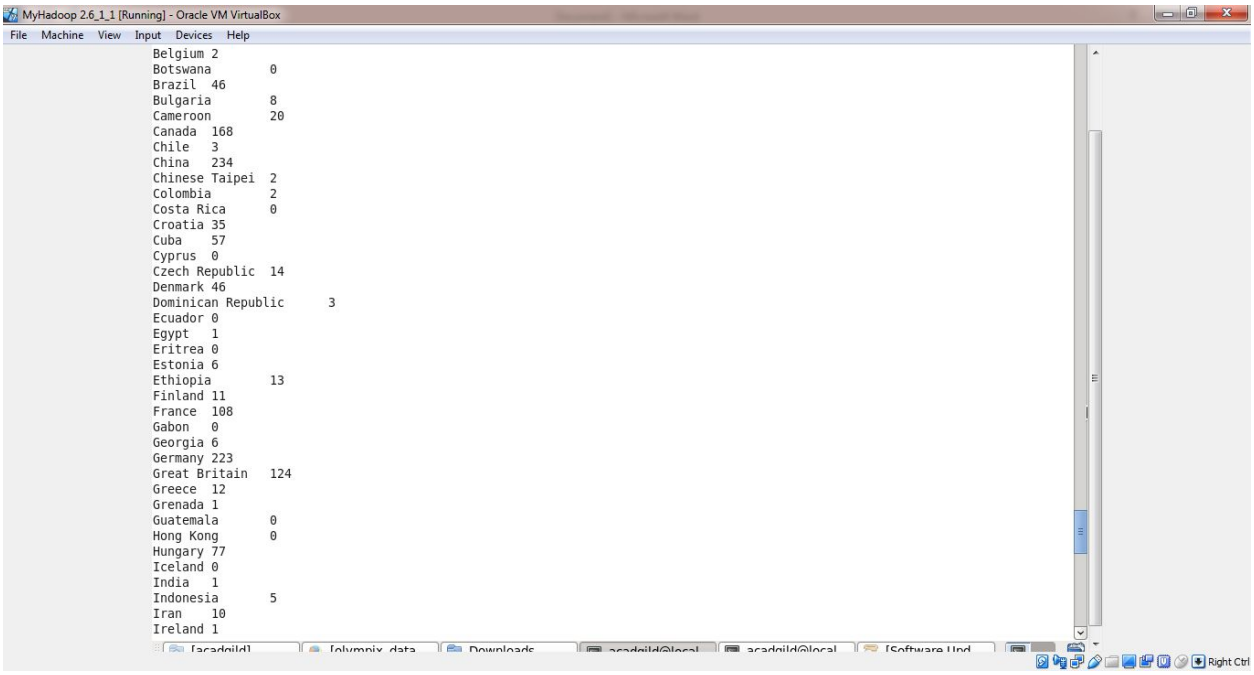

MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Belgium 18
Botswana 1
Brazil 221
Bulgaria 41
Cameroon 20
Canada 370
Chile 22
China 530
Chinese Taipei 20
Colombia 13
Costa Rica 2
Croatia 81
Cuba 188
Cyprus 1
Czech Republic 81
Denmark 89
Dominican Republic 5
Ecuador 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
[acadgild]
[olympix_data....
Downloads
acadgild@local...
acadgild@local...
[Software Upd...
Right Ctrl

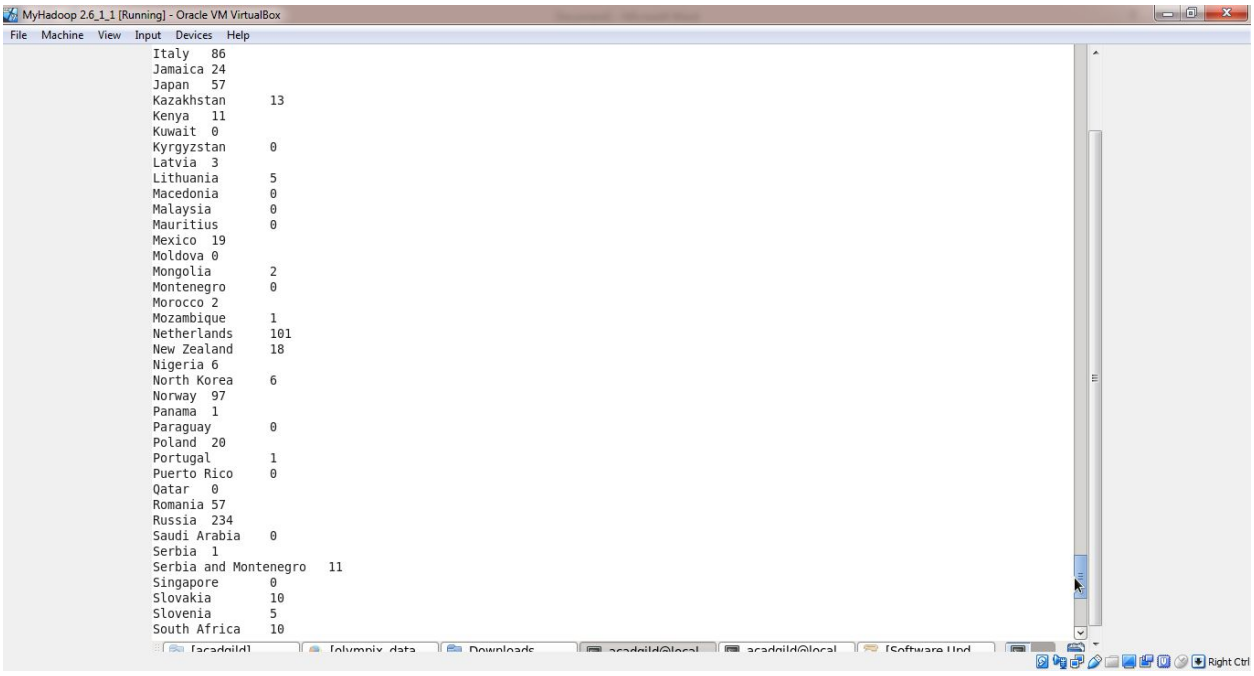

MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Israel 4
Italy 331
Jamaica 80
Japan 282
Kazakhstan 42
Kenya 39
Kuwait 2
Kyrgyzstan 3
Latvia 17
Lithuania 30
Macedonia 1
Malaysia 3
Mauritius 1
Mexico 38
Moldova 5
Mongolia 10
Montenegro 14
Morocco 11
Mozambique 1
Netherlands 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
[acadgild]
[olympix_data....
Downloads
acadgild@local...
acadgild@local...
[Software Upd...
Right Ctrl

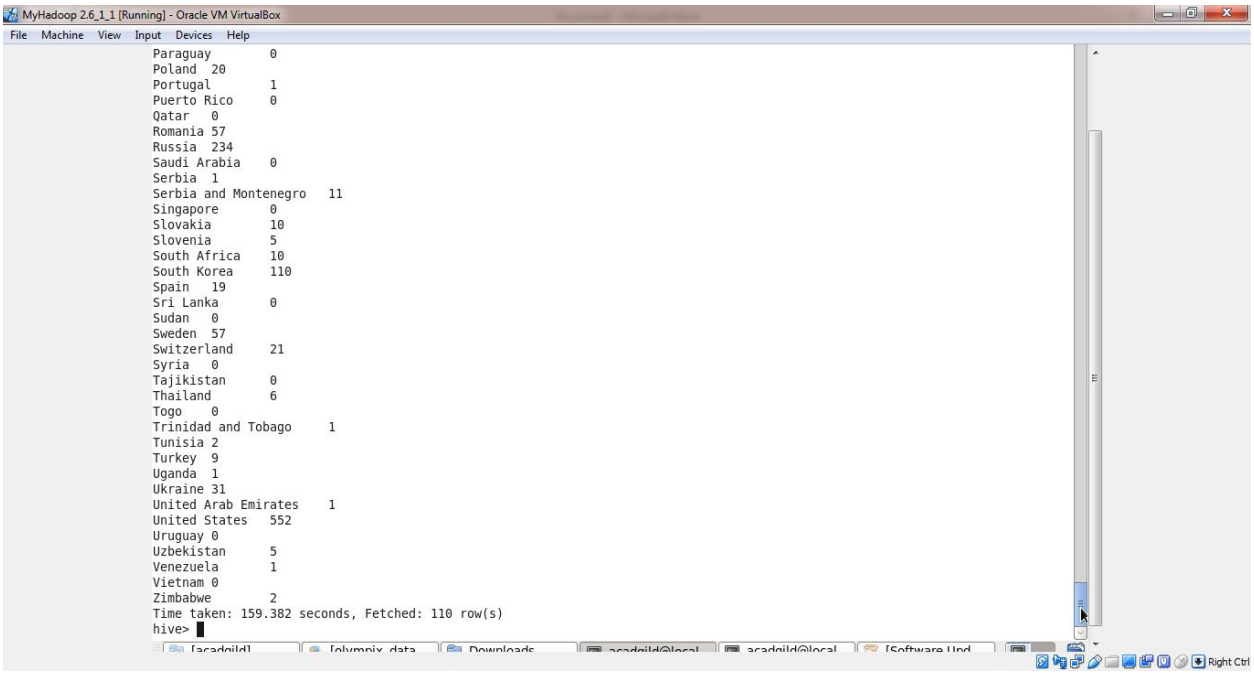


4. Write a Hive program to find the number of gold medals each country won.




MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
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
Right Ctrl


MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
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
Right Ctrl



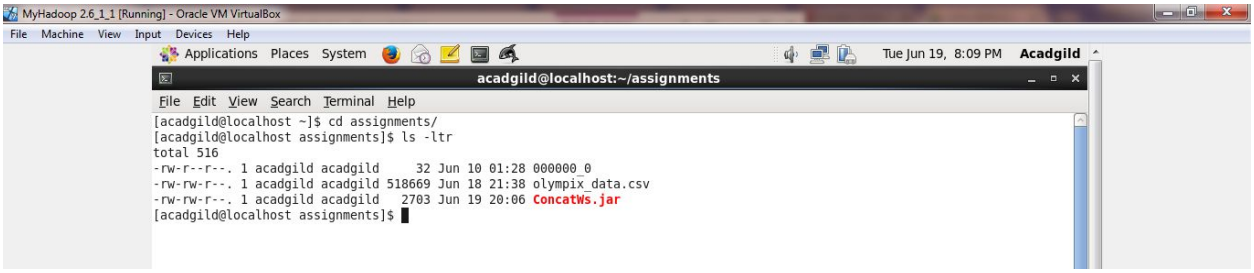
Task 2

Write a hive UDF that implements functionality of string concat_ws(string SEP, array<string>).

This UDF will accept two arguments, one string and one array of string.

It will return a single string where all the elements of the array are separated by the SEP.

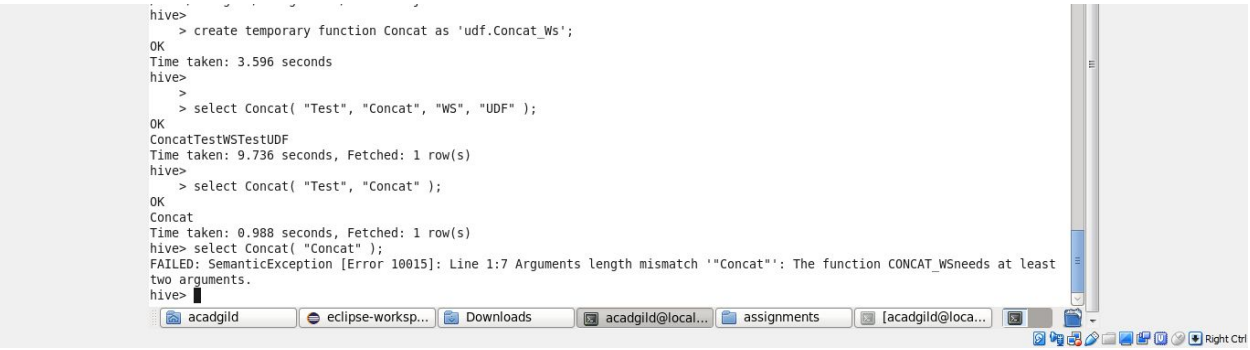
Created a jar for the above with name ConcatWs.jar.



Added jar to hive using **Add jar** command and verified if jar is added using **list jars** command.

Create a temporary function for the class Concat_Ws with name Concat.

Run the function with some examples.



Task 3

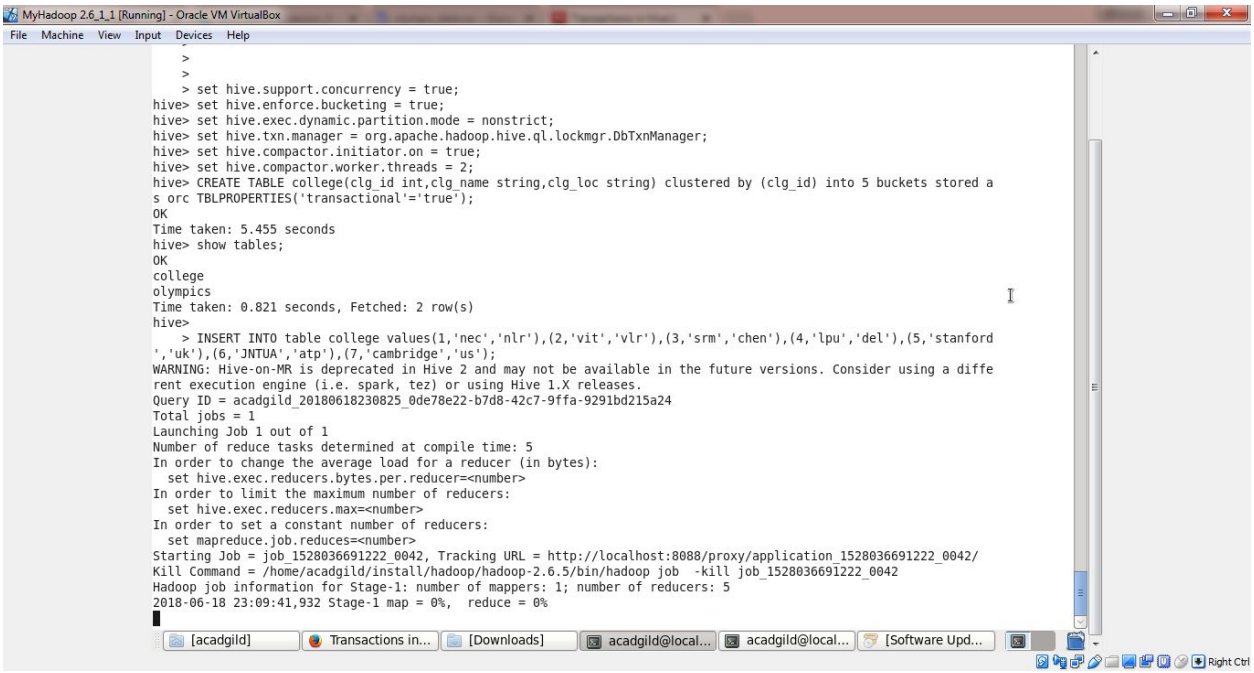
Link: https://acadgild.com/blog/transactions-in-hive/

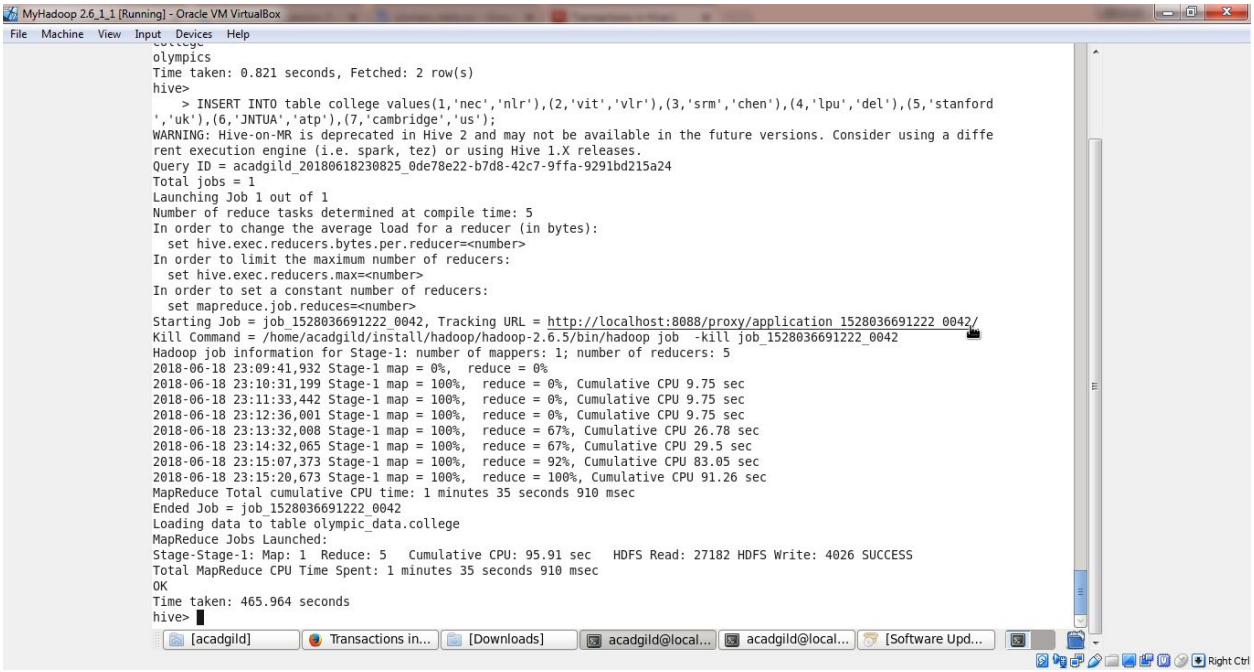
Refer the above given link for transactions in Hive and implement the operations given in the blog using your own sample data set and send us the screenshot.

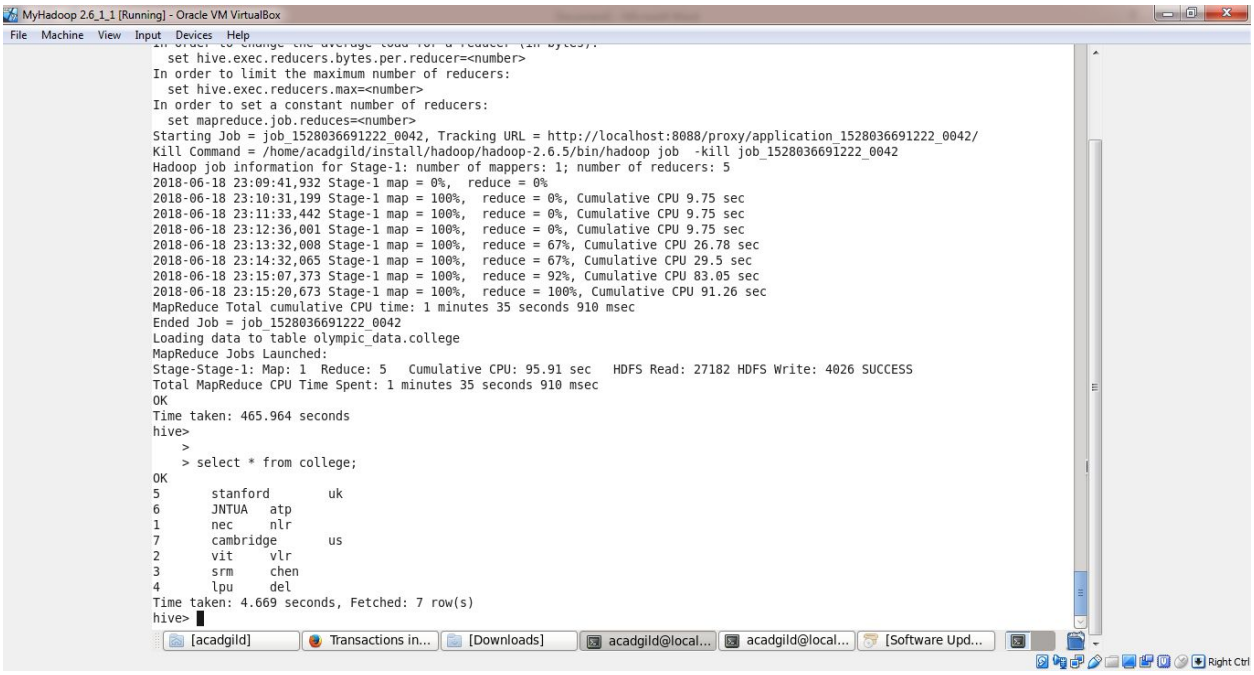
Below properties need to be set in hive shell so that insert, update and delete operations will work perfectly.

Create a table College with columns clg_id, clg_name, clg_loc columns.

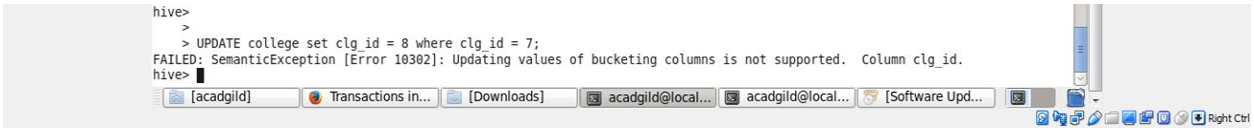
Insert some values to college table as below.


MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
> set hive.support.concurrency = true;
hive> set hive.enforce.bucketing = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
hive> set hive.txn.manager = org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;
hive> set hive.compactor.initiator.on = true;
hive> set hive.compactor.worker.threads = 2;
hive> CREATE TABLE college(clg_id int,clg_name string,clg_loc string) clustered by (clg_id) into 5 buckets stored a
s orc TBLPROPERTIES('transactional'='true');
OK
Time taken: 5.455 seconds
hive> show tables;
OK
college
olympics
Time taken: 0.821 seconds, Fetched: 2 row(s)
hive>
> INSERT INTO table college values(1,'nec','nlr'),(2,'vit','vlr'),(3,'srm','chen'),(4,'lpu','del'),(5,'stanford
','uk'),(6,'JNTUA','atp'),(7,'cambridge','us');
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a diffe
rent execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180618230825_0de78e22-b7d8-42c7-9ffa-9291bd215a24
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 5
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_1528036691222_0042, Tracking URL = http://localhost:8088/proxy/application_1528036691222_0042/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1528036691222_0042
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 5
2018-06-18 23:09:41,932 Stage-1 map = 0%, reduce = 0%
[acadgild]
Transactions in...
[Downloads]
acadgild@local...
acadgild@local...
[Software Upd...
Right Ctrl

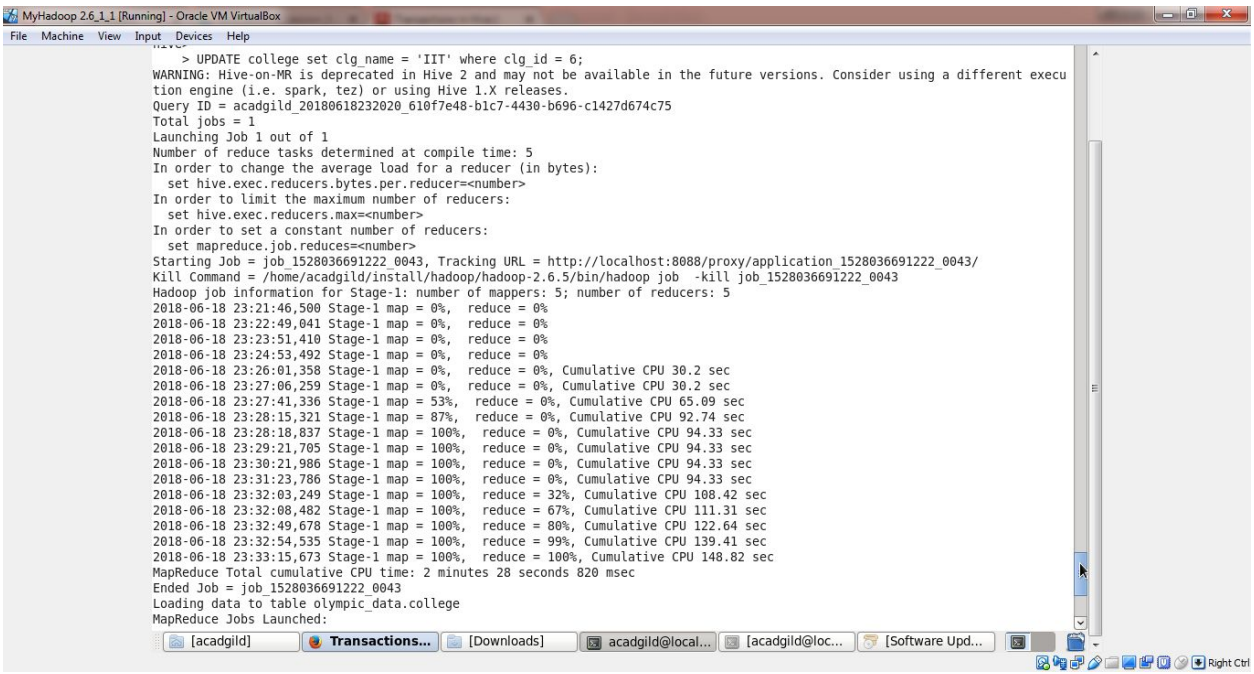

MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
olympics
Time taken: 0.821 seconds, Fetched: 2 row(s)
hive>
> INSERT INTO table college values(1,'nec','nlr'),(2,'vit','vlr'),(3,'srm','chen'),(4,'lpu','del'),(5,'stanford
','uk'),(6,'JNTUA','atp'),(7,'cambridge','us');
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a diffe
rent execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180618230825_0de78e22-b7d8-42c7-9ffa-9291bd215a24
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 5
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_1528036691222_0042, Tracking URL = http://localhost:8088/proxy/application_1528036691222_0042/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1528036691222_0042
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 5
2018-06-18 23:09:41,932 Stage-1 map = 0%, reduce = 0%
2018-06-18 23:10:31,199 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.75 sec
2018-06-18 23:11:33,442 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.75 sec
2018-06-18 23:12:36,001 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 9.75 sec
2018-06-18 23:13:32,008 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 26.78 sec
2018-06-18 23:14:32,065 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 29.5 sec
2018-06-18 23:15:07,373 Stage-1 map = 100%, reduce = 92%, Cumulative CPU 83.05 sec
2018-06-18 23:15:20,673 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 91.26 sec
MapReduce Total cumulative CPU time: 1 minutes 35 seconds 910 msec
Ended Job = job_1528036691222_0042
Loading data to table olympic_data.college
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 5 Cumulative CPU: 95.91 sec HDFS Read: 27182 HDFS Write: 4026 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 35 seconds 910 msec
OK
Time taken: 465.964 seconds
hive>
[acadgild]
Transactions in...
[Downloads]
acadgild@local...
acadgild@local...
[Software Upd...
Right Ctrl



Below is an update command on a bucketed column which shows error that it is not supported.

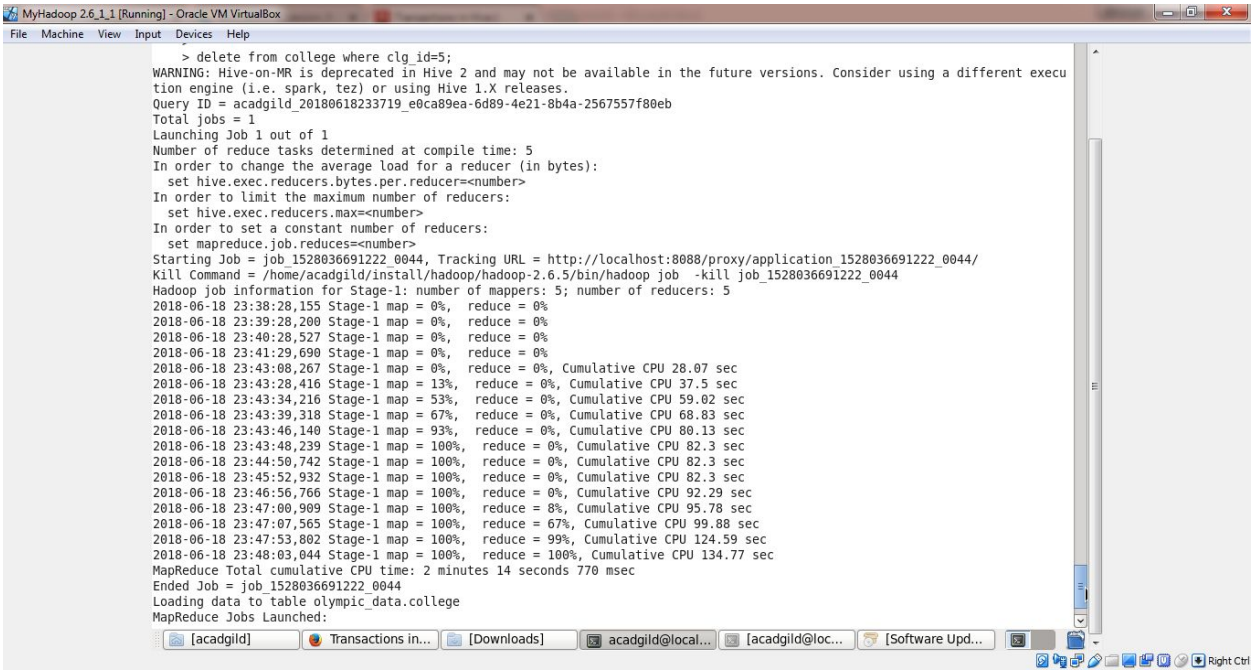


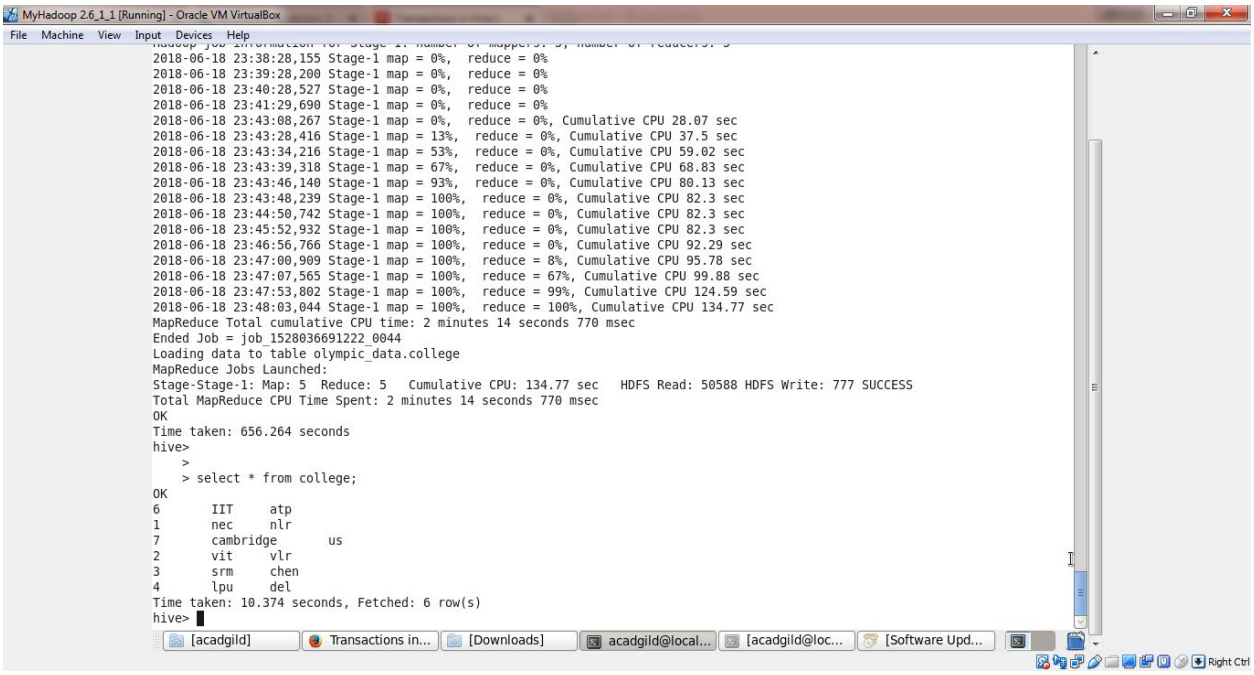
Update command on college table for column clg_name.



```
hive>
    >
    > select * from college;
OK
5       stanford        uk
6       IIT     atp
1       nec     nlr
7       cambridge       us
2       vit     vlr
3       srm     chen
4       lpu     del
Time taken: 2.906 seconds, Fetched: 7 row(s)
hive>
```

Delete command on college table which has column clg_id = 5.




MyHadoop 2.6_1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
2018-06-18 23:38:28,155 Stage-1 map = 0%, reduce = 0%
2018-06-18 23:39:28,200 Stage-1 map = 0%, reduce = 0%
2018-06-18 23:40:28,527 Stage-1 map = 0%, reduce = 0%
2018-06-18 23:41:29,690 Stage-1 map = 0%, reduce = 0%
2018-06-18 23:43:08,267 Stage-1 map = 0%, reduce = 0%, Cumulative CPU 28.07 sec
2018-06-18 23:43:28,416 Stage-1 map = 13%, reduce = 0%, Cumulative CPU 37.5 sec
2018-06-18 23:43:34,216 Stage-1 map = 53%, reduce = 0%, Cumulative CPU 59.02 sec
2018-06-18 23:43:39,318 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 68.83 sec
2018-06-18 23:43:46,140 Stage-1 map = 93%, reduce = 0%, Cumulative CPU 80.13 sec
2018-06-18 23:43:48,239 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 82.3 sec
2018-06-18 23:44:50,742 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 82.3 sec
2018-06-18 23:45:52,932 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 82.3 sec
2018-06-18 23:46:56,766 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 92.29 sec
2018-06-18 23:47:00,909 Stage-1 map = 100%, reduce = 8%, Cumulative CPU 95.78 sec
2018-06-18 23:47:07,565 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 99.88 sec
2018-06-18 23:47:53,802 Stage-1 map = 100%, reduce = 99%, Cumulative CPU 124.59 sec
2018-06-18 23:48:03,044 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 134.77 sec
MapReduce Total cumulative CPU time: 2 minutes 14 seconds 770 msec
Ended Job = job_1528036691222_0044
Loading data to table olympic_data.college
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5 Reduce: 5 Cumulative CPU: 134.77 sec HDFS Read: 50588 HDFS Write: 777 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 14 seconds 770 msec
OK
Time taken: 656.264 seconds
hive>
>
> select * from college;
OK
6 IIT atp
1 nec nlr
7 cambridge us
2 vit vlr
3 srm chen
4 lpu del
Time taken: 10.374 seconds, Fetched: 6 row(s)
hive>
[acadgild]
Transactions in...
[Downloads]
acadgild@local...
[acadgild@loc...
[Software Upd...
Right Ctrl