Task 1

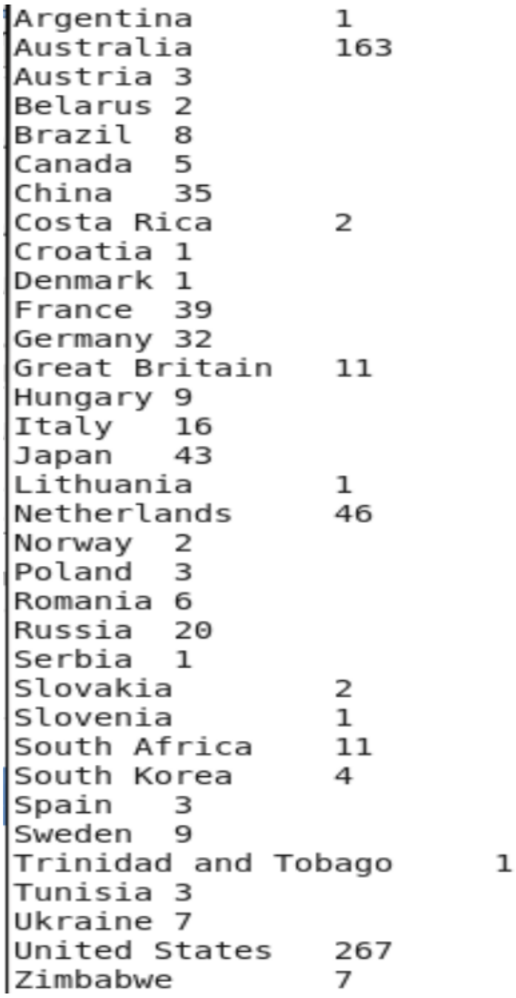
----

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

Code:

```
SELECT country,SUM(total_medals) as medalcount FROM olympics WHERE sport='Swimming'
GROUP BY country;
```

Output:



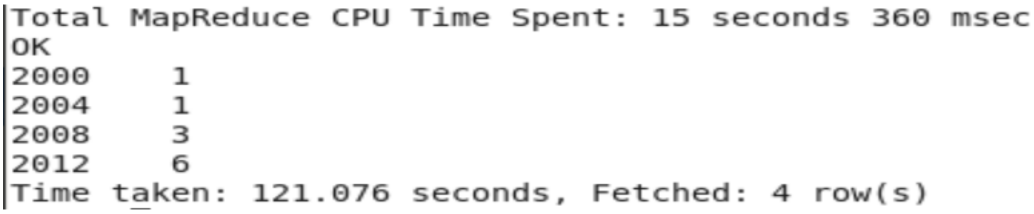
```
Argentina       1
Australia       163
Austria 3
Belarus 2
Brazil  8
Canada  5
China   35
Costa Rica      2
Croatia 1
Denmark 1
France  39
Germany 32
Great Britain   11
Hungary 9
Italy   16
Japan   43
Lithuania       1
Netherlands     46
Norway  2
Poland  3
Romania 6
Russia  20
Serbia  1
Slovakia        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
```

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

Code:

```
SELECT year,SUM(total_medals) FROM olympics WHERE country='India'
GROUP BY year ORDER BY year;
```

Output:



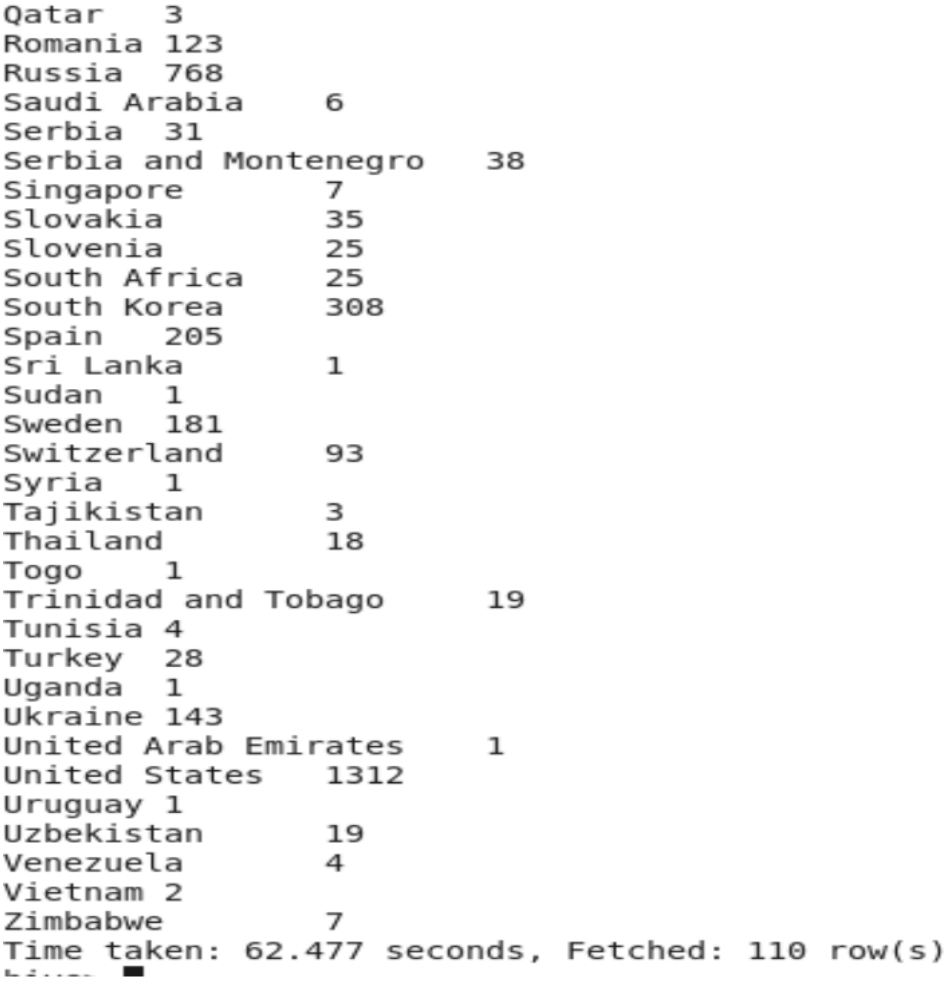
```
Total MapReduce CPU Time Spent: 15 seconds 360 msec
OK
2000    1
2004    1
2008    3
2012    6
Time taken: 121.076 seconds, Fetched: 4 row(s)
```

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

Code:

SELECT country,SUM(total_medals) as medalcount FROM olympics
GROUP BY country;

Output: showing a part of the result set.



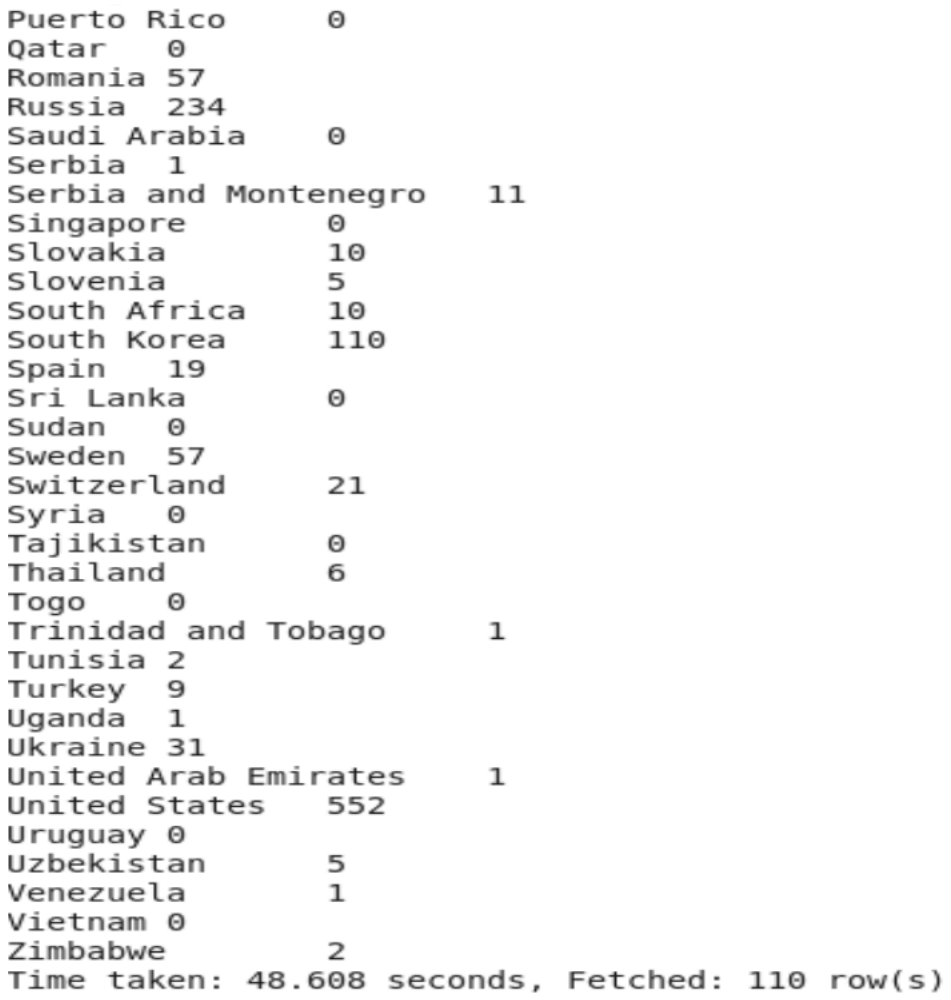
```
Qatar   3
Romania 123
Russia  768
Saudi Arabia    6
Serbia  31
Serbia and Montenegro   38
Singapore       7
Slovakia        35
Slovenia        25
South Africa    25
South Korea     308
Spain   205
Sri Lanka       1
Sudan   1
Sweden  181
Switzerland     93
Syria   1
Tajikistan      3
Thailand        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
Time taken: 62.477 seconds, Fetched: 110 row(s)
hive> █
```

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

Code:

SELECT country,SUM(gold_medals) as medalcount FROM olympics
GROUP BY country;

Output: showing part of result set



```
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
Time taken: 48.608 seconds, Fetched: 110 row(s)
```

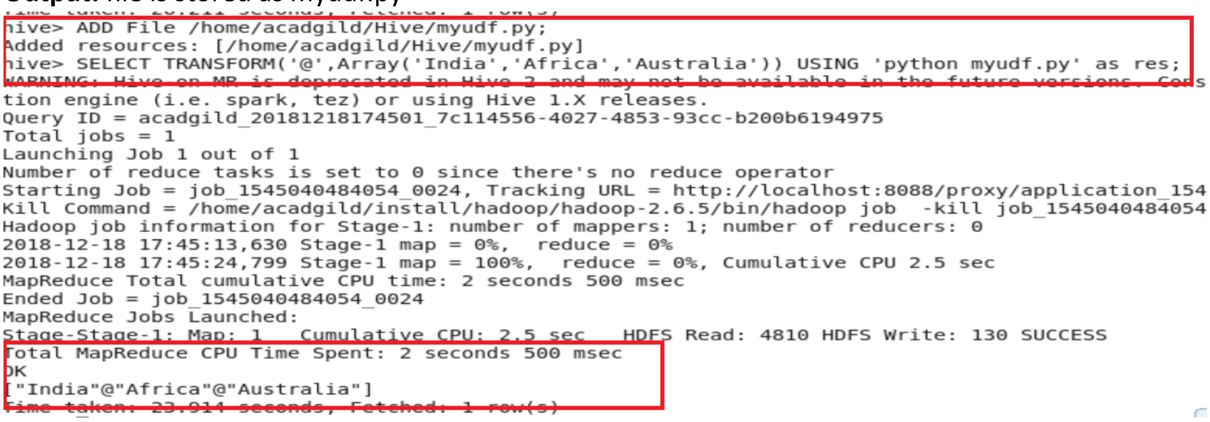
Task 2

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.

Code:

```
import sys
for line in sys.stdin:
 line=line.strip()
 (delim,str_array) = line.split(' ')
 str_array1=str_array.split(',')
 print str_array1
 res=''
 for i,element in enumerate(str_array1):
   if i<len(str_array1)-1:
       res+=element+delim
  else:
       res+=element
print str(res)
```

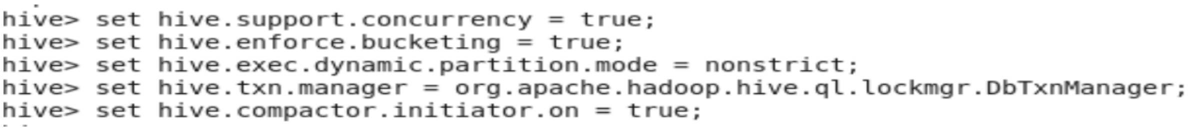
Output: file is stored as myudf.py



```
Time taken: 20.211 seconds, Fetched: 1 row(s)
hive> ADD File /home/acadgild/Hive/myudf.py;
Added resources: [/home/acadgild/Hive/myudf.py]
hive> SELECT TRANSFORM('@',Array('India','Africa','Australia')) USING 'python myudf.py' as res;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Cons
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181218174501_7c114556-4027-4853-93cc-b200b6194975
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1545040484054_0024, Tracking URL = http://localhost:8088/proxy/application_154
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1545040484054
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-12-18 17:45:13,630 Stage-1 map = 0%,  reduce = 0%
2018-12-18 17:45:24,799 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
MapReduce Total cumulative CPU time: 2 seconds 500 msec
Ended Job = job_1545040484054_0024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.5 sec   HDFS Read: 4810 HDFS Write: 130 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 500 msec
OK
["India"@"Africa"@"Australia"]
Time taken: 23.914 seconds, Fetched: 1 row(s)
```

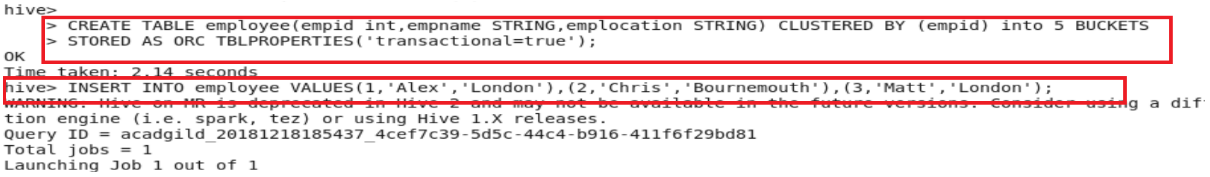
Task 3

Setting below properties to activate row level transactions in hive-



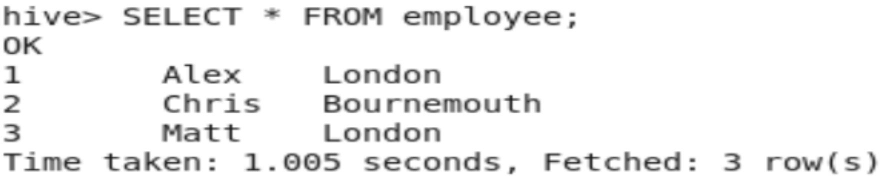
```
hive> set hive.support.concurrency = true;
hive> set hive.enforce.bucketing = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
hive> set hive.txn.manager = org.apache.hadoop.hive.ql.lockmgr.DbTxnManager;
hive> set hive.compactor.initiator.on = true;
```

Created employee table and inserted 3 dummy rows-



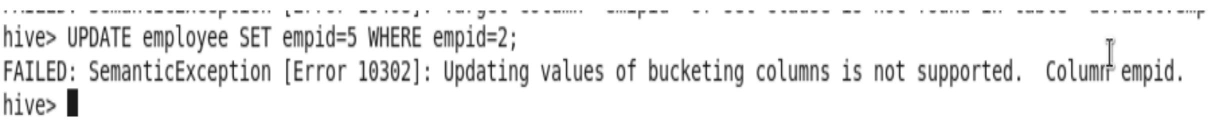
```
hive>
    > CREATE TABLE employee(empid int,empname STRING,emplocation STRING) CLUSTERED BY (empid) into 5 BUCKETS
    > STORED AS ORC TBLPROPERTIES('transactional=true');
OK
Time taken: 2.14 seconds
hive> INSERT INTO employee VALUES(1,'Alex','London'),(2,'Chris','Bournemouth'),(3,'Matt','London');
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a dif
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181218185437_4cef7c39-5d5c-44c4-b916-411f6f29bd81
Total jobs = 1
Launching Job 1 out of 1
```

Selecting data from employee table-



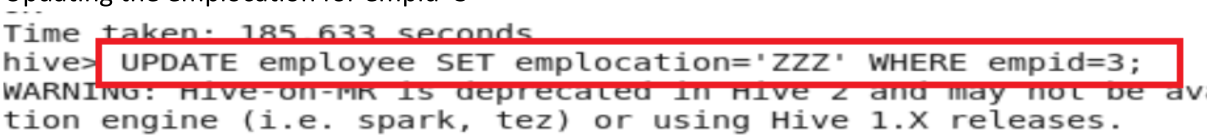
```
hive> SELECT * FROM employee;
OK
1       Alex    London
2       Chris   Bournemouth
3       Matt    London
Time taken: 1.005 seconds, Fetched: 3 row(s)
```

Throwing error while trying to update the bucketing column-



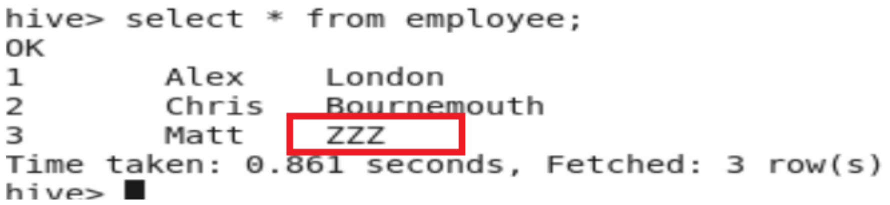
```
hive> UPDATE employee SET empid=5 WHERE empid=2;
FAILED: SemanticException [Error 10302]: Updating values of bucketing columns is not supported.  Column empid.
hive>
```

Updating the emplocation for empid=3



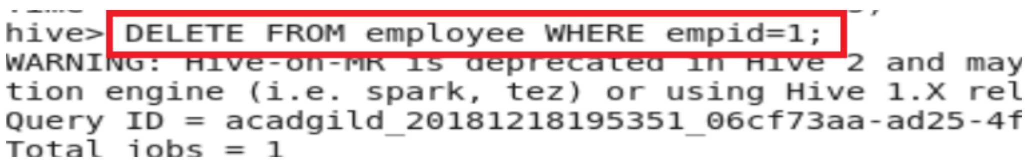
```
Time taken: 185.633 seconds
hive> UPDATE employee SET emplocation='ZZZ' WHERE empid=3;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be av
tion engine (i.e. spark, tez) or using Hive 1.X releases.
```

Updated value-



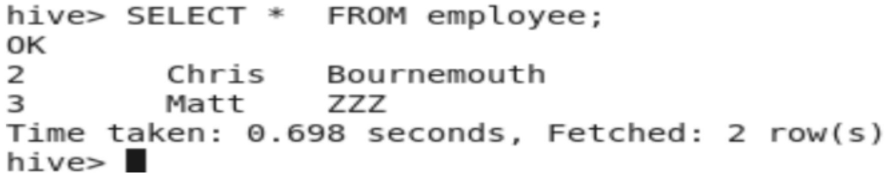
```
hive> select * from employee;
OK
1        Alex     London
2        Chris    Bournemouth
3        Matt     ZZZ
Time taken: 0.861 seconds, Fetched: 3 row(s)
hive> █
```

Deleting a row with empid=1



```
hive> DELETE FROM employee WHERE empid=1;
WARNING: Hive-on-MR is deprecated in Hive 2 and may
tion engine (i.e. spark, tez) or using Hive 1.X rel
Query ID = acadgild_20181218195351_06cf73aa-ad25-4f
Total jobs = 1
```

Output: record with empid=1 got deleted



```
hive> SELECT *  FROM employee;
OK
2        Chris    Bournemouth
3        Matt     ZZZ
Time taken: 0.698 seconds, Fetched: 2 row(s)
hive> █
```