

Assignment 9.1 Advance Hive.

Table Creation:

create table olympic (athlete 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;

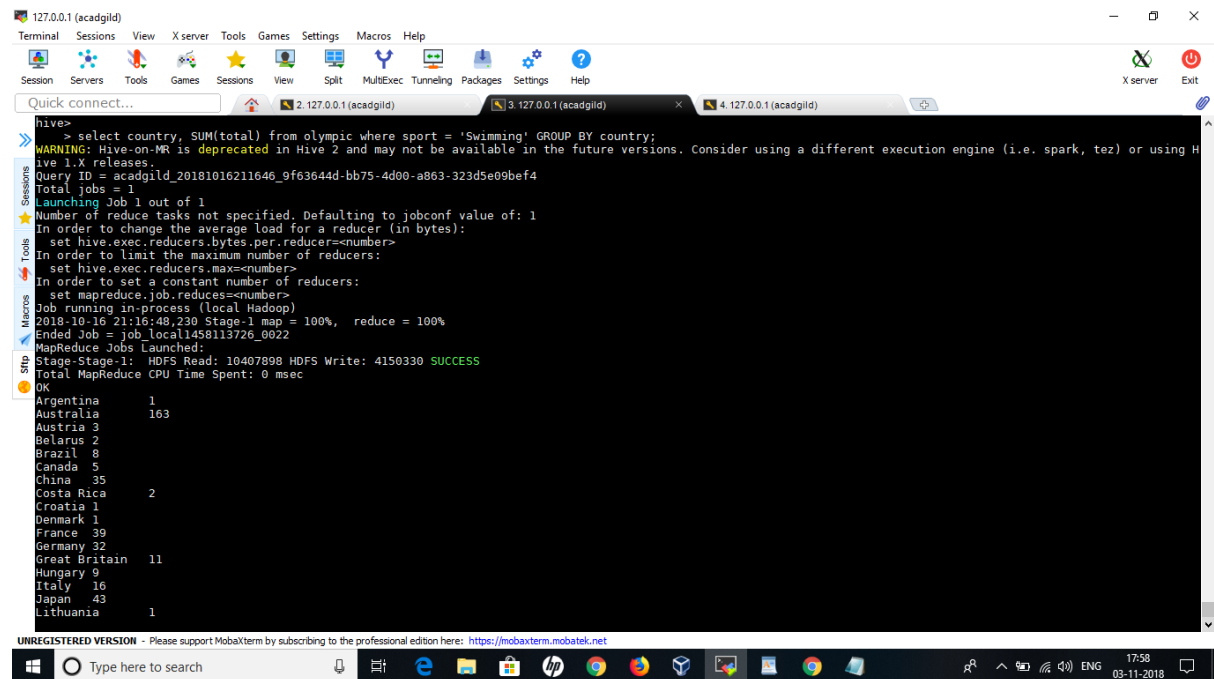
```
127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
hive> create table olympic (athlete STRING,age INT,country STRING,year STRING,closing STRING,sport STRING,gold INT,silver INT,bronze INT,total INT) row format delimited fields terminated by '\t' stored as textfile;
OK
Time taken: 0.183 seconds
hive> load data local inpath '/home/acadgild/Desktop/olympix_data.csv' into table olympic;
Loading data to table custom.olympic
OK
Time taken: 0.763 seconds
hive>
```

```
127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
hive> select * from olympic;
OK
Michael Phelps 23 United States 2008 08-24-08 Swimming 8 0 0 8
Michael Phelps 19 United States 2004 08-29-04 Swimming 6 0 2 8
Michael Phelps 27 United States 2012 08-12-12 Swimming 4 2 0 6
Natalie Coughlin 25 United States 2008 08-24-08 Swimming 1 2 3 6
Aleksey Nemov 24 Russia 2000 10-01-00 Gymnastics 2 1 3 6
Alicia Coutts 24 Australia 2012 08-12-12 Swimming 1 3 1 5
Missy Franklin 17 United States 2012 08-12-12 Swimming 4 0 1 5
Ryan Lochte 27 United States 2012 08-12-12 Swimming 2 2 1 5
Allison Schmitt 22 United States 2012 08-12-12 Swimming 3 1 1 5
Natalie Coughlin 21 United States 2004 08-29-04 Swimming 2 2 1 5
Ian Thorpe 17 Australia 2000 10-01-00 Swimming 3 2 0 5
Dara Torres 33 United States 2000 10-01-00 Swimming 2 0 3 5
Cindy Klassen 26 Canada 2006 02-26-06 Speed Skating 1 2 2 5
Nastia Liukin 18 United States 2008 08-24-08 Gymnastics 1 3 1 5
Marit Bjergen 29 Norway 2010 02-28-10 Cross Country Skiing 3 1 1 5
Sun Yang 20 China 2012 08-12-12 Swimming 1 1 4 5
Kirsty Coventry 24 Zimbabwe 2008 08-24-08 Swimming 1 3 0 4
Libby Lenton-Trickett 23 Australia 2008 08-24-08 Swimming 2 1 1 4
Ryan Lochte 24 United States 2008 08-24-08 Swimming 2 0 2 4
Inge de Bruijn 30 Netherlands 2004 08-29-04 Swimming 1 1 2 4
Petria Thomas 28 Australia 2004 08-29-04 Swimming 3 1 0 4
Ian Thorpe 21 Australia 2004 08-29-04 Swimming 2 1 1 4
Inge de Bruijn 27 Netherlands 2000 10-01-00 Swimming 3 1 0 4
Gary Hall Jr. 25 United States 2000 10-01-00 Swimming 2 1 1 4
Michael Klim 23 Australia 2000 10-01-00 Swimming 2 2 0 4
Susie O'Neill 27 Australia 2000 10-01-00 Swimming 1 3 0 4
Jenny Thompson 27 United States 2000 10-01-00 Swimming 3 0 1 4
Pieter van den Hoogenband 22 Netherlands 2000 10-01-00 Swimming 2 0 2 4
An Hyeon-Su 20 South Korea 2006 02-26-06 Short-Track Speed Skating 3 0 1 4
Aliya Mustafina 17 Russia 2012 08-12-12 Gymnastics 1 1 2 4
Shawn Johnson 16 United States 2008 08-24-08 Gymnastics 1 3 0 4
Dmitry Sautin 26 Russia 2000 10-01-00 Diving 1 1 2 4
Leonien Zijlward-van Moorsel 30 Netherlands 2000 10-01-00 Cycling 3 1 0 4
Petter Northug Jr. 24 Norway 2010 02-28-10 Cross Country Skiing 2 1 1 4
Ole Einar Bjoerdalen 28 Norway 2002 02-24-02 Biathlon 4 0 0 4
Janica Kostelic 20 Croatia 2002 02-24-02 Alpine Skiing 3 1 0 4
```

Task 1

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

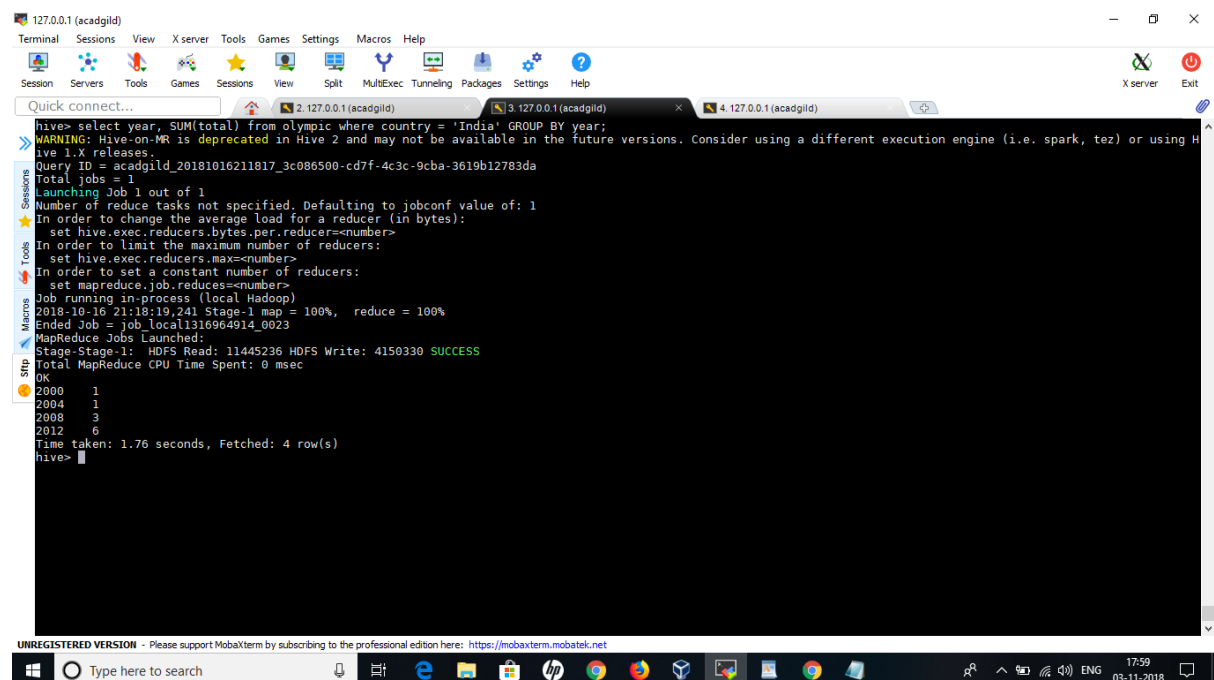
Solution: *select country, SUM(total) from olympic where sport = 'Swimming' GROUP BY country;*



```
127.0.0.1 (acadgild)
Terminal Sessions View Xserver Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
> hive>
> select country, SUM(total) from olympic where sport = 'Swimming' GROUP BY country;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181016211646_9f63644d-bb75-4d00-a863-323d5e09bef4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Defaulting to jobconf value of: 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>
Job running in-process (local Hadoop)
2018-10-16 21:16:48,230 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1458113726_0022
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 10407898 HDFS Write: 4150330 SUCCESS
Total MapReduce CPU Time Spent: 0 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
```

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

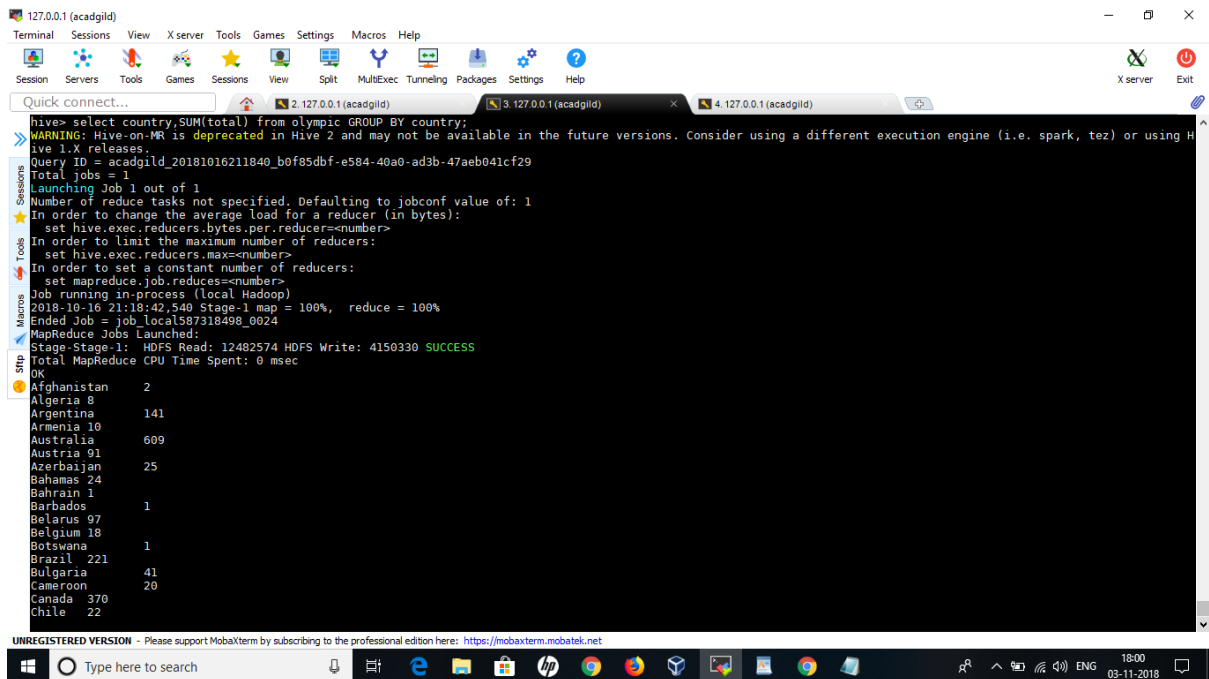
Solution: *select year, SUM(total) from olympic where country = 'India' GROUP BY year;*



```
127.0.0.1 (acadgild)
Terminal Sessions View Xserver Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help
Quick connect...
> 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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181016211817_3c086500-cd7f-4c3c-9cba-3619b12783da
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Defaulting to jobconf value of: 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>
Job running in-process (local Hadoop)
2018-10-16 21:18:19,241 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1316964914_0023
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 11445236 HDFS Write: 4150330 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
2000      1
2004      1
2008      3
2012      6
Time taken: 1.76 seconds, Fetched: 4 row(s)
hive>
```

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

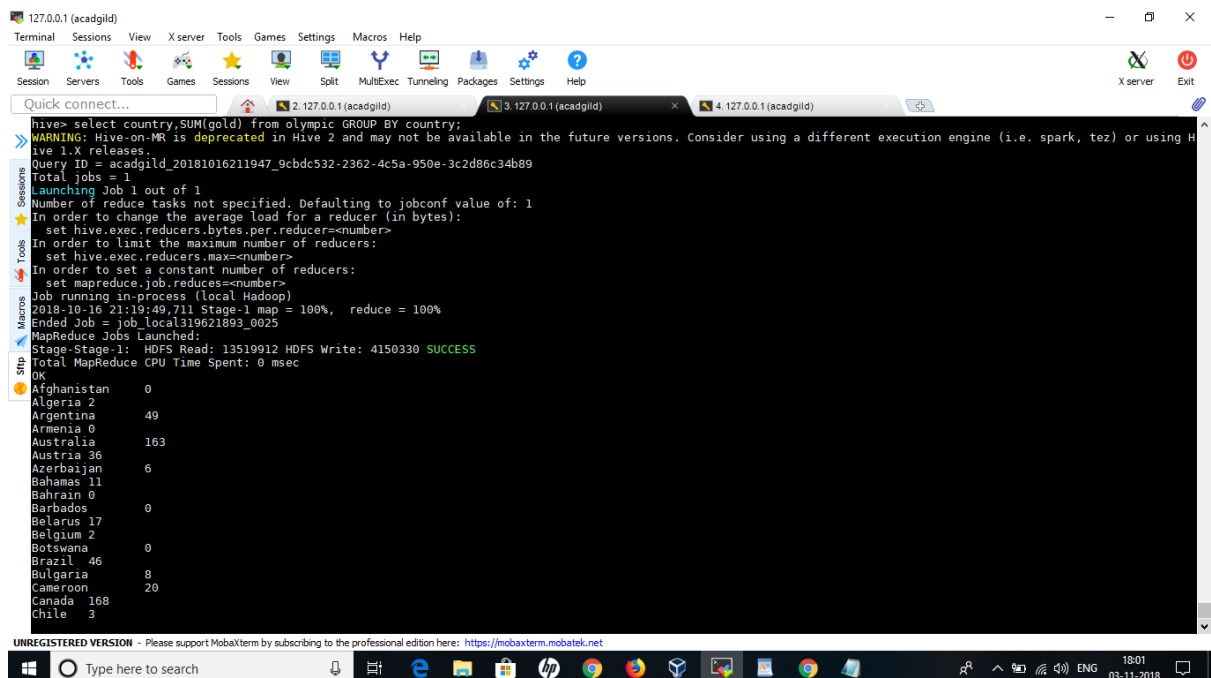
Solution: *select country,SUM(total) from olympic GROUP BY country;*



```
Hive> select country,SUM(total) from olympic GROUP BY country;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181016211840_b0f85dbf-e584-40a0-ad3b-47aeb041cf29
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Defaulting to jobconf value of: 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.reducers=<number>
Job running in-process (local Hadoop)
2018-10-16 21:18:42,540 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local587318498_0024
MapReduce Jobs Launched:
Stage:Stage-1:  HDFS Read: 12402574 HDFS Write: 4150330 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
country      total
-----
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
```

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

Solution: *select country,SUM(gold) from olympic GROUP BY country;*



```
Hive> select country,SUM(gold) from olympic GROUP BY country;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181016211947_9cbdc532-2362-4c5a-950e-3c2d86c34b89
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Defaulting to jobconf value of: 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.reducers=<number>
Job running in-process (local Hadoop)
2018-10-16 21:19:49,711 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local319621893_0025
MapReduce Jobs Launched:
Stage:Stage-1:  HDFS Read: 13519912 HDFS Write: 4150330 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
country      gold
-----
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
```

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.

Solution:

```
package com.udfconcatws;

import java.util.ArrayList;
import org.apache.hadoop.hive.ql.exec.UDF;

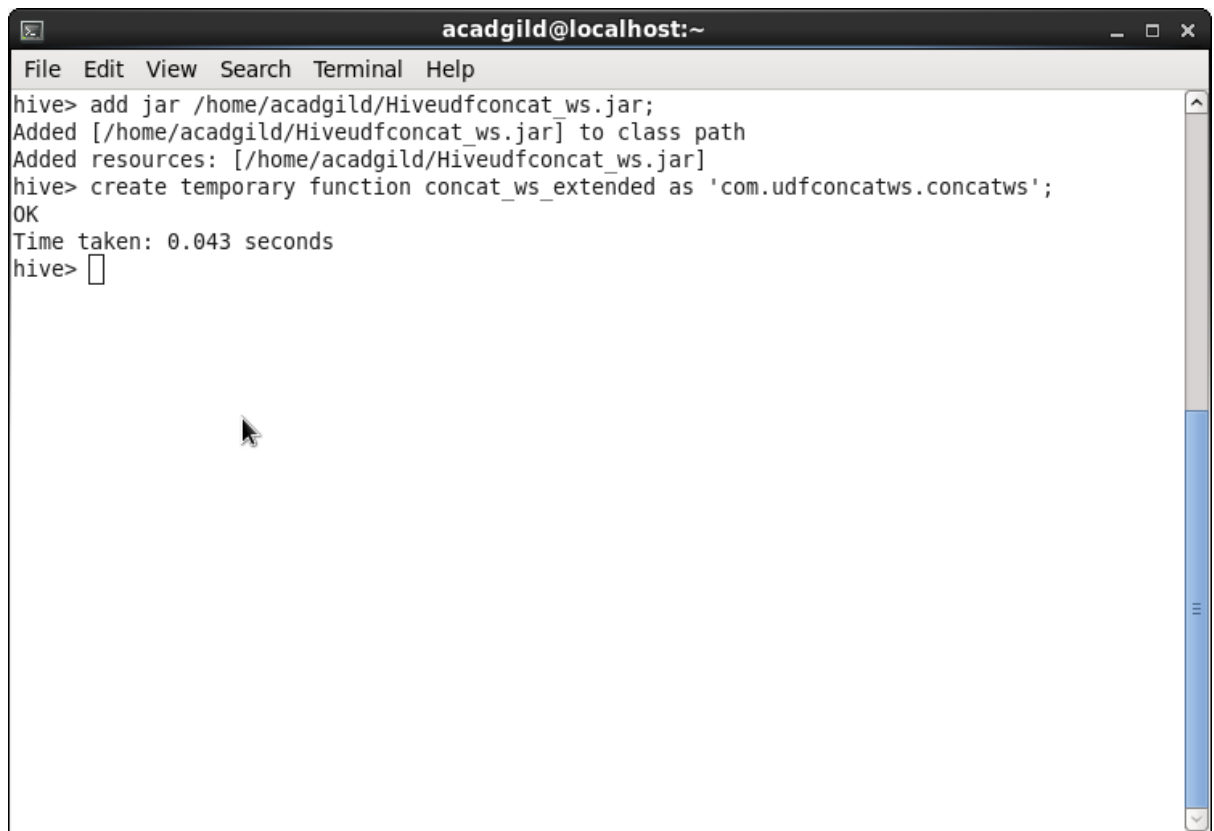
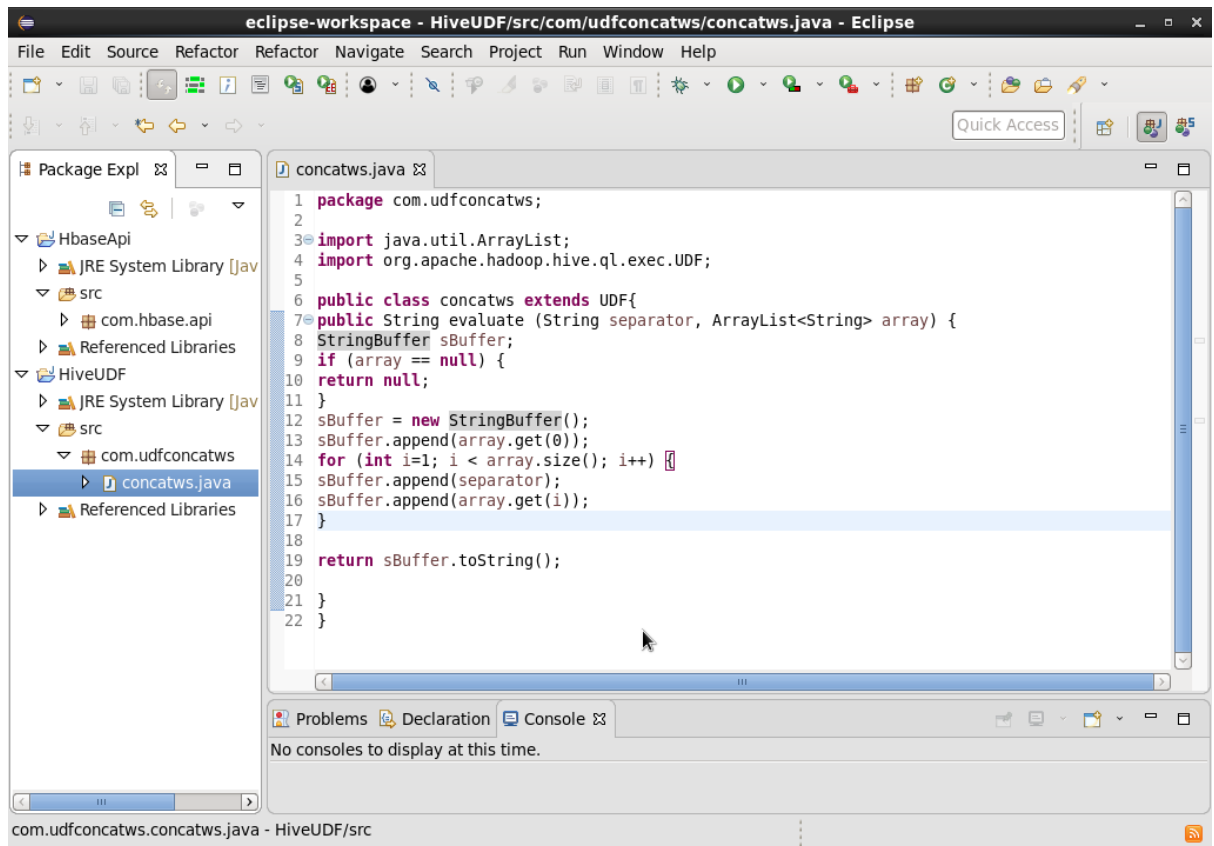
public class concatws extends UDF{

    public String evaluate (String separator, ArrayList<String> array) {

        StringBuffer sBuffer;

        if (array == null) {
            return null;
        }

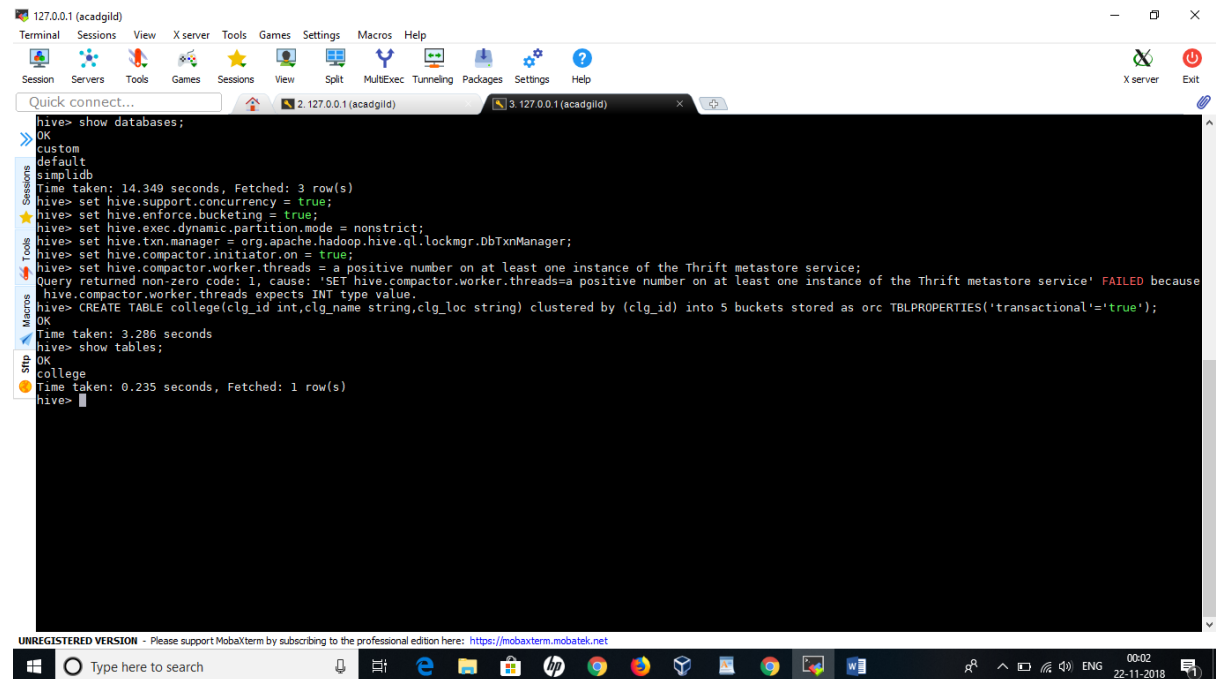
        sBuffer = new StringBuffer();
        sBuffer.append(array.get(0));
        for (int i=1; i < array.size(); i++) {
            sBuffer.append(separator);
            sBuffer.append(array.get(i));
        }
        return sBuffer.toString();
    }
}
```



Task 3: ACID Properties in Hive:

Solution: Creating a Table That Supports ACID Hive Transactions

CMD: `CREATE TABLE college(clg_id int,clg_name string,clg_loc string) clustered by (clg_id) into 5 buckets stored as orc TBLPROPERTIES('transactional'='true');`



The screenshot shows a terminal window titled '127.0.0.1 (acadgild)' with a menu bar (Terminal, Sessions, View, X server, Tools, Games, Settings, Macros, Help) and a toolbar. The terminal displays the following commands and output:

```
hive> show databases;
OK
>
custom
default
sampledb
Time taken: 14.349 seconds, Fetched: 3 row(s)
hive> 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 = a positive number on at least one instance of the Thrift metastore service;
Query returned non-zero code: 1, cause: 'SET hive.compactor.worker.threads=a positive number on at least one instance of the Thrift metastore service' FAILED because
hive.compactor.worker.threads expects INT type value.
hive> CREATE TABLE college(clg_id int,clg_name string,clg_loc string) clustered by (clg_id) into 5 buckets stored as orc TBLPROPERTIES('transactional'='true');
OK
Time taken: 3.286 seconds
hive> show tables;
OK
college
Time taken: 0.235 seconds, Fetched: 1 row(s)
hive>
```

At the bottom of the terminal window, there is a message: "UNREGISTERED VERSION - Please support Mobalterm by subscribing to the professional edition here: <https://mobalterm.mcbatek.net>". The Windows taskbar is visible at the bottom of the screen.

Inserting Data into the Hive Table:

CMD: `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');`

The screenshot shows a MobaXterm window with a terminal session. The user has executed a Hive INSERT statement into a table named 'college'. The output shows the job was launched successfully and the data was loaded. The user then executed a SELECT statement to view the data, which is displayed as a table with 7 rows.

```
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 different execution engine (i.e. spark, tez) or using H
ive 1.X releases.
Query ID = acadgild_20181119043438_b055b3b9-9467-4133-936c-03b4180d4e30
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>
Job running in-process (local Hadoop)
2018-11-19 04:34:48,585 Stage-1 map = 0%, reduce = 0%
2018-11-19 04:34:49,668 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local364725743_0001
Loading data to table default.college
MapReduce Jobs Launched:
Stage-Stage1: HDFS Read: 492 HDFS Write: 12595 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 14.176 seconds
hive> select * from college;
OK
5      stanford      uk
6      JNTUA      atp
1      nec      nlr
2      vit      vlr
7      cambridge      us
3      srm      chen
4      lpu      del
Time taken: 0.666 seconds, Fetched: 7 row(s)
hive>
```

Re-insert the Same Value again:

The screenshot shows a MobaXterm window with a terminal session. The user has executed a Hive INSERT statement into a table named 'college'. The output shows the job was launched successfully and the data was loaded. The user then executed a SELECT statement to view the data, which is displayed as a table with 7 rows.

```
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 different execution engine (i.e. spark, tez) or using H
ive 1.X releases.
Query ID = acadgild_20181119043556_5ff02460-8c61-46ae-97bc-e48ded56a43f
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>
Job running in-process (local Hadoop)
2018-11-19 04:35:58,781 Stage-1 map = 100%, reduce = 100%
Ended Job = job_local772997562_0002
Loading data to table default.college
MapReduce Jobs Launched:
Stage-Stage1: HDFS Read: 29172 HDFS Write: 37093 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 5.141 seconds
hive> select * from college;
OK
5      stanford      uk
5      stanford      uk
6      JNTUA      atp
1      nec      nlr
6      JNTUA      atp
1      nec      nlr
7      cambridge      us
2      vit      vlr
7      cambridge      us
2      vit      vlr
3      srm      chen
3      srm      chen
4      lpu      del
4      lpu      del
```

Update the Data in Hive Table:

CMD: *UPDATE college set clg_name = 'IIT' where clg_id = 6;*

```
127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Quick connect... 2 127.0.0.1 (acadgild) 3 127.0.0.1 (acadgild)
>> hive> UPDATE college set clg_name = 'IIT' where clg_id = 6;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181119043839_94f066ed-5dc1-408d-a2b8-19ffec23f6b7
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>
Job running in-process (local Hadoop)
2018-11-19 04:38:42.142 Stage-1 map = 100%,  reduce = 0%
2018-11-19 04:38:43.172 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local197971906_0003
Loading data to table default.college
MapReduce Jobs Launched:
Stage-Stage1:  HDFS Read: 207461 HDFS Write: 85301 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 5.936 seconds
hive> select * from college;
OK
  clg_id  clg_name  clg_address
-----
5        stanford  uk
5        stanford  uk
6        IIT      atp
6        nec      nlr
6        IIT      atp
6        nec      nlr
7        cambridge  us
2        vit      vlr
7        cambridge  us
2        vit      vlr
2        srm      chen
3        srm      chen
4        lpu      del
4        lpu      del
Time taken: 0.446 seconds, Fetched: 12 row(s)
```

Delete the Row from Hive Table:

CMD: *delete from college where clg_id=5;*

```
127.0.0.1 (acadgild)
Terminal Sessions View X server Tools Games Settings Macros Help
Quick connect... 2 127.0.0.1 (acadgild) 3 127.0.0.1 (acadgild)
>> hive> delete from college where clg_id=5;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181119044014_b320c06d-cf7c-4f01-9854-c1f1b34178c8
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>
Job running in-process (local Hadoop)
2018-11-19 04:40:16.853 Stage-1 map = 100%,  reduce = 0%
2018-11-19 04:40:17.874 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local651447377_0004
Loading data to table default.college
MapReduce Jobs Launched:
Stage-Stage1:  HDFS Read: 397249 HDFS Write: 94655 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 4.374 seconds
hive> select * from college;
OK
  clg_id  clg_name  clg_address
-----
6        IIT      atp
6        nec      nlr
6        IIT      atp
6        nec      nlr
7        cambridge  us
2        vit      vlr
7        cambridge  us
2        vit      vlr
2        srm      chen
3        srm      chen
4        lpu      del
4        lpu      del
Time taken: 0.446 seconds, Fetched: 12 row(s)
```