Name : **Sarthak Santosh Shinde**
Student Id : 21261592
Email Id : sarthak.shinde4@mail.dcu.ie

Link For Git Repository :
https://github.com/sarthakshinde1998/CloudAssignment

Data Description : This Data contains attributes like User ID , Body , Score ,Tags etc.This data is taken from **social medial** like Platform and provides us the various details of Users.

Task 1 : Get Data from Stack Exchange.To download the data I used the Data Explorer feature of Stack Exchange using the given link.

Link : https://data.stackexchange.com/stackoverflow/query/new

Technology used : I used SQL queries to acquire Data from Data Explorer of Stack Exchange.

Queries :

1)select top 50000 * from posts ORDER BY posts.ViewCount DESC

2)select top 50000 * from post where posts.ViewCount < 127754 ORDER by posts.ViewCount DESC

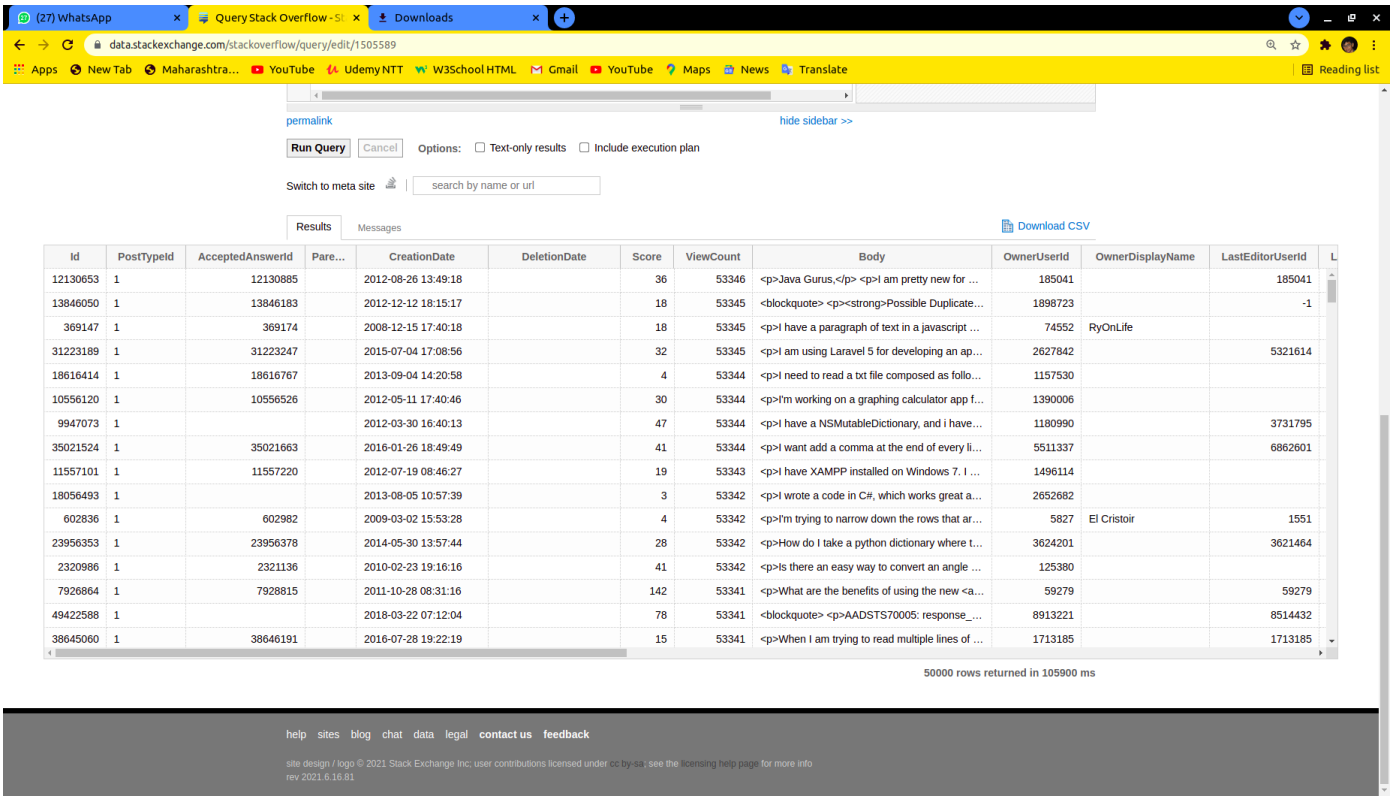
3)select top 50000 * from posts where posts.ViewCount < 74785 ORDER BY posts.ViewCount DESC

4)select top 50000 * from posts where posts.ViewCount < 41425 order by posts.ViewCount

Query_Discription : When we run the 1st given query on StackExchange we get the top 50000 posts according to the View Count per Post.After which we have to run the second query according to the last recorded view count of that data set which was obtained in the last query and so on.Here on StackExchange we can download at max 50000 posts in a single query.And Since we have to acquire the top 200000 posts we need to run 4 queries in all.For better understanding I have attached the the sample output screenshots of the Resulted Querry.

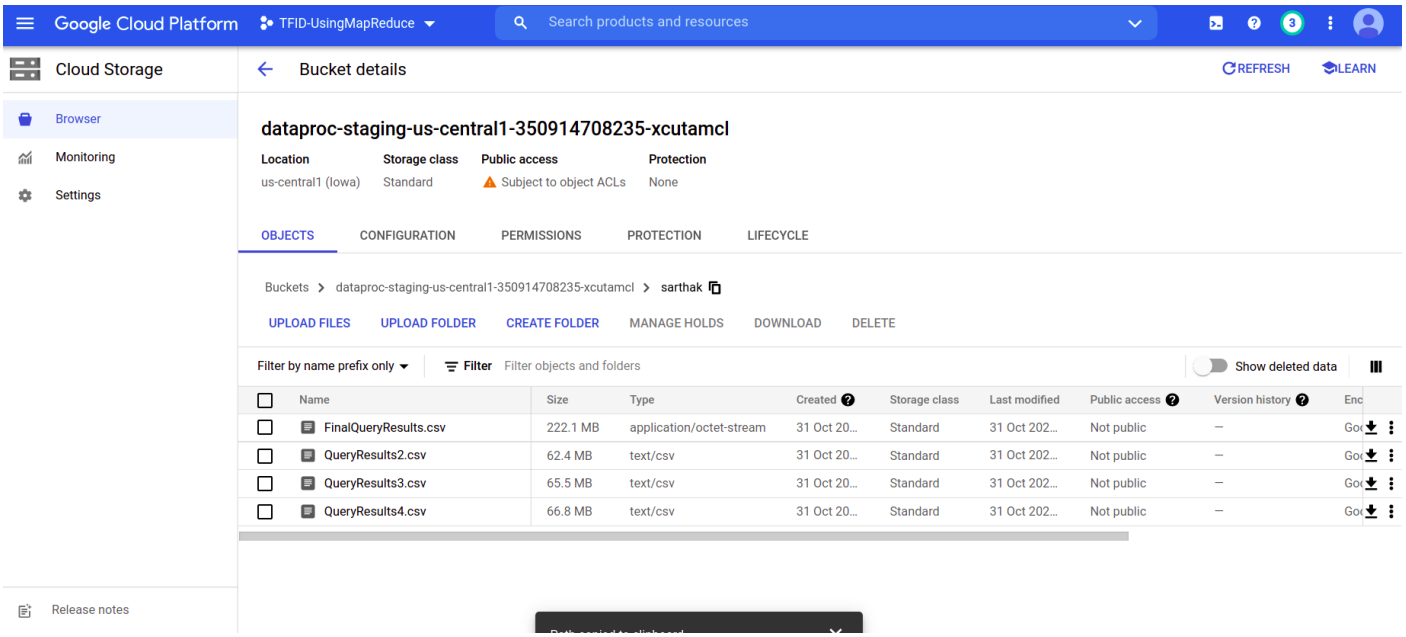
Id	PostTypeId	AcceptedAnswerId	Pare…	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDisplayName	LastEditorUserId
1311578	1	1311594		2009-08-21 12:09:56		36	74802	I'm trying to open multiple files at once wit…	152598		1525
14480616	1	14481087		2013-01-23 13:15:02		21	74802	I'm getting the following error <pre><…	1426157		4223
5027687	1	5027699		2011-02-17 10:26:46		67	74801	I have a table filled with a lot of rows and …	192310		745
5582155	1	5582192		2011-04-07 13:47:49		29	74799	i have xml what i get as byte array, whats …	655134		
10687099	1	10687158		2012-05-21 14:23:29		88	74798	How can I test a URL if it is a relative or a…	235334		
12534694	1	12534904		2012-09-21 16:47:01		316	74797	In my Visual Studio 2012 Solution Explor…	131270		27564
2583139	1	7493020		2010-04-06 06:52:42		24	74796	I'd like to have a submit button that submi…	65387		653
5939353	1	5939393		2011-05-09 15:51:21		17	74796	I am developing an ASP.NET MVC 3 web …	675082		
3130375	1	3130425		2010-06-28 06:06:54		64	74795	I'm writing a script to backup a database. …	39905		68626
49964093	1			2018-04-22 08:42:36		50	74794	I am trying to run a webpage using pytho…	9208277		44209
5699137	1	5699150		2011-04-18 06:07:29		14	74794	<pre><code>$stringText = "[TEST-1] test task…	506516		
16033448	1	16033595		2013-04-16 09:32:17		19	74794	I'm trying to get orders from an orderview.…	1635767		16357
5159065	1	5159405		2011-03-01 19:01:51		50	74792	Looking to add in vertical space between …	632729		3167
6509600	1	6509769		2011-06-28 16:12:08		250	74790	How can I resolve this warning? <blo…	19875		12612
998662	1			2009-06-15 22:06:02		48	74786	I hear this term sometimes and am wond…	31327		584
13468286	1	13468456		2012-11-20 07:17:56		48	74786	Consider the following Linux kernel dump…	1031417		38894

Figure(Sample_Output1)



Id	PostTypeId	AcceptedAnswerId	Pare…	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDisplayName	LastEditorUserId
12130653	1	12130885		2012-08-26 13:49:18		36	53346	Java Gurus, I am pretty new for …	185041		185041
13846050	1	13846183		2012-12-12 18:15:17		18	53345	<blockquote> Possible Duplicate…	1898723		-1
369147	1	369174		2008-12-15 17:40:18		18	53345	I have a paragraph of text in a javascript …	74552	RyOnLife	
31223189	1	31223247		2015-07-04 17:08:56		32	53345	I am using Laravel 5 for developing an ap…	2627842		5321614
18616414	1	18616767		2013-09-04 14:20:58		4	53344	I need to read a txt file composed as follo…	1157530		
10556120	1	10556526		2012-05-11 17:40:46		30	53344	I'm working on a graphing calculator app f…	1390006		
9947073	1			2012-03-30 16:40:13		47	53344	I have a NSMutableDictionary, and i have…	1180990		3731795
35021524	1	35021663		2016-01-26 18:49:49		41	53344	I want add a comma at the end of every li…	5511337		6862601
11557101	1	11557220		2012-07-19 08:46:27		19	53343	I have XAMPP installed on Windows 7. I …	1496114		
18056493	1			2013-08-05 10:57:39		3	53342	I wrote a code in C#, which works great a…	2652682		
602836	1	602982		2009-03-02 15:53:28		4	53342	I'm trying to narrow down the rows that ar…	5827	El Cristoir	1551
23956353	1	23956378		2014-05-30 13:57:44		28	53342	How do I take a python dictionary where t…	3624201		3621464
2320986	1	2321136		2010-02-23 19:16:16		41	53342	Is there an easy way to convert an angle …	125380		
7926864	1	7928815		2011-10-28 08:31:16		142	53341	What are the benefits of using the new <a…	59279		59279
49422588	1			2018-03-22 07:12:04		78	53341	<blockquote> AADSTS70005: response_…	8913221		8514432
38645060	1	38646191		2016-07-28 19:22:19		15	53341	When I am trying to read multiple lines of …	1713185		1713185

Figure(Sample_Output2)

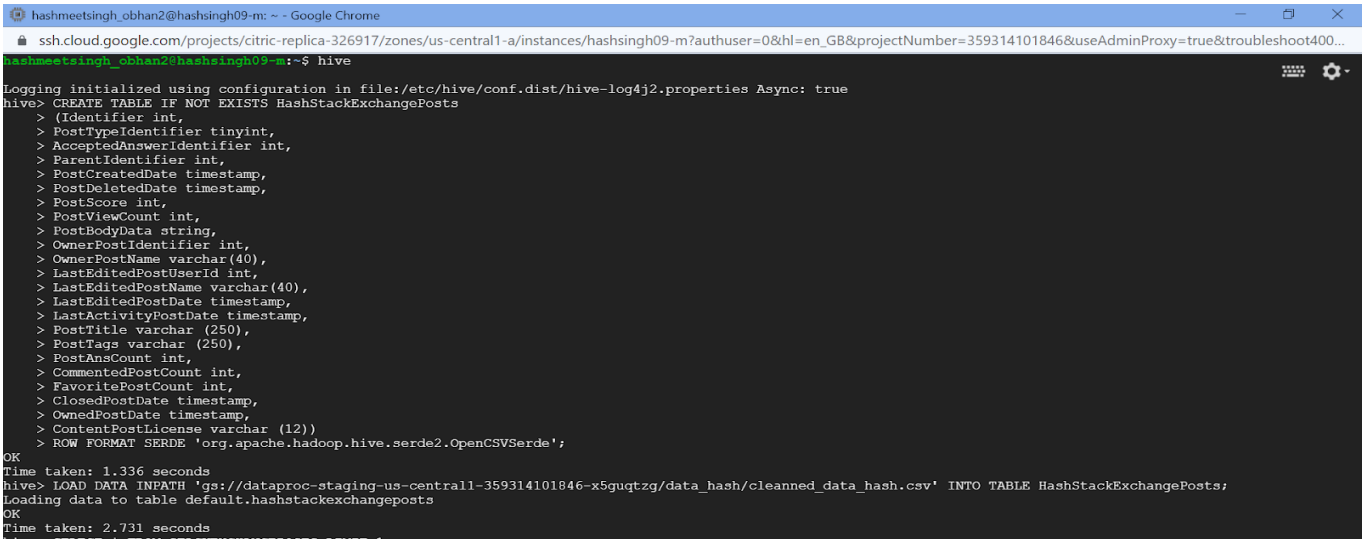


Figure(Successfully uploaded all four data files and Merge them into one on GCP)

Task 2 : Load Data into Chosen Technology(Hive).

Once the data is uploaded and Cleaned create table on HIVE which is pre-installed on GCP.Then I loaded the data in the table created on HIVE after which I created a View Similar to the Table Structure.

Technology Used : I used HIVE because it was easier to execute the task comparatively to other Technologies and less time consuming and on top that the use of SQL queries in HIVE makes it more familiar to me as I have used SQL in the past for Development and Experiments.



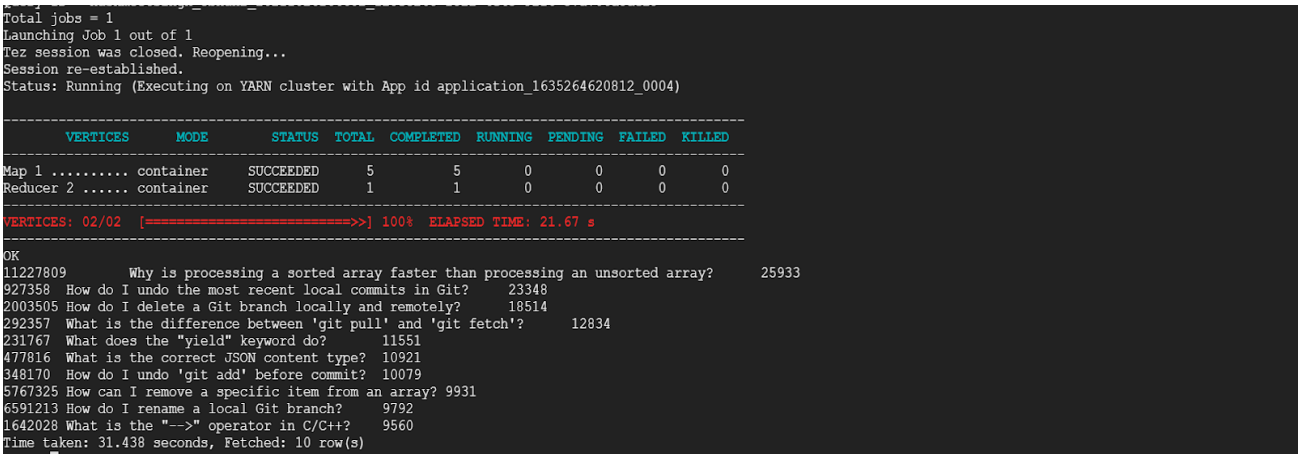
Command to Load Data : LOAD DATA INPATH "gs://dataproc-staging-us-central1-350914708235-xcutamcl/sarthak/FinalQueryResults.csv" INTO TABLE SarthakDB;

Task 3 : Run the Query Data Using HIVE.

Technology Used : Similar to Task 2 , I have Used HIVE to execute all the task 3 queries.

a)Top 10 posts by Score (Run the Below Query to Obtain the following attached output)

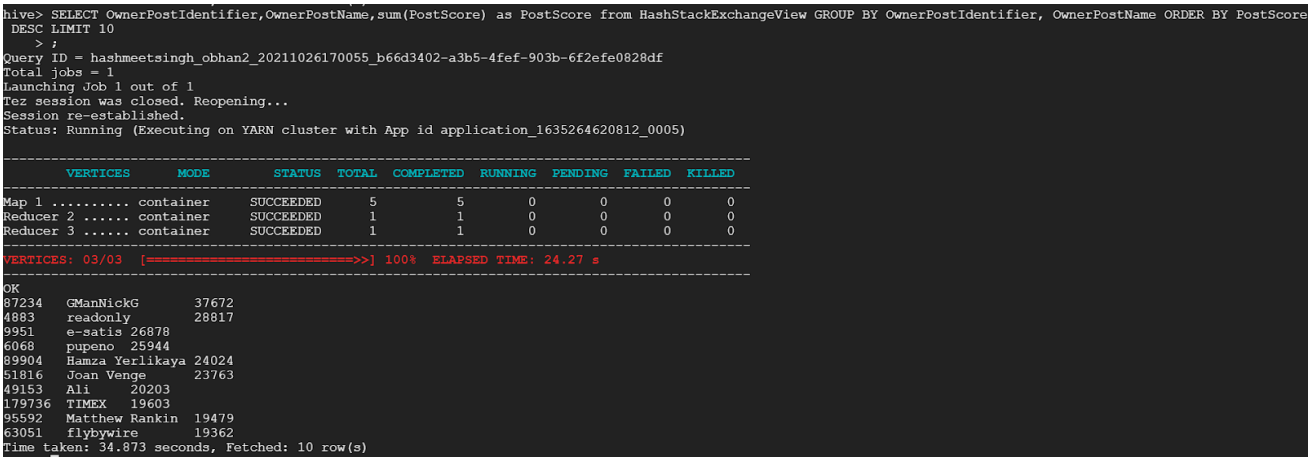
SELECT Id, Title, Score from SarthakDBView ORDER BY Score DESC LIMIT 10;



Figure(Output of the Given Query)

b)Top 10 users by Post Score (Run the Below Query to Obtain the following attached output)

SELECT OwnerUserId,OwnerDisplayName,sum(Score) as Score from SarthakDBView GROUP BY OwnerUserId, OwnerDisplayName ORDER BY Score DESC LIMIT 10;



Figure(Output of the Given Query)

c)The number of distinct users that have the word “Cloud” in on of their Posts.

```
SELECT COUNT(DISTINCT OwnerUserId) as UniqueUsers FROM
SarthakDBView WHERE Title LIKE '% cloud %' OR Body LIKE '% cloud
%';
```

```
hive> SELECT COUNT(DISTINCT OwnerPostIdentifier) as TotalDistinctUsers FROM HashStackExchangeView WHERE PostTitle LIKE '% cloud %' OR PostBodyData LIKE '% cloud %';

Query ID = hashmeetsingh_obhan2_20211026171011_7364cfc4-8390-4cfc-bf13-15f2016f8cd8
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_1635264620812_0006)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..... container     SUCCEEDED      5          5        0        0       0       0
Reducer 2 ..... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ..... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 23.64 s
--------------------------------------------------------------------------------
OK
248
Time taken: 33.976 seconds, Fetched: 1 row(s)
```

Figure(Number of Distinct User with the Word Cloud in one of their Posts)

Task 4 : Calculate the per-user TF-IDF of the top 10 terms for each of the top 10 users.

Technology used : I have used Python to find the top 10 terms for each of the 10 users as it was simple to implement the code in python and also the code is more efficient.

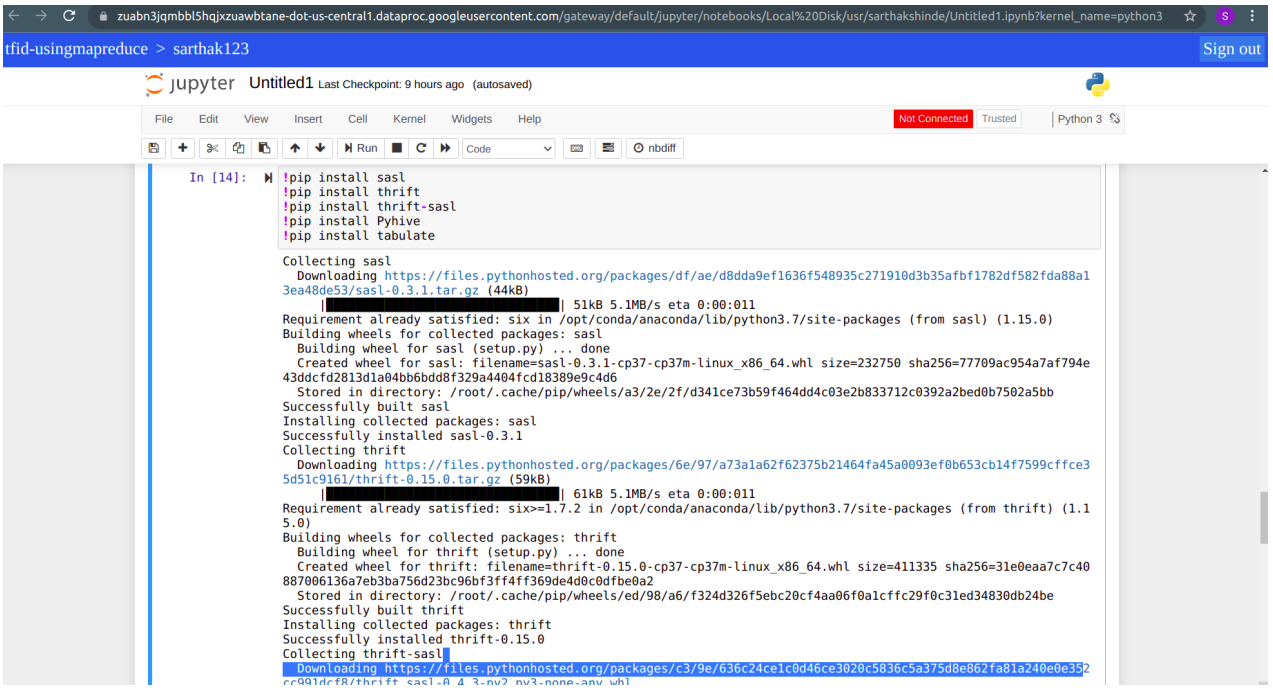
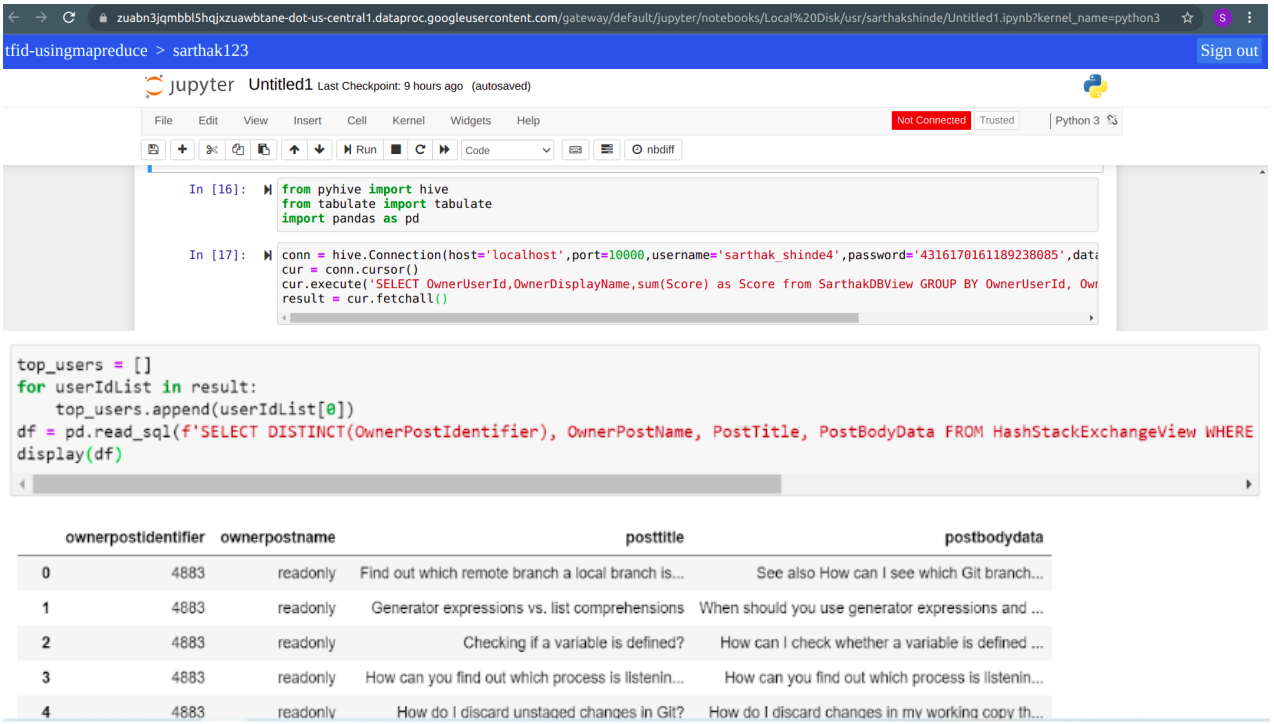


Figure : Installing the Required Python Libraries



Figure(Successfully Establishing connection with HIVE Database and found the top 10 Users)

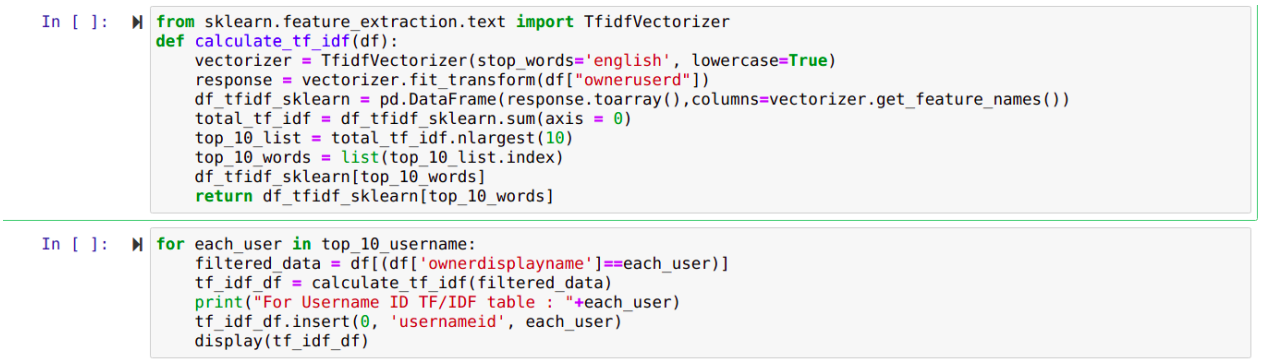


Figure (Successfully executed the top 10 terms for each of the 10 users)

Explanation of the Code Snippets: After installing the python libraries required and establishing connecting with HIVE database I found the top 10 user details based on post scores.After that I searched for the OwnerUserId in details that I fetched earlier for 10 users.Then I used sklearn library to find the TF_IDF for each user.