



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 sample output screenshots of the Resulted Query.

Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount	Body	OwnerUserId	OwnerDisplayName	LastEditorUserId
1311578	1	1311594		2009-08-21 12:09:56		36	74802	<p>I'm trying to open multiple files at once wit...	152598		152598
14480616	1	14481087		2013-01-23 13:15:02		21	74802	<p>I'm getting the following error</p> <pre><...</pre>	1426157		42235
5027687	1	5027699		2011-02-17 10:26:46		67	74801	<p>I have a table filled with a lot of rows and ...	192310		7458
5582155	1	5582192		2011-04-07 13:47:49		29	74799	<p>i have xml what i get as byte array, whats ...	655134		
10687099	1	10687158		2012-05-21 14:23:29		88	74798	<p>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	<p>In my Visual Studio 2012 Solution Explor...	131270		27564
2583139	1	7493020		2010-04-06 06:52:42		24	74796	<p>I'd like to have a submit button that submi...	65387		65387
5939353	1	5939393		2011-05-09 15:51:21		17	74796	<p>I am developing an ASP.NET MVC 3 web ...	675082		
3130375	1	3130425		2010-06-28 06:06:54		64	74795	<p>I'm writing a script to backup a database. ...	39905		68626
49964093	1			2018-04-22 08:42:36		50	74794	<p>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	<p>I'm trying to get orders from an orderview...	1635767		1635767
5159065	1	5159405		2011-03-01 19:01:51		50	74792	<p>Looking to add in vertical space between ...	632729		3167
6509600	1	6509769		2011-06-28 16:12:08		250	74790	<p>How can I resolve this warning?</p> <blo...	19875		12617
998662	1			2009-06-15 22:06:02		48	74786	<p>I hear this term sometimes and am wond...	31327		584
13468286	1	13468456		2012-11-20 07:17:56		48	74786	<p>Consider the following Linux kernel dump...	1031417		38894

Figure (Sample_Output1)

permalink [hide sidebar >>](#)

Run Query Options: ☐ Text-only results ☐ Include execution plan

Switch to meta site

Results [Messages](#) [Download CSV](#)

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

50000 rows returned in 105900 ms

help sites blog chat data legal **contact us** feedback

site design / logo © 2021 Stack Exchange Inc; user contributions licensed under cc by-sa see the licensing help page for more info
rev 2021.6.16.81

Figure (Sample_Output2)

Bucket details

dataproc-staging-us-central1-350914708235-xcutamcl

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Subject to object ACLs	None

OBJECTS | CONFIGURATION | PERMISSIONS | PROTECTION | LIFECYCLE

Buckets > dataproc-staging-us-central1-350914708235-xcutamcl > sarthak

UPLOAD FILES | UPLOAD FOLDER | CREATE FOLDER | MANAGE HOLDS | DOWNLOAD | DELETE

Filter by name prefix only | Filter | Filter objects and folders | Show deleted data

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Enc
FinalQueryResults.csv	222.1 MB	application/octet-stream	31 Oct 20...	Standard	31 Oct 202...	Not public	—	Go
QueryResults2.csv	62.4 MB	text/csv	31 Oct 20...	Standard	31 Oct 202...	Not public	—	Go
QueryResults3.csv	65.5 MB	text/csv	31 Oct 20...	Standard	31 Oct 202...	Not public	—	Go
QueryResults4.csv	66.8 MB	text/csv	31 Oct 20...	Standard	31 Oct 202...	Not public	—	Go

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.

```

hashmeetsingh_obhan2@hashsingh09-m: ~ - Google Chrome
ssh.cloud.google.com/projects/citric-replica-326917/zones/us-central1-a/instances/hashsingh09-m?authuser=0&hl=en_GB&projectNumber=359314101846&useAdminProxy=true&troubleshoot400...
hashmeetsingh_obhan2@hashsingh09-m:~$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> CREATE TABLE IF NOT EXISTS HashStackExchangePosts
> (Identifier int,
> PostTypeIdentifier tinyint,
> AcceptedAnswerIdentifier int,
> ParentIdentifier int,
> PostCreatedDate timestamp,
> PostDeletedDate timestamp,
> PostScore int,
> PostViewCount int,
> PostBodyData string,
> OwnerPostIdentifier int,
> OwnerPostName varchar(40),
> LastEditedPostUserId int,
> LastEditedPostName varchar(40),
> LastEditedPostDate timestamp,
> LastActivityPostDate timestamp,
> PostTitle varchar (250),
> PostTags varchar (250),
> PostAnsCount int,
> CommentedPostCount int,
> FavoritePostCount int,
> ClosedPostDate timestamp,
> OwnedPostDate timestamp,
> ContentPostLicense varchar (12))
> ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde';
OK
Time taken: 1.336 seconds
hive> LOAD DATA INPATH 'gs://dataproc-staging-us-central1-350914101846-x5guqtzg/data_hash/cleaned_data_hash.csv' INTO TABLE HashStackExchangePosts;
Loading data to table default.hashstackexchangeposts
OK
Time taken: 2.731 seconds

```

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;
```

```
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_0004)

-----
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
-----
VERTICES: 02/02  [=====] 100%  ELAPSED TIME: 21.67 s
-----
OK
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
Time taken: 31.438 seconds, Fetched: 10 row(s)
```

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;
```

```
hive> SELECT OwnerPostIdentifier,OwnerPostName,sum(PostScore) as PostScore from HashStackExchangeView GROUP BY OwnerPostIdentifier, OwnerPostName ORDER BY PostScore
DESC LIMIT 10
>
Query ID = hashmeetsingh_obhan2_20211026170055_b66d3402-a3b5-4fef-903b-6f2efe0828df
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_0005)

-----
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: 24.27 s
-----
OK
87234  GManNickG      37672
4883   readonly      28817
9951   e-satis 26878
6068   pupeno 25944
89904  Hamza Yerlikaya 24024
51816  Joan Venge     23763
49153  Ali           20203
179736 TIMEEX        19603
95592  Matthew Rankin 19479
63051  flybywire     19362
Time taken: 34.873 seconds, Fetched: 10 row(s)
```

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
SarathakDBView 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	0
Reducer 2	container	SUCCEEDED	1	1	0	0	0	0	0
Reducer 3	container	SUCCEEDED	1	1	0	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)

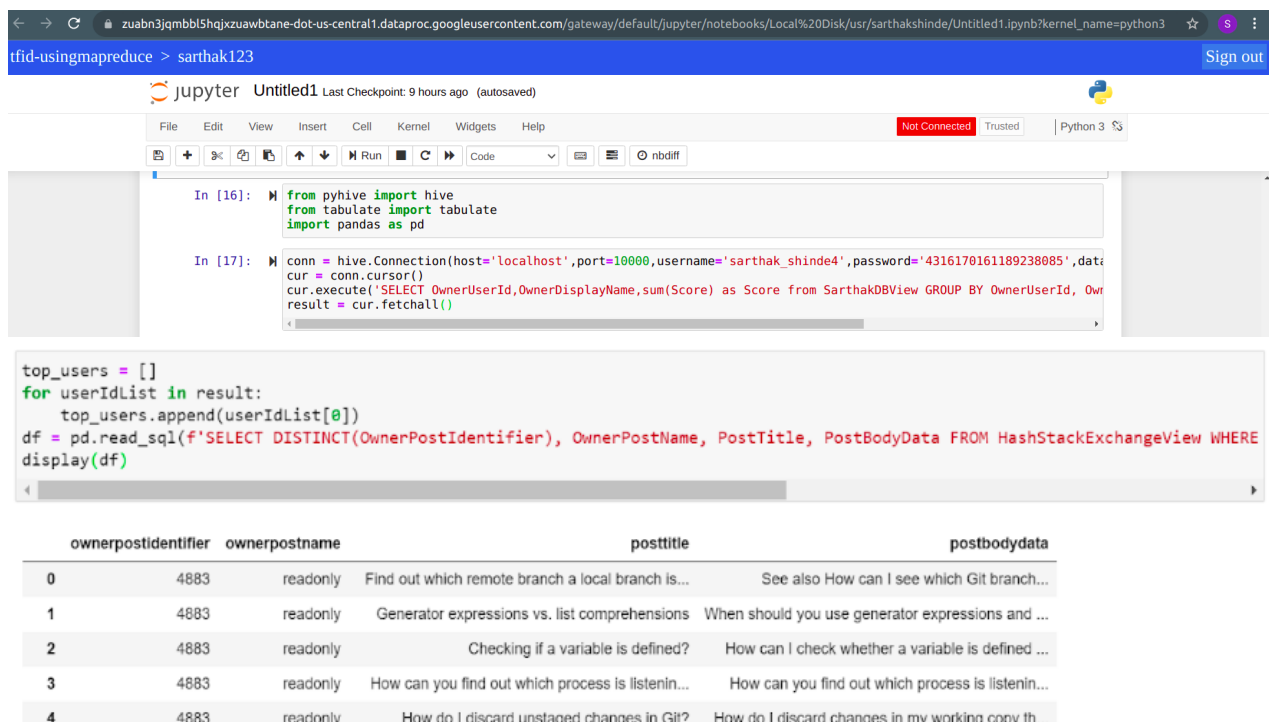
Task 4 : Calculate the per-user TF-IDF of the top 10 terms for each of the top 10 users.

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.

```
In [14]: pip install sas
pip install thrift
pip install thrift-sasl
pip install Pyhive
pip install tabulate

Collecting sas
  Downloading https://files.pythonhosted.org/packages/df/ae/d8da9ef1636f548935c271910d3b35afb1782df582fda88a13ea48de53/sas-0.3.1.tar.gz (44kB)
    [51kB 5.1MB/s eta 0:00:01]
Requirement already satisfied: six in /opt/conda/anaconda/lib/python3.7/site-packages (from sas) (1.15.0)
Building wheels for collected packages: sas
  Building wheel for sas (setup.py) ... done
  Created wheel for sas: filename=sas-0.3.1-cp37-cp37m-linux_x86_64.whl size=232750 sha256=77709ac954af794e43ddcfd2813da04bb6bdd8f329a4404fcd18389e9c4d6
  Stored in directory: /root/.cache/pip/wheels/a3/2e/2f/d341ce73b59f464dd4c03e2b833712c0392a2bed0b7502a5bb
Successfully built sas
Installing collected packages: sas
Successfully installed sas-0.3.1
Collecting thrift
  Downloading https://files.pythonhosted.org/packages/6e/97/a73a1a62f62375b21464fa45a0093ef0b653cb14f7599cffe35d51c9161/thrift-0.15.0.tar.gz (59kB)
    [61kB 5.1MB/s eta 0:00:01]
Requirement already satisfied: six>=1.7.2 in /opt/conda/anaconda/lib/python3.7/site-packages (from thrift) (1.15.0)
Building wheels for collected packages: thrift
  Building wheel for thrift (setup.py) ... done
  Created wheel for thrift: filename=thrift-0.15.0-cp37-cp37m-linux_x86_64.whl size=411335 sha256=31e0ea7c7c40887006136a7eb3ba756d23bc96bf3ff4f369de4d0c0dfbe0a2
  Stored in directory: /root/.cache/pip/wheels/ed/98/a6/f324d326f5ebc20cf4aa06f0a1cffe29f0c31ed34830db24be
Successfully built thrift
Installing collected packages: thrift
Successfully installed thrift-0.15.0
Collecting thrift-sasl
  Downloading https://files.pythonhosted.org/packages/c3/9e/636c24ce1c0d46ce3020c5836c5a375d8e862fa81a240e0e352cc0q1drf8/thrift-sasl-0.4.3-nv2-nv3-nnp-nv whl
```

Figure : Installing the Required Python Libraries



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

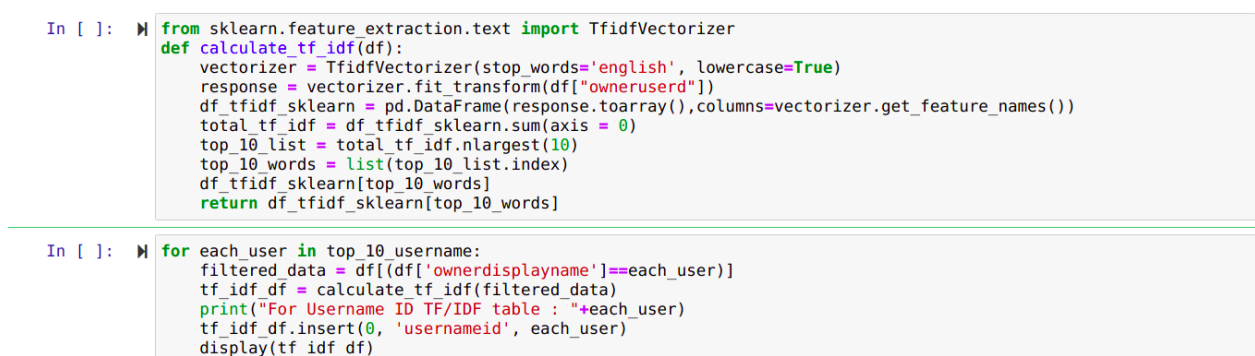


Figure (Successfully executed the top 10 terms for each of the 10 users)

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.

