Cloud Assignment - 1

Name: Prerna Kamboj
Student Id: 20210716
Email: prerna.kamboj2@mail.dcu.ie

1. GET DATA FROM STACK EXCHANGE:

Performed the following queries to get 200,000 records.
Queries:

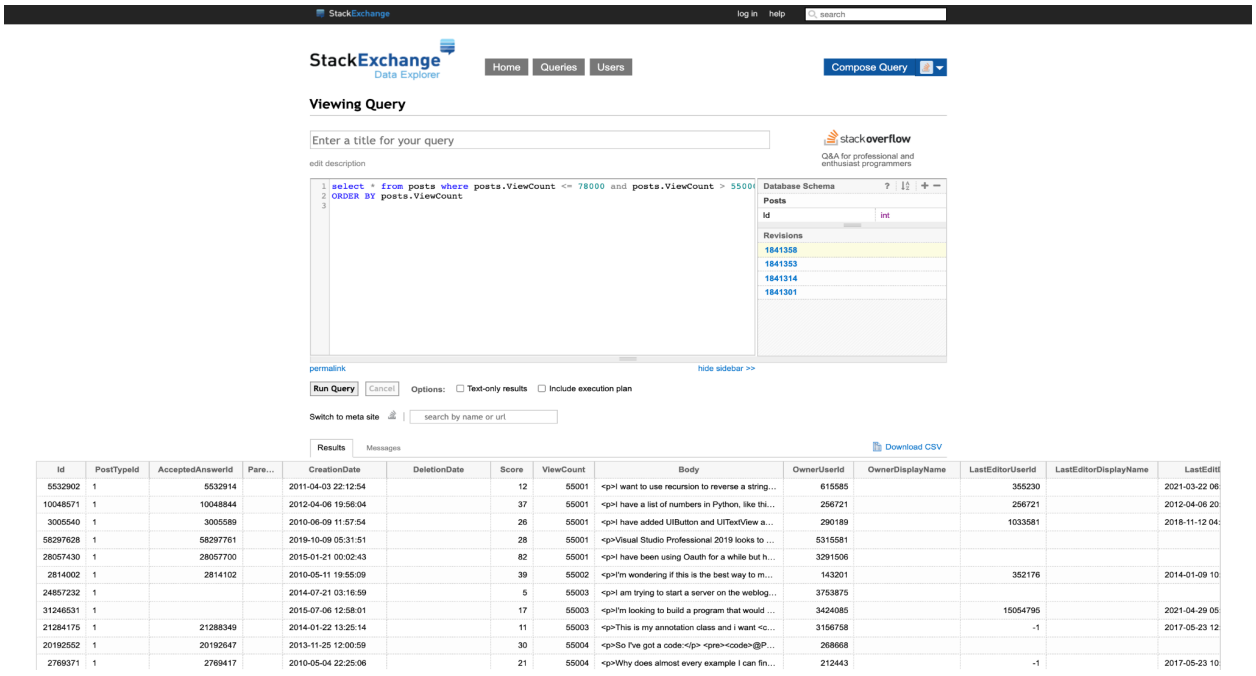
select * from posts where posts.ViewCount > 130000
ORDER BY posts.ViewCount

select * from posts where posts.ViewCount <= 130000 and posts.ViewCount > 78000
ORDER BY posts.ViewCount

select * from posts where posts.ViewCount <= 78000 and posts.ViewCount > 55000
ORDER BY posts.ViewCount

select * from posts where posts.ViewCount <= 55000 and posts.ViewCount > 42500
ORDER BY posts.ViewCount

select * from posts where posts.ViewCount <= 42500 and posts.ViewCount > 40500
ORDER BY posts.ViewCount



Id	PostTypeId	AcceptedAnswerId	Pare...	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDisplayName	LastEditorUserId	LastEditorDisplayName	LastEditI
5532902	1	5532914		2011-04-03 22:12:54		12	55001	I want to use recursion to reverse a string…	615585		355230		2021-03-22 06
10048571	1	10048844		2012-04-06 19:56:04		37	55001	I have a list of numbers in Python, like thi…	256721		256721		2012-04-06 20
3005540	1	3005589		2010-06-09 11:57:54		26	55001	I have added UIButton and UITextView a…	290189		1033581		2018-11-12 04
58297628	1	58297761		2019-10-09 05:31:51		28	55001	Visual Studio Professional 2019 looks to …	5315581				
28057430	1	28057700		2015-01-21 00:02:43		82	55001	I have been using Oauth for a while but h…	3291506				
2814002	1	2814102		2010-05-11 19:55:09		39	55002	I'm wondering if this is the best way to m…	143201		352176		2014-01-09 10
24857232	1			2014-07-21 03:16:59		5	55003	I am trying to start a server on the weblog…	3753875				
31246531	1			2015-07-06 12:58:01		17	55003	I'm looking to build a program that would …	3424085		15054795		2021-04-29 05
21284175	1	21288349		2014-01-22 13:25:14		11	55003	This is my annotation class and i want <c…	3156758		-1		2017-05-23 12
20192552	1	20192647		2013-11-25 12:00:59		30	55004	So I've got a code: <pre><code>@P…	268668				
2769371	1	2769417		2010-05-04 22:25:06		21	55004	Why does almost every example I can fin…	212443		-1		2017-05-23 10

2. SORTING, MERGING AND CLEANING OF THE FILES :

The csv files were first sorted in the descending order with respect to the ViewCount and then data of QueryResults5.csv was dropped to bring it to the total count of 200,000 records

Sorting of the files:

In [ ]:
```
#### Sorting of Fourth CSV
```

In [7]:
```
import pandas as pd
csvData3 = pd.read_csv('/Users/prernakamboj/Downloads/data/QueryResults4.csv')
csvData3.sort_values(["ViewCount"], ascending=[False], inplace=True)
csvData3.to_csv('/Users/prernakamboj/Downloads/data/QueryResults4.csv')
```

In [ ]:
```
#### Sorting of Fifth CSV
```

In [ ]:
```
import pandas as pd
csvData = pd.read_csv('/Users/prernakamboj/Downloads/data/QueryResults5.csv')
csvData.sort_values(["ViewCount"], ascending=[False], inplace=True)
csvData.drop(csvData.tail(5087).index,inplace=True)
csvData.to_csv('/Users/prernakamboj/Downloads/data/QueryResults5.csv')
```

After the files were sorted they were merged into a csv file **Merged.csv**

```
#### Merging of all 5 csv files
```

```
import pandas as pd
from glob import glob
Querydata_files = sorted(glob('/Users/prernakamboj/Downloads/data/QueryResults*.csv'))
merged_file = pd.concat(pd.read_csv(Querydatafiles)
                        for Querydatafiles in Querydata_files)
merged_file.to_csv('/Users/prernakamboj/Downloads/data/Merged.csv')
```

The record count was checked after the merging of the csv files

```
len(merged_file)
```

```
200000
```

Data Cleaning:

The data cleaning was done using a python script. The procedure began by reading the csv file using pandas and then the csv file was converted into a dataframe and further dataframe was converted into a Dictionary. After that, we looped through the dictionary and values from the columns were picked, thereafter commas ,extra spaces and newlines were removed and then the cleaned data was again put back to the dictionary as shown in the code below. These steps were followed for all the mentioned rows and after that data was stored back creating a new csv (Merged_Cleaned.csv).

```
import pandas as pd
import re

def main():
    data_csv = pd.read_csv('/Users/prernakamboj/Downloads/data/Merged.csv', usecols=['Id', 'Score', 'Body', 'OwnerUserI
    data_dict = data_csv.to_dict()

    count = len(data_dict['Id'])


    for i in range(count):
        data_dict['Body'][i] =  re.sub(r',+', '', data_dict['Body'][i])
        data_dict['Title'][i] = re.sub(r',+', '', data_dict['Title'][i])
        data_dict['Tags'][i] =  re.sub(r',+', '', data_dict['Tags'][i])

        data_dict['Body'][i] =  re.sub(r'(  )+', '', data_dict['Body'][i])
        data_dict['Title'][i] = re.sub(r'(  )+', '', data_dict['Title'][i])
        data_dict['Tags'][i] =  re.sub(r'(  )+', '', data_dict['Tags'][i])

        data_dict['Body'][i] =  re.sub(r'\n+', '', data_dict['Body'][i])
        data_dict['Title'][i] = re.sub(r'\n+', '', data_dict['Title'][i])
        data_dict['Tags'][i] =  re.sub(r'\n+', '', data_dict['Tags'][i])


    pd.DataFrame.from_dict(data_dict).to_csv('/Users/prernakamboj/Downloads/data/Merged_Cleaned.csv', index = False)


if __name__ == "__main__":
    main()
```

3. QUERY THEM WITH HIVE :

Explanation :-

An external table named postsdata1 is created using a merged and cleaned csv file.

Query :-

CREATE EXTERNAL TABLE postsdata1(id BIGINT,score BIGINT,body STRING,owneruserid BIGINT,title STRING,tags STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION 'gs://da-ass1-bucket/';

Output :-

```
0: jdbc:hive2://localhost:10000/default> CREATE EXTERNAL TABLE postsdata1
. . . . . . . . . . . . . . . . . . . .>        (id BIGINT,
. . . . . . . . . . . . . . . . . . . .>         score BIGINT,
. . . . . . . . . . . . . . . . . . . .>         body STRING,
. . . . . . . . . . . . . . . . . . . .>         owneruserid BIGINT,
. . . . . . . . . . . . . . . . . . . .>         title STRING,
. . . . . . . . . . . . . . . . . . . .>         tags STRING)
. . . . . . . . . . . . . . . . . . . .>       ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
. . . . . . . . . . . . . . . . . . . .>       LOCATION 'gs://da-ass1-bucket/';
No rows affected (1.051 seconds)
0: jdbc:hive2://localhost:10000/default>
```

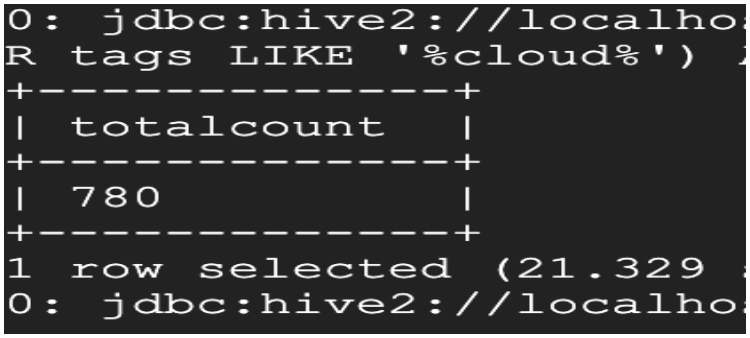
a) The top ten posts by score.
Explanation:- The data of top 10 users is extracted with respect to Score and printed along with the Id and Title.
Query:-
SELECT Id, Title, Score from postdata1 ORDER BY Score desc limit 10;

```
+-----------+-------------------------------------------------------+--------+
|    id     |                         title                         | score  |
+-----------+-------------------------------------------------------+--------+
| 11227809  | Why is processing a sorted array faster than processing an unsorted array? | 25933  |
| 927358    | How do I undo the most recent local commits in Git? | 23348  |
| 2003505   | How do I delete a Git branch locally and remotely? | 18514  |
| 292357    | What is the difference between 'git pull' and 'git fetch'? | 12834  |
| 231767    | "What does the ""yield"" keyword do?"                 | 11551  |
| 477816    | What is the correct JSON content type?                | 10921  |
| 348170    | How do I undo 'git add' before commit?                | 10079  |
| 5767325   | How can I remove a specific item from an array?       | 9931   |
| 6591213   | How do I rename a local Git branch?                   | 9792   |
| 1642028   | "What is the ""-->"" operator in C/C++?"              | 9560   |
+-----------+-------------------------------------------------------+--------+
10 rows selected (10.893 seconds)
0: jdbc:hive2://localhost:10000/default>
```

b) The top 10 users by post score.
Explanation:- The data of top 10 users is extracted with respect to post score which was displayed as TotalScore along with the OwnerUserId.
Query:-
SELECT OwnerUserId, SUM(Score) AS TotalScore FROM postdata1 GROUP BY OwnerUserId ORDER BY TotalScore DESC LIMIT 10;

```
0: jdbc:hive2://localhost:10000/default> SELECT OwnerUserId, SUM(Score) AS TotalScore FRO
BY TotalScore DESC LIMIT 10;
+--------------+-------------+
| owneruserid  | totalscore  |
+--------------+-------------+
| 87234        | 37672       |
| 4883         | 28817       |
| 9951         | 26799       |
| 6068         | 25944       |
| 89904        | 24024       |
| 51816        | 23719       |
| 49153        | 20203       |
| 179736       | 19530       |
| 95592        | 19479       |
| 63051        | 19345       |
+--------------+-------------+
10 rows selected (20.548 seconds)
```

c) The number of distinct users, who used the word “cloud” in one of their posts.
Explanation:- The number of distinct users who used the word count in their posts are extracted from the data and their final count is printed as TotalCount.
Query:-
SELECT COUNT(DISTINCT OwnerUserId) AS TotalCount FROM postsdata1 WHERE (body LIKE '%cloud%' OR title LIKE '%cloud%' OR tags LIKE '%cloud%');



REFERENCES :-

1.https://www.w3schools.com/python/python_regex.asp

2. Video reference for hive installation on GCP :- https://www.youtube.com/watch?v=Nr0vc-OIH1k

3. https://hackersandslackers.com/pandas-dataframe-drop/