Dublin City University
School of Computing
CA675 Cloud Technologies
Assignment - 1

- **Student Details: -**

Student Name	Soumi Mitra
Student ID Number	20210300
Student Email ID	Soumi.mitra2@mail.dcu.ie
Programme of Study	MSc in Computing – Data Analytics (Full Time)
Module Code	CA675
Date of Submission	23-11-2020

❖ Get data from Stack Exchange

```
select top 50000 * from posts where posts.ViewCount>15000 and posts.ViewCount < 20000
ORDER BY posts.ViewCount   ---- saved as Set1.csv

select top 50000 * from posts where posts.ViewCount>16575 and posts.ViewCount < 20000
ORDER BY posts.ViewCount    ---- saved as Set2.csv

select top 50000 * from posts where posts.ViewCount>18478 and posts.ViewCount < 30000
ORDER BY posts.ViewCount  ---- saved as Set3.csv

select top 50000 * from posts where posts.ViewCount>20868 and posts.ViewCount < 30000
ORDER BY posts.ViewCount   ---- saved as Set4.csv
```

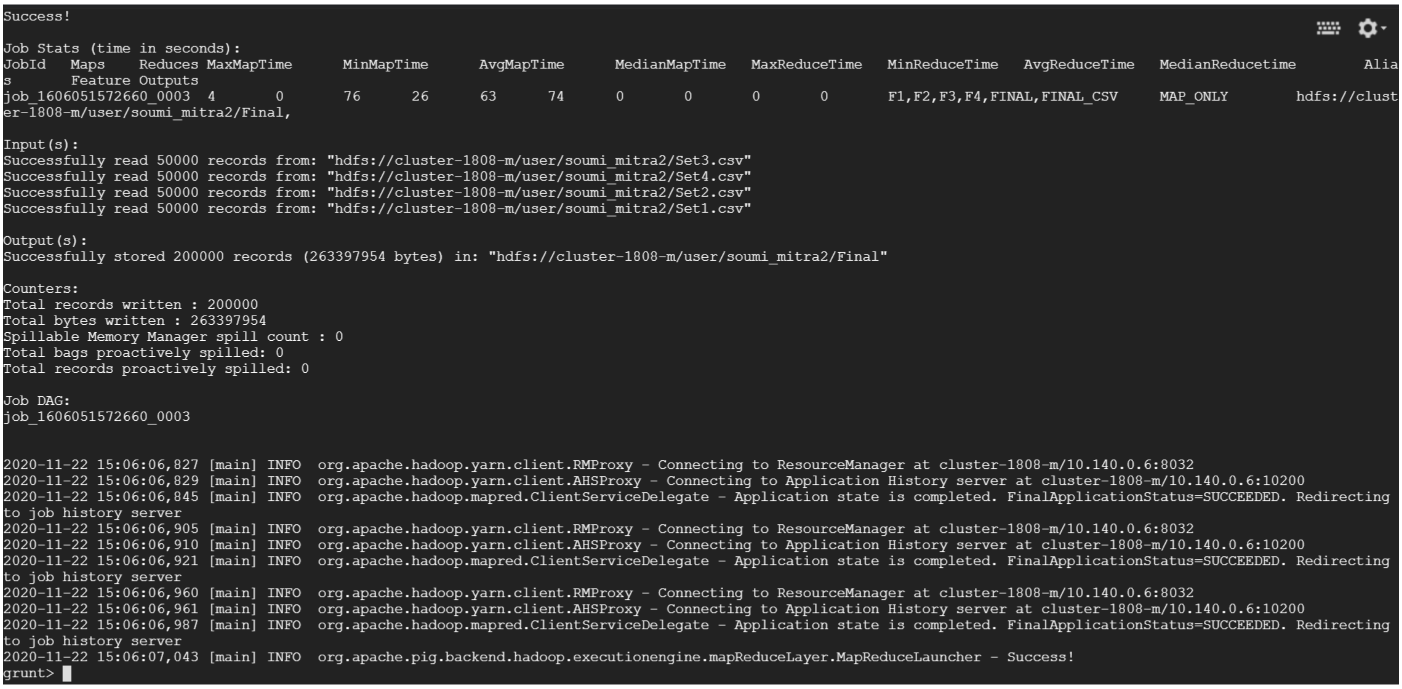
❖ Loading the CSV files using PIG without header

```
F1 = load 'hdfs://cluster-1808-m/user/soumi_mitra2/Set1.csv' using
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER') as
(Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:Datetime, DeletionDate:Datetime,
Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray,
LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:Datetime, LastActivityDate:Datetime,
Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int,
ClosedDate:Datetime, CommunityOwnedDate:Datetime);

F2 = load 'hdfs://cluster-1808-m/user/soumi_mitra2/Set2.csv' using
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER') as
(Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:Datetime, DeletionDate:Datetime,
Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray,
LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:Datetime, LastActivityDate:Datetime,
Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int,
ClosedDate:Datetime, CommunityOwnedDate:Datetime);
```

```
F3 = load 'hdfs://cluster-1808-m/user/soumi_mitra2/Set3.csv' using
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER') as
(Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:Datetime, DeletionDate:Datetime,
Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray,
LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:Datetime, LastActivityDate:Datetime,
Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int,
ClosedDate:Datetime, CommunityOwnedDate:Datetime);

F4 = load 'hdfs://cluster-1808-m/user/soumi_mitra2/Set4.csv' using
org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER') as
(Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:Datetime, DeletionDate:Datetime,
Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDisplayName:chararray,
LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:Datetime, LastActivityDate:Datetime,
Title:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int,
ClosedDate:Datetime, CommunityOwnedDate:Datetime);

FINAL = UNION F1,F2,F3,F4;

FINAL_CSV = FOREACH FINAL GENERATE Id as Id, PostTypeId as PostTypeId,  Score as Score, ViewCount as
ViewCount,  REPLACE(REPLACE(Body,'\n',''),',','') as Body, OwnerUserId as OwnerUserId,
REPLACE(OwnerDisplayName,'"','') as OwnerDisplayName, Title as Title, AnswerCount as AnswerCount;
```

- **Storing the data using PIG**

```
STORE FINAL_CSV INTO 'hdfs://cluster-1808-m/user/soumi_mitra2/Final' USING
org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE');
```



```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alia
s       Feature Outputs
job_1606051572660_0003  4       0       76      26      63      74      0       0       0       0       F1,F2,F3,F4,FINAL,FINAL_CSV     MAP_ONLY        hdfs://clust
er-1808-m/user/soumi_mitra2/Final,

Input(s):
Successfully read 50000 records from: "hdfs://cluster-1808-m/user/soumi_mitra2/Set3.csv"
Successfully read 50000 records from: "hdfs://cluster-1808-m/user/soumi_mitra2/Set4.csv"
Successfully read 50000 records from: "hdfs://cluster-1808-m/user/soumi_mitra2/Set2.csv"
Successfully read 50000 records from: "hdfs://cluster-1808-m/user/soumi_mitra2/Set1.csv"

Output(s):
Successfully stored 200000 records (263397954 bytes) in: "hdfs://cluster-1808-m/user/soumi_mitra2/Final"

Counters:
Total records written : 200000
Total bytes written : 263397954
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1606051572660_0003


2020-11-22 15:06:06,827 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-1808-m/10.140.0.6:8032
2020-11-22 15:06:06,829 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-1808-m/10.140.0.6:10200
2020-11-22 15:06:06,845 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
to job history server
2020-11-22 15:06:06,905 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-1808-m/10.140.0.6:8032
2020-11-22 15:06:06,910 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-1808-m/10.140.0.6:10200
2020-11-22 15:06:06,921 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
to job history server
2020-11-22 15:06:06,960 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-1808-m/10.140.0.6:8032
2020-11-22 15:06:06,961 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-1808-m/10.140.0.6:10200
2020-11-22 15:06:06,987 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
to job history server
2020-11-22 15:06:07,043 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> 
```

- **Query them with HIVE**

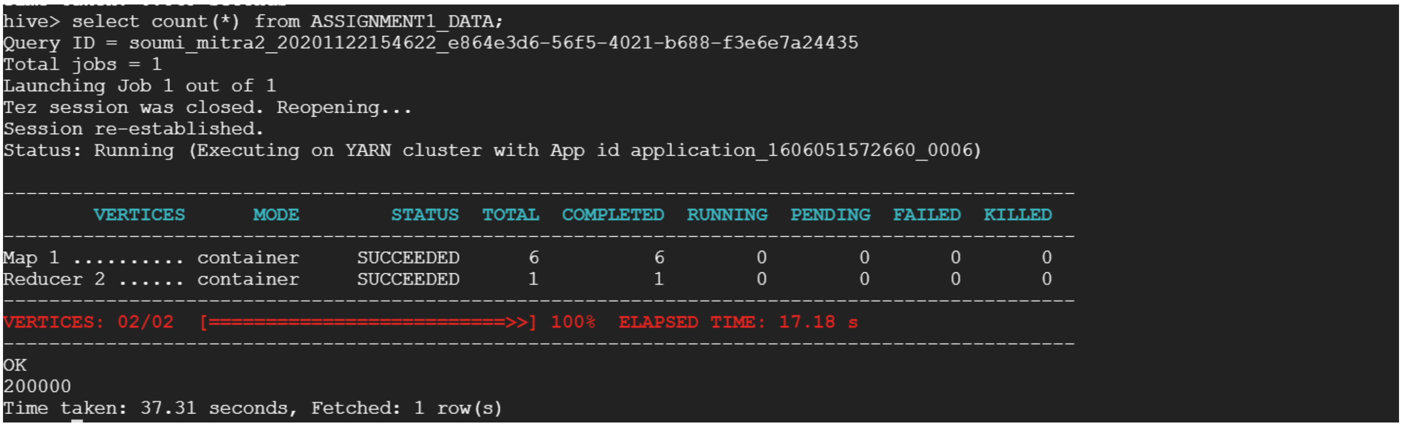
--- Database is created

Create database cloud_assignment;

---- table is created with the data stored with PIG

```
Create table final_data (
Id int,
PostTypeId int,
Score int,
ViewCount int,
Body string,
OwnerUserId int,
OwnerDisplayName string,
Title string,
AnswerCount int)
row format delimited
fields terminated by ','
location '/user/soumi_mitra2/Final1';
```

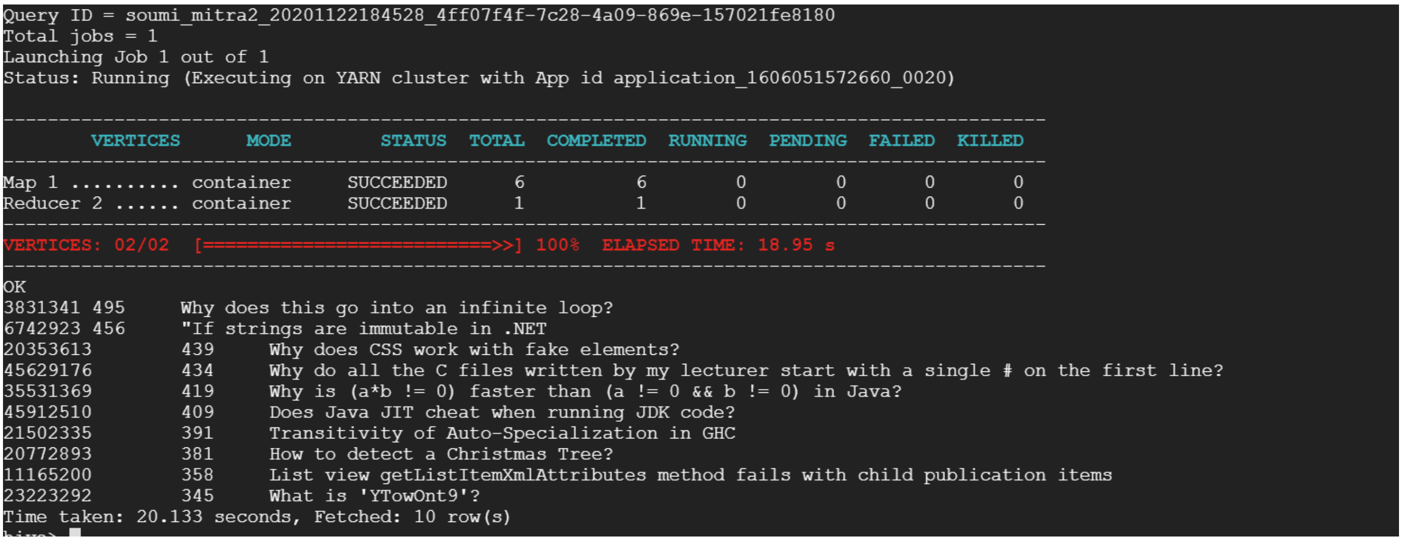
```
Select count(*) from final_data;
```



```
hive> select count(*) from ASSIGNMENT1_DATA;
Query ID = soumi_mitra2_20201122154622_e864e3d6-56f5-4021-b688-f3e6e7a24435
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1606051572660_0006)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 17.18 s
----------------------------------------------------------------------------------------------
OK
200000
Time taken: 37.31 seconds, Fetched: 1 row(s)
```

---- **Quering Top 10 posts by score**

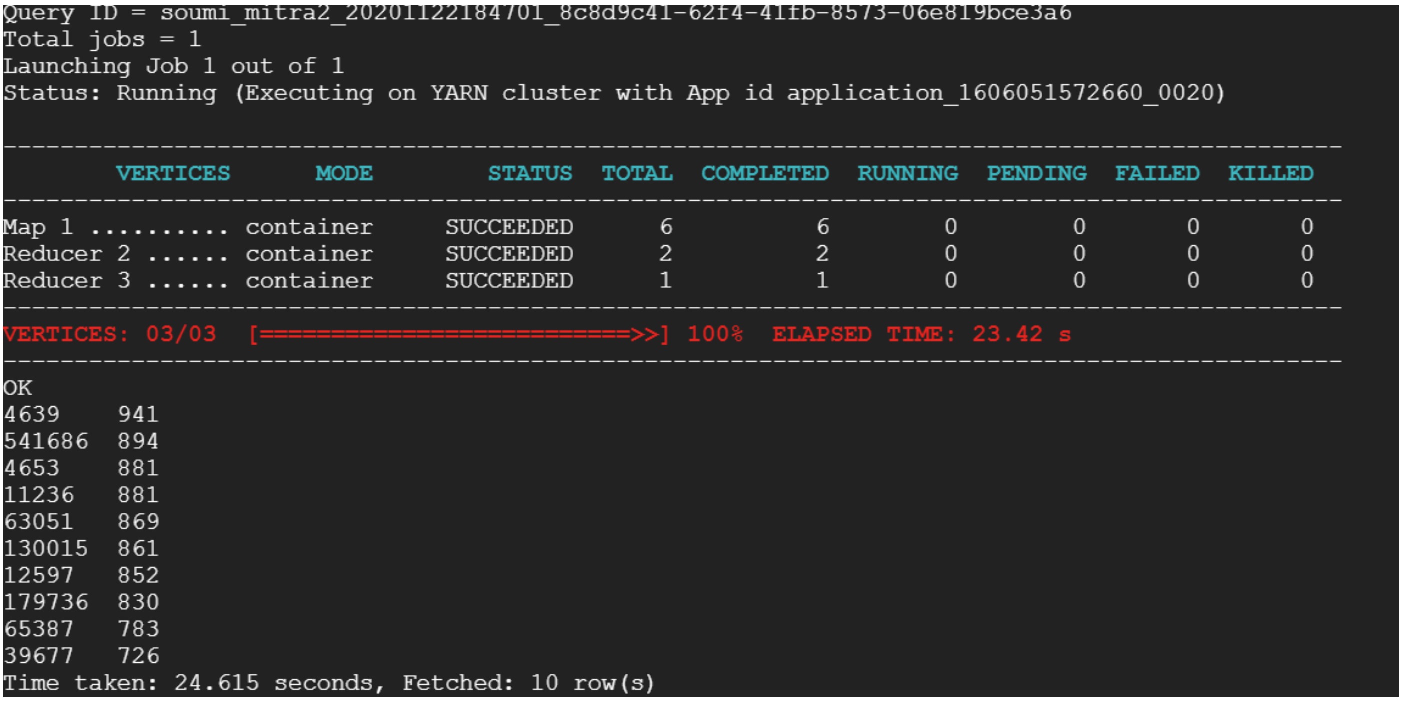
```
SELECT Id,Score,Title
FROM final_data
ORDER BY Score
DESC LIMIT 10;
```



```
Query ID = soumi_mitra2_20201122184528_4ff07f4f-7c28-4a09-869e-157021fe8180
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1606051572660_0020)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 18.95 s
----------------------------------------------------------------------------------------------
OK
3831341 495     Why does this go into an infinite loop?
6742923 456     "If strings are immutable in .NET
20353613        439     Why does CSS work with fake elements?
45629176        434     Why do all the C files written by my lecturer start with a single # on the first line?
35531369        419     Why is (a*b != 0) faster than (a != 0 && b != 0) in Java?
45912510        409     Does Java JIT cheat when running JDK code?
21502335        391     Transitivity of Auto-Specialization in GHC
20772893        381     How to detect a Christmas Tree?
11165200        358     List view getListItemXmlAttributes method fails with child publication items
23223292        345     What is 'YTowOnt9'?
Time taken: 20.133 seconds, Fetched: 10 row(s)
hive>
```

--- **The Top 10 users by Post Score**

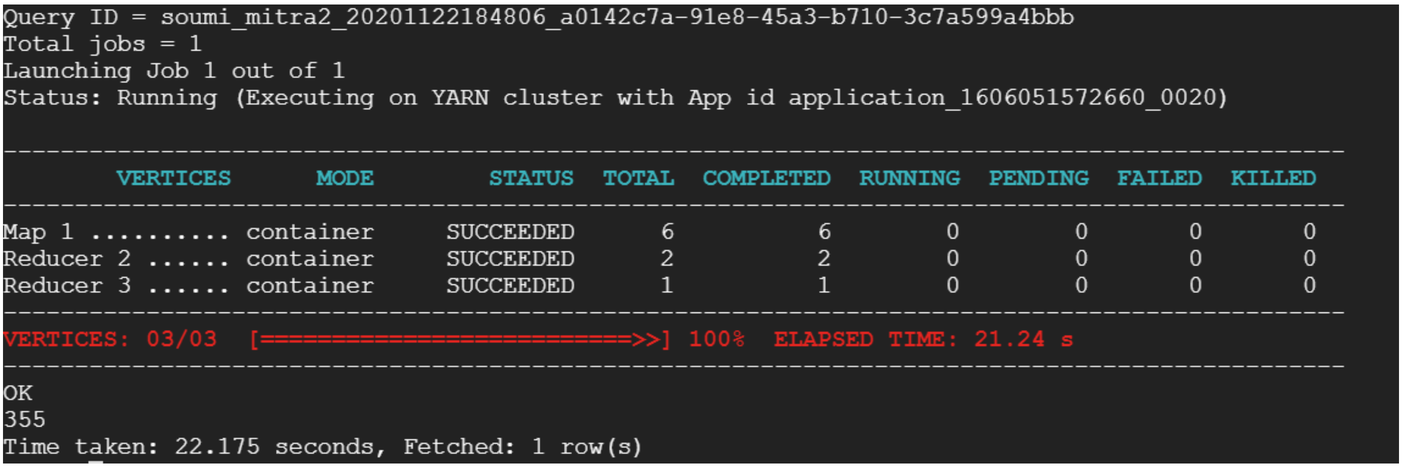
SELECT owneruserid, sum(score) as score
FROM final_data
WHERE owneruserid is not null
GROUP BY owneruserid
ORDER BY Score
DESC LIMIT 10;



```
Query ID = soumi_mitra2_20201122184701_8c8d9c41-62f4-41fb-8573-06e819bce3a6
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1606051572660_0020)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 23.42 s
----------------------------------------------------------------------------------------------
OK
4639    941
541686  894
4653    881
11236   881
63051   869
130015  861
12597   852
179736  830
65387   783
39677   726
Time taken: 24.615 seconds, Fetched: 10 row(s)
```

--- **The number of distinct users, who used the word “Hadoop” in one of their posts**

SELECT COUNT(*) FROM (
SELECT DISTINCT(OwnerUserId)
FROM final_data
WHERE (Body REGEXP 'hadoop')
OR (Title REGEXP 'hadoop')) as A;



```
Query ID = soumi_mitra2_20201122184806_a0142c7a-91e8-45a3-b710-3c7a599a4bbb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1606051572660_0020)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 21.24 s
----------------------------------------------------------------------------------------------
OK
355
Time taken: 22.175 seconds, Fetched: 1 row(s)
```

- Reference: https://data.stackexchange.com/