DATA ANALYSIS - CLOUD TECHNOLOGIES

STACK EXCHANGE DATA ANALYSIS (MAPREDUCE/PIG/HIVE) – ASSIGNMENT 1

CA675 CLOUD TECHNOLOGIES ASSIGNMENT – COLIN HEHIR 20213371

DETAILS

Name: *Colin Hehir*

Student ID: *20213371*

Email: *colin.hehir3@mail.dcu.ie*

GitHub Repository: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis

ABSTRACT

The goal of this project was to use Pig to conduct ETL on a diversified series of questions on a dataset from Stack Exchange and then use Hive in HDFS to examine and query to get specified datasets.

Task 1 - Acquire the top 200,000 posts by View Count.

Task 2 & 3 - Use Pig/Hive/MapReduce - Extract, Transform and Load the data as applicable to get:

1. The top 10 posts by score
2. The top 10 users by post score
3. The number of distinct users, who used the word “cloud” in one of their posts.

Task 4 - Use MapReduce/Pig/Hive to calculate the per-user TF-IDF of the top 10 terms for each of the top 10 users.

Stack Exchange is an internet community for answers to questions on a variety of topics and offers an open-source tool for running arbitrary queries against public data from the Stack Exchange network. Features include query editing for all public Stack Exchange sites. After downloading the data as per Task 1, it was input into Apache Pig for transformation and cleaning. When the data was ready, it was imported into the Hadoop file system and queried with Apache Hive. A Data-Proc cluster was used to implement this method on Google Cloud Platform (GCP).

TASK 1 - DATA ACQUISITION AND COLLECTION

The goal of Task 1 was to acquire data from Stack Exchange of the top 200,000 posts by View Count. Firstly, we identified what the lowest value of View Count there was in the top 200,000 posts. We then ran a query as demonstrated below and identified there were 200,002 posts with a View Count greater than 41231, including 199,999 posts with a View Count greater than 41232. This meant it was impossible to collect exactly the top 200,000 posts as there were three posts that had a View Count of 41232.

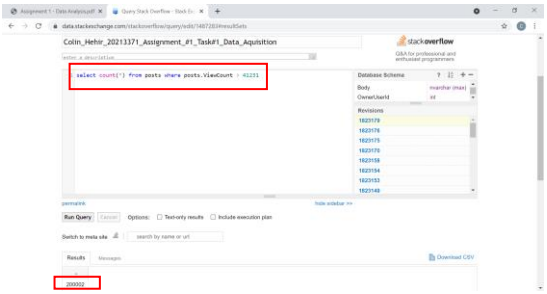


Figure 1: 200,002 rows returned when querying rows with View Count greater than 41231.

```
Count Query #1: select count(*) from posts
       where posts.ViewCount > 41231

Count Query #2: select count(*) from posts
       where posts.ViewCount > 41232
```

Due to the inherent functionality restrictions of Stack Exchange Data Explorer, an individual can only download a maximum of 50,000 records at a time. This meant we had to run 5 queries as can be seen below from the Stack Exchange site to get the 200,002 posts identified as above.

```
Query #1: select top 50000* from posts
where posts.ViewCount > 41231 order by
posts.ViewCount DESC

Query #2: select top 50000 * from posts
where posts.ViewCount <= 127078 order by
posts.ViewCount DESC

Query #3:  select top 50000 * from posts
where posts.ViewCount <= 74420 order by
posts.ViewCount DESC

Query #4: select top 50000 * from posts
where posts.ViewCount <= 53077 order by
posts.ViewCount DESC

Query #5: select top 50000 * from posts
where posts.ViewCount <= 41234 and
posts.ViewCount > 41231 order by
posts.ViewCount DESC
```

Please note the upper bound also has an equal to sign (<=) for Queries #2, #3, #4 and #5. This is to include the lowest view count value in the previous query as there may be posts with the same view count value, but it was not included in the dataset due to the limit in selecting the top 50,000 posts. Otherwise, using a less than sign (<) may exclude posts that are in the top 200,002 posts by View Count. As a result, there may be duplicates in the total dataset, however these will be identified and removed in the data processing, cleaning and integration section of this assignment and verified with a count of 200,002 for completeness and accuracy.

TECHNOLOGIES USED

GOOGLE CLOUD PLATFORM GCP

GCP is a flexible, open, secure and cost-effective way to utilize services such as DataProc as you can integrate your data lake meta store with open-source data clusters you build. Therefore, I created a Hadoop Cluster on GCP to perform the subsequent tasks.

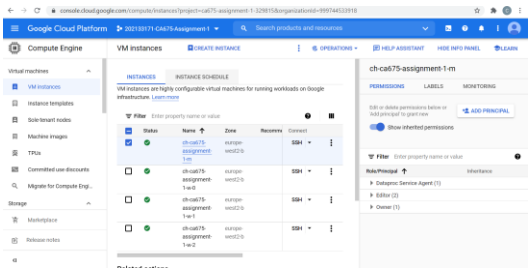


Figure 2: My GCP Hadoop Cluster for this assignment.

APACHE HADOOP (MAPREDUCE)

I chose to use Apache Hadoop as it is an open-source software framework that is reliable, scalable and fault-tolerant. I used it to analyze and process the big data I had and to enable distributed computing of data in a cluster on Google DataProc.

APACHE PIG

Pig was utilized to perform ETL and data processing on the data collected. Its ability to perform this made it a vital tool for this project.

APACHE HIVE

I used Hive to get specific answers from the respective dataset as it supports SQL-style queries and works off a structured schema built around tables, rows, columns and queries.

TASK 2 PERFORM ETL

After uploading the 5 .csv files to my cluster's default bucket using the 'Upload Folder' button, I then used the following command to load the data into HDFS. (Note X refers to file number)

```
$ hadoop fs -put /home/colin_hehir3/
Stack_Exchange_Query_Results_X.csv
/Data_Acquisition_Collection
```

After loading and merging the data using Pig utilizing the org.apache.pig.piggybank.storage. CSVExcelStorage function, I then used the following commands to remove any duplicates and verify that the new dataset has 200,002 values.

```
grunt> StackExchangeData_Total_Column_
Filter = FOREACH StackExchangeData_Total
GENERATE Id,Score,ViewCount,Body,
OwnerUserId,OwnerDisplayName,Title;

grunt> StackExchangeData_Total_Grouped =
GROUP StackExchangeData_Total_Column_
Filter BY Id;

grunt> StackExchangeData_Total_Distinct_ID
= FOREACH StackExchangeData_Total_Grouped
{result = TOP(1, 0, $1);GENERATE
FLATTEN(result);}
```

```
OK
200002
Time taken: 15.57 seconds, Fetched: 1 row(s)
hive>
```

Figure 3: Count (*) function on dataset showing 200,002 instances.

In order to prevent skewing and I then performed multiple commands to perform data cleaning and processing in order to remove unwanted symbols and data artefacts such as HTML tags and line breaks. This was completed on the 'Body' and 'Title' columns with the following generic code:

```
grunt> StackExchangeData_Cleaning_(X) =
FOREACH StackExchangeData_Cleaning_(X-1)
GENERATE Id,Score,ViewCount, REPLACE (Body
,'(Artefact)','') as Body,OwnerUserId,
OwnerDisplayName,REPLACE(Title,'(Artefact)
','')as Title;
```

The new data was then re-loaded using the following path: '*/Final_Dataset/part-r-00000*'.

TASK 3 QUERY DATA

A table was created in Hive before loading the data from in the previous section in Pig including only necessary columns.

```
hive> CREATE TABLE FINALDATASET (Id int, Score int, ViewCount int, Body string
TERMINATED BY ',' TBLPROPERTIES('skip.header.line.count'='1');
OK
Time taken: 1.72 seconds
hive> LOAD DATA INPATH '/Final_Dataset/part-r-00000' INTO TABLE FINALDATASET;
```

Figure 4: Creation of table and loading of data in Hive.

1. TOP 10 POSTS BY SCORE

To get the top ten posts from the Stack Exchange dataset by score, we selected the Id, Score and Title attributes in our query from our table created in hive and sorted it by the highest score, while also using the limit function to give the top 10 values.

ID	Score
11227809	25893
927358	23274
2003505	18451
292357	12796
231767	11512
477816	10894
348170	10045
5767325	9877
6591213	9747
1642028	9539

```
hive> select Id,
Score, Title from
DATASET sort by
Score desc limit
10;
```

```
Id      score   title
11227809        25893   Why is p
927358  23274   How do I undo th
2003505 18451   How do I delete
292357  12796   What is the diff
231767  11512   What does the yi
477816  10894   What is the corr
348170  10045   How do I undo 'g
5767325 9877    How can I remove
6591213 9747    How do I rename
1642028 9539    What is the -- o
Time taken: 18.117 seconds, Fetc
hive>
```

Table 1 and Figure 5: Output for top 10 posts by score.

2. TOP 10 USERS BY POST SCORE

In order to identify the top ten users by their post score from the Stack Exchange dataset, we selected the OwnerUserId attribute and created a TotalUserScore column in our query from our table created in hive and grouped it by OwnerUserId before ordering it by the highest TotalUserScore, while also using the limit function to give the top 10 values.

```
owneruserid     totaluserscore
87234   37606
4883    28155
9951    26728
6068    25860
89904   23949
51816   23632
179736  19415
95592   19413
63051   18738
49153   18541
Time taken: 12.875 seconds, Fetched: 10 row(s)
hive>
```

Figure 6: Output for top 10 users by post-score.

OwnerUserId	TotalUserScore
87234	37606
4883	28155
9951	26728
6068	25860
89904	23949
51816	23632
179736	19415
95592	19413
63051	18738
49153	18541

```
hive> select
OwnerUserId,
SUM(Score) as
TotalUser Score from
DATASET where
OwnerUserId is not
null group by
OwnerUserId order by
TotalUser Score desc
limit 10;
```

Table 2: Output for top 10 users by their total post-score.

3. USERS WHO USED "CLOUD"

For this task, I created a query for two scenarios based on different assumptions. I made use of the Lower () function in order to identify the word cloud even if it may contain upper case letters, while in assumption B I included white space, full stop, hyphen etc. to cover the full scope of different scenarios I where it might be in a sentence. Only 'Body' and 'Title' columns were considered whereas the 'Tags' column was excluded, as the user didn't actually use the word as such in that case in a post.

Assumption A: Match word cloud in any instance > Number of distinct users = **710**

```
hive> select count (distinct OwnerUserId)
from DATASET where LOWER(Body) like
LOWER('%cloud%') or LOWER(Title) like
LOWER('%cloud%');
```

Assumption B: Match word cloud solely on its own - including hyphenated compound words. E.g. - 'cloud-technology'. > Number of distinct users = **290**

```
hive> select count (distinct OwnerUserId)
from DATASET where LOWER(Body) like
LOWER('% cloud %') or LOWER(Body) like
LOWER('%cloud %') or LOWER(Body) like
LOWER(' %cloud.%') or LOWER(Body) like
LOWER(' %cloud-%') or LOWER(Body) like
LOWER(' %cloud*%') or LOWER(Title) like
LOWER('% cloud %') or LOWER(Title) like
LOWER('%cloud %') or LOWER(Title) like
LOWER('% cloud.%') or LOWER(Title) like
LOWER('% cloud-%') or LOWER(Title) like
LOWER('% cloud*%');
```

TASK 4 - PER-USER TF-IDF

CALCULATE TF-IDF OF THE TOP 10 TERMS FOR THE TOP 10 USERS

The TF-IDF is essentially a metric used to see how common or important a word is in a document, where in this case we are looking at the words used by the top 10 users identified in Task 3.2 previously.

Like in Task 2, the data was loaded and cleaned in Pig before filtering on the 10 top users identified by their respective ID, in order to create the new dataset for this task. In order to calculate the TF-IDF for words in both the 'Body' and 'Title' column, the CONCAT () function was used to create the column 'Body_Title'.

```
grunt> StackExchangeData_Total_Task4 =
FILTER StackExchangeData_Total_Task4 BY
(OwnerUserId == 87234 or OwnerUserId ==
4883 or OwnerUserId == 9951 or OwnerUserId
== 6068 or OwnerUserId == 6068 or
OwnerUserId == 89904 or OwnerUserId ==
51816 or OwnerUserId == 179736 or
OwnerUserId == 95592 or OwnerUserId ==
63051 or OwnerUserId == 4915);
```

The calculation of the TF-IDF was conducted in three phases using three mapper files and three reducer files, with a further mapper to giving the output results in the final phase of the implementation. These files were sourced from [10] which were modified from [12]. Changes were made by updating the 'stop words' with the ones in [15]. Due to the consistent logic in the code, no other changes were made to the mappers and reducers.

Using the Hadoop JAR command, we were able to bundle the programs together and run the code in the following generic format:

```
$ hadoop jar /usr/lib/hadoop/hadoop-
streaming.jar -file /home/colin_hehir3
/mapperX.py /home/colin_hehir3/reducerX.py
-mapper "python mapperX.py" -reducer
"reducer.py" -input /user/colin_hehir3/
StackExchangeData_Total_Task4_Output/part-
m-*/ -output /user/colin_hehir3/ch_outputX
```

The results of the TF-IDF calculation were then loaded into a Hive table from the mapper/reducer files created.

```
hive> load data inpath
'/user/colin_hehir3/ch_final_output/part-
0000*' into table
CH_StackExchangeData_Total_Task4;
```

Since we are looking for the top 10 TF-IDF records per user, our Hive query essentially ranked each word from 1-10 in descending order by their respective TF-IDF value and grouped them by the user who posted it.

For each row in the dataset, RANK() was used to form a position of hierarchy for each user's TF-IDF values. The PARTITION BY clause was used to define the subset of data in a partition, which in this case was the user's ID and then ORDER BY to rank the highest value of results in descending order.

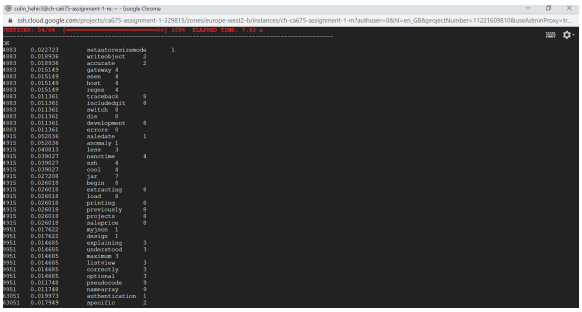
DISTRIBUTE BY and SORT BY essentially clustered the dataset by the user, on OwnerUserId.

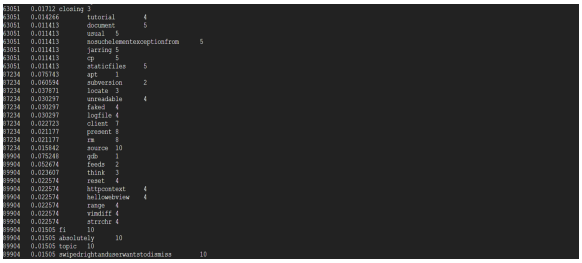
The WHERE clause was (<=) to include rank #10 also. Due to the fact some words may have the same TF-IDF value, they may also have had the same rank number – resulting in the query giving more than 10 words per user in this scenario.

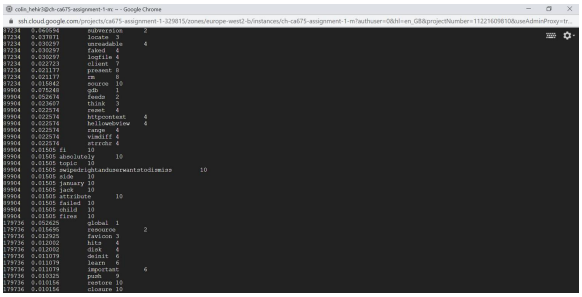
The ORDER BY function was used on the user's ID first in order to group each user together, followed by the rank to output the words from 1-10 based on descending value.

```
hive> SELECT * FROM (SELECT
OwnerUserId,tfidf, word, rank() over
(PARTITION BY userId ORDER BY tfidf DESC)
as tfidf rank FROM ch_tfidf_task4
DISTRIBUTE BY userId SORT BY OwnerUserId
desc) a WHERE rank <= 10 ORDER BY
OwnerUserId, tfidfrank;
```

The result of the query is as follows:







Figures 7, 8 and 9: TF-IDF of the top 10 terms for the top 10 users.

Where the TF-IDF value was calculated via the formula:

$$TF - IDF = \frac{n}{N} \times \log\left(\frac{D}{m}\right)$$

Where n *= # times word is in a document and* $N = \sum n$.

CONCLUSION

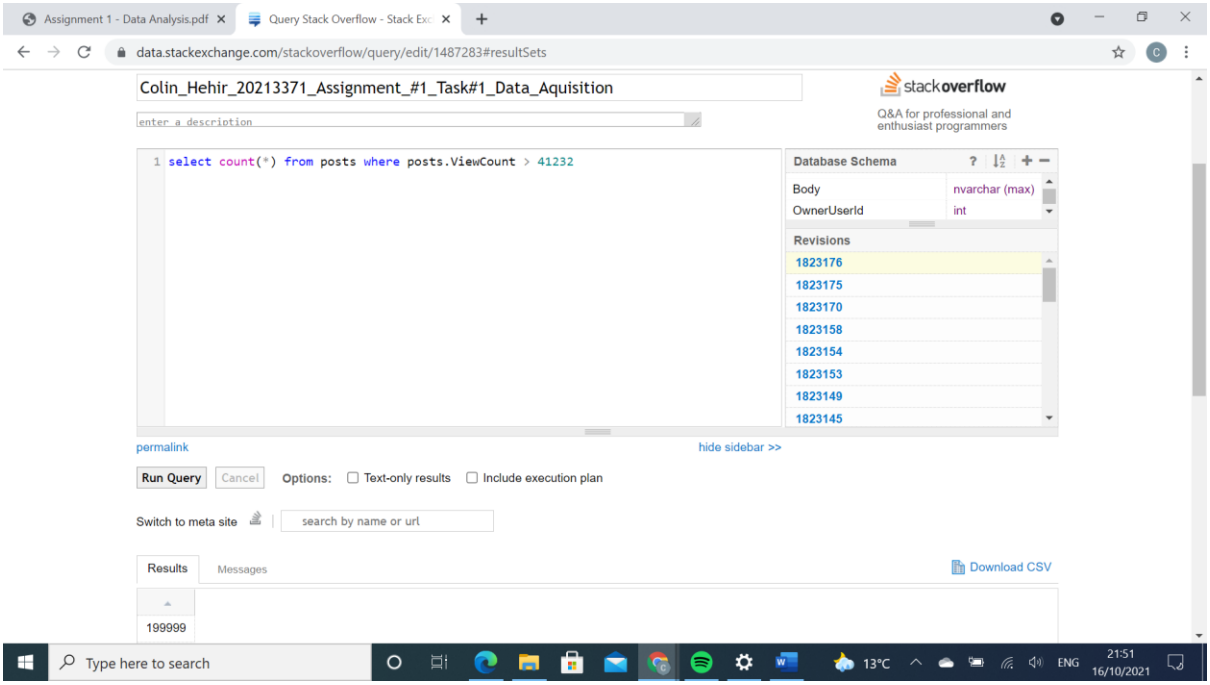
Over the course of this assignment, I acquired and collected data from Stack Exchange Data Explorer before performing ETL and data cleaning/processing in Apache Pig. Queries were run in Apache Hive in order to conduct data analysis, while MapReduce algorithms were used from Python scripts in Apache Hadoop. This was all implemented in a cluster created in Google Cloud Platform (GCP) via DataProc.

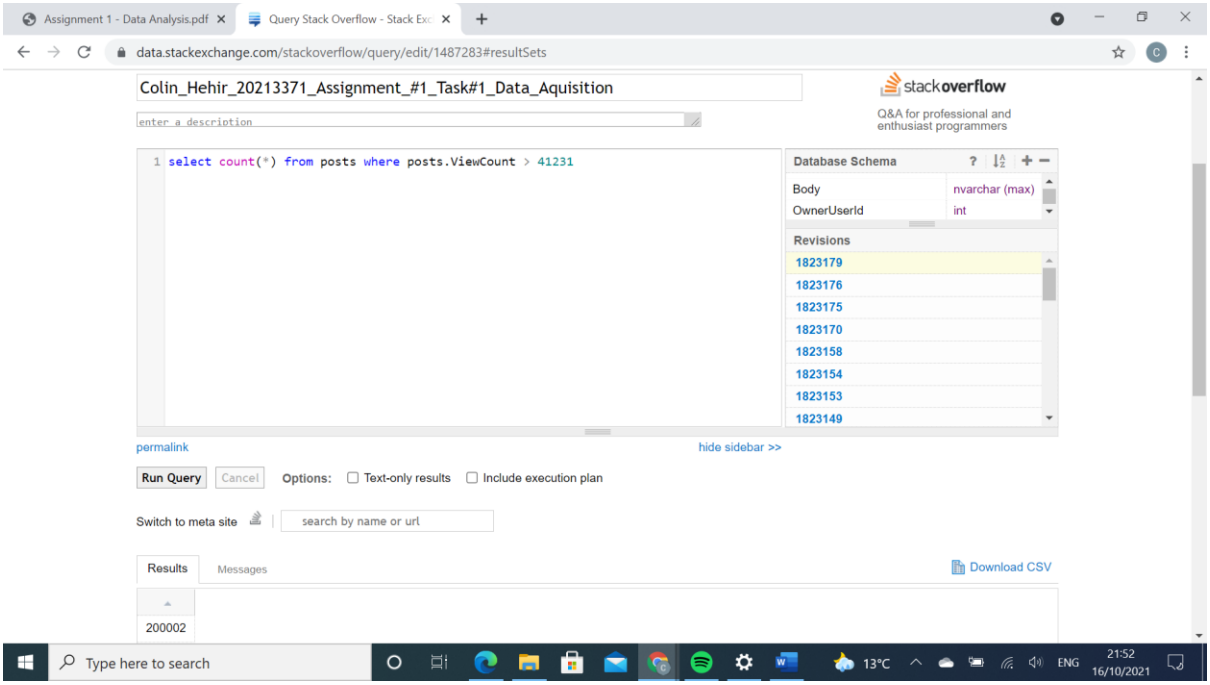
REFERENCES

1) Kesavulu, M (2021). *LOOP DCU: CA675 Cloud Technologies.* [online] Available at: https://loop.dcu.ie/course/view.php?id=56668 (Accessed Multiple Times).
2) IntelligentHQ (2021). *Effective Use of Cloud Computing in Education.* [image] https://www.intelligenthq.com/effective-use-cloud-computing-education/ (Accessed: 25 October 2021).
3) Stack Exchange (2021). *Stack Exchange Data Explorer* [online] Available at: https://data.stackexchange.com/stackoverflow/query/new (Accessed: 16 October 2021).
4) Kumar, M. (2016). Google cloud platform: a powerful big data analytics cloud platform. *Int J Res Appl Sci Eng Technol*, *4*(11), 387-392.
5) Nandimath, J., Banerjee, E., Patil, A., Kakade, P., Vaidya, S., & Chaturvedi, D. (2013, August). Big data analysis using Apache Hadoop. In *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)* (pp. 700-703). IEEE.
6) Fuad, A., Erwin, A., & Ipung, H. P. (2014, September). Processing performance on apache pig, apache hive and MySQL cluster. In *Proceedings of International Conference on Information, Communication Technology and System (ICTS) 2014* (pp. 297-302). IEEE.
7) Kumar, R., Gupta, N., Charu, S., Bansal, S., & Yadav, K. (2014). Comparison of SQL with HiveQL. *International Journal for Research in Technological Studies*, *1*(9), 2348-1439.
8) Hive, A. (2013). Apache Hive.
9) Dittrich, J., & Quiané-Ruiz, J. A. (2012). Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment*, *5*(12), 2014-2015.
10) Vijay Khair, A. (2020). *CA675_Cloud_Technologies_Assignment_1* [online] Available at: https://gitlab.computing.dcu.ie/khaira2/ca675_cloud_technologies_assignment_1 (Accessed: 24 October 2021).
11) Unni, S. (2020). *CA675-Assignment-1-20211114* [online] Available at: https://github.com/srjth19/CA675-Assignment-1-20211114 (Accessed: 24 October 2021).
12) Patel, D. (2017). *TF-IDF-implementation-using-map-reduce-Hadoop-python-* [online] Available at: https://github.com/devangpatel01/TF-IDF-implementation-using-map-reduce-Hadoop-python- (Accessed: 24 October 2021).
13) Higgins, D. (2020). *CA675 Cloud Technologies* Available at: https://gitlab.computing.dcu.ie/higgid23/ca675-cloud-technologies (Accessed: 24 October 2021).
14) Bellgutte Ramesh, A. (2019). *PigHiveOnStackExchangeData* Available at: https://github.com/arunabellgutteramesh/PigHiveOnStackExchangeData (Accessed: 24 October 2021).
15) RANKS NL. (2021). *English Stopwords* Available at: https://www.ranks.nl/stopwords (Accessed: 24 October 2021).

APPENDIX A: DATA ACQUISITION FROM STACK EXCHANGE

Please refer to the 'Data Acquisition from Stack Exchange Queries.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis





Assignment 1 - Data Analysis.pdf | Query Stack Overflow - Stack Exc

data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets

Colin_Hehir_20213371_Assignment_#1_Task#1_Data_Aquisition

enter a description

stackoverflow
Q&A for professional and enthusiast programmers

```
1 select top 50000* from posts where posts.ViewCount > 41231 order by posts.ViewCount DESC
```

permalink << show sidebar

Run Query | Cancel | Options: Text-only results | Include execution plan

Switch to meta site | search by name or url

Results | Messages | Download CSV

Id	PostTypeId	AcceptedAnswerId	Pare...	CreationDate	DeletionDate	Score	ViewCount	Body
927358	1	927386		2009-05-29 18:09:14		23274	10001585	I accidentally committed the wro…

Type here to search 13°C ENG 22:05 16/10/2021

Assignment 1 - Data Analysis.pdf | Query Stack Overflow - Stack Exc

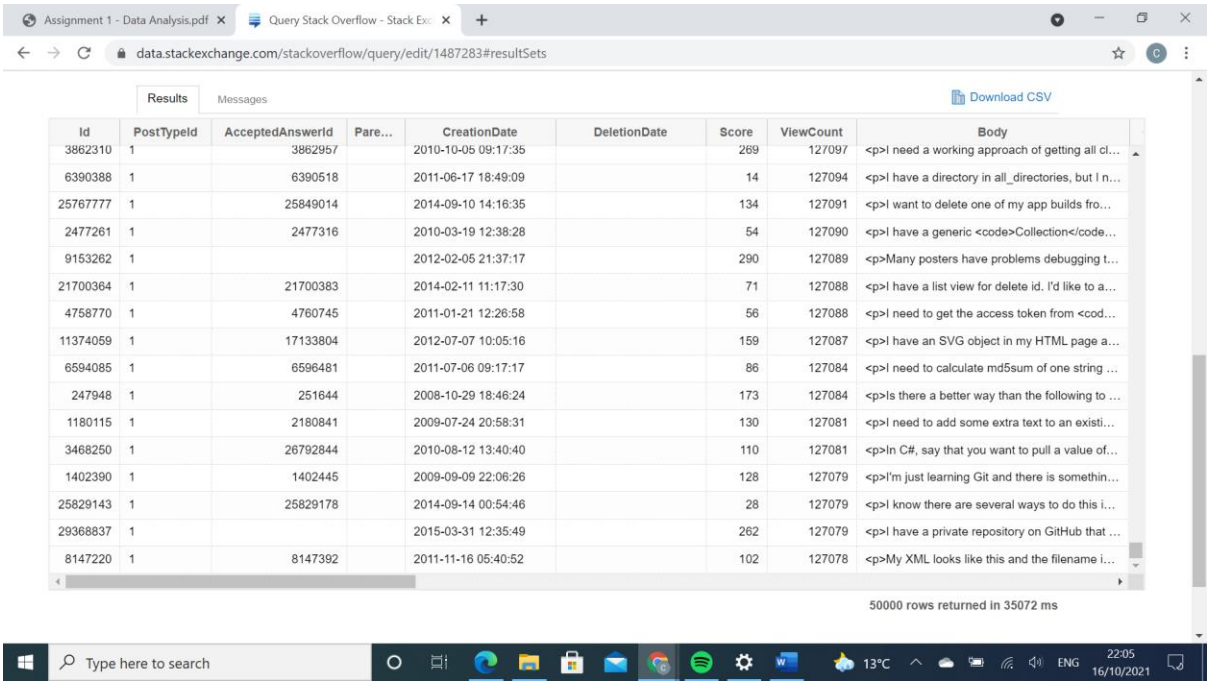
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets

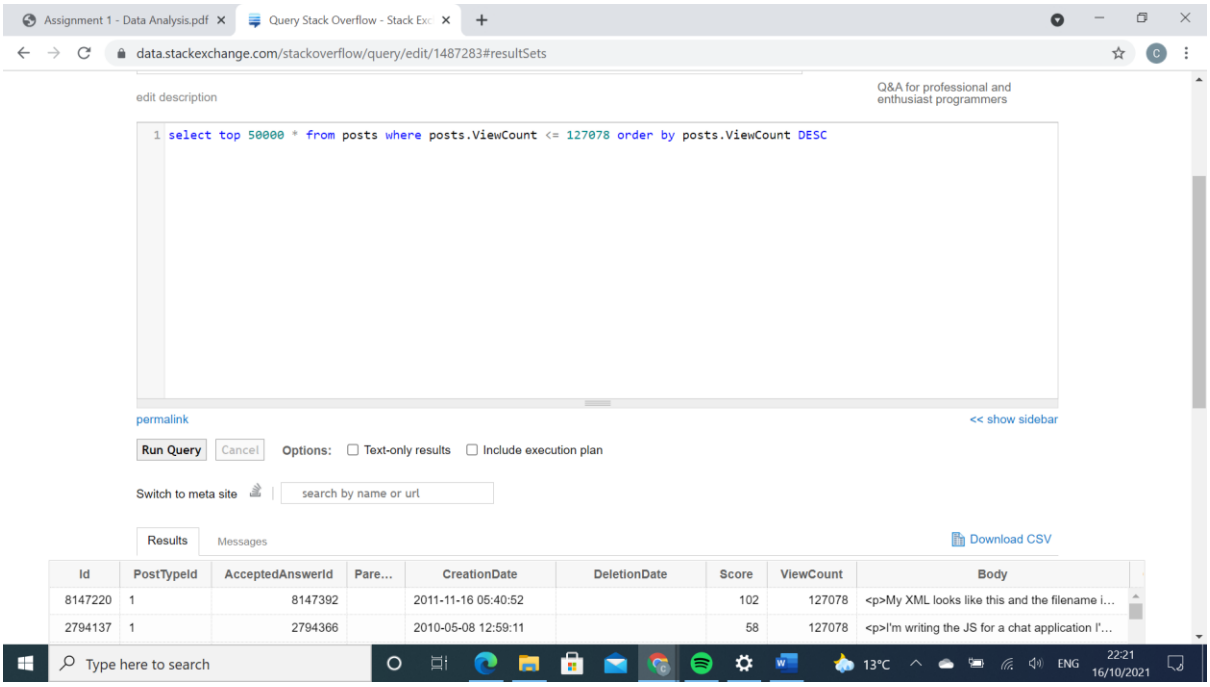
Results | Messages | Download CSV

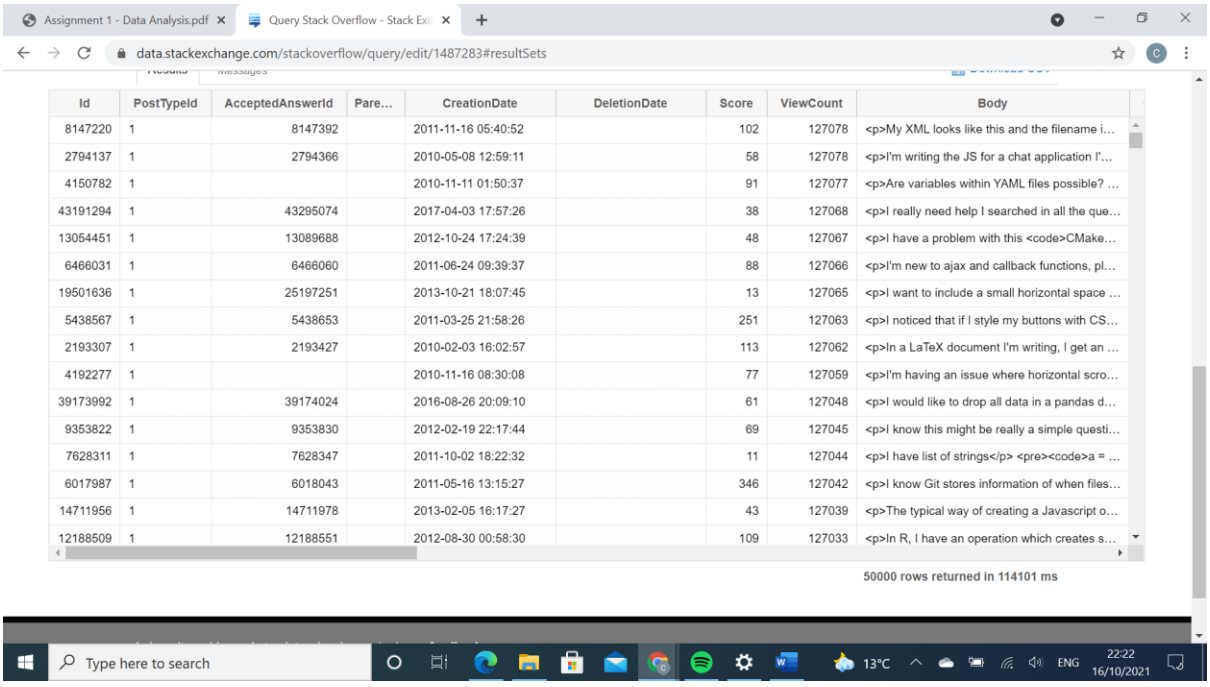
Id	PostTypeId	AcceptedAnswerId	Pare...	CreationDate	DeletionDate	Score	ViewCount	Body
927358	1	927386		2009-05-29 18:09:14		23274	10001585	I accidentally committed the wro…
2003505	1	2003515		2010-01-05 01:12:15		18451	9233751	I want to delete a branch both locally and …
5767325	1	5767357		2011-04-23 22:17:18		9877	8860159	I have an array of numbers and I'm using …
16956810	1	16957078		2013-06-06 08:06:45		6270	8177226	I'm trying to find a way to scan my entire …
2906582	1	2906586		2010-05-25 16:39:47		2294	7760798	I would like to create an HTML button that…
4114095	1	4114122		2010-11-06 16:58:14		7621	7565860	How do I revert from my current state to a…
503093	1	506004		2009-02-02 12:54:16		7710	6918255	How can I redirect the user from one pag…
1789945	1	1789952		2009-11-24 13:04:29		7418	6836325	Usually I would expect a <code>String.co…
5585779	1	5585800		2011-04-07 18:27:54		3275	6383401	How can I convert a <code>String</code…
1783405	1	1783426		2009-11-23 14:23:46		7521	6217855	Somebody pushed a branch called <code…
1125968	1	8888015		2009-07-14 14:58:15		8258	6073747	How do I force an overwrite of local files o…
3207219	1			2010-07-08 19:31:22		3466	5983823	How can I list all files of a directory in Pyt…
4366730	1	4366748		2010-12-06 13:14:05		2660	5828599	Consider: <pre><code>$a = 'How ar…
3437059	1	3437070		2010-08-09 02:52:50		3596	5254761	I'm looking for a <code>string.contains</c…
20035101	1	20035319		2013-11-17 19:29:06		2877	5102090	<blockquote> Mod note</strong…
1200621	1	1200646		2009-07-29 14:22:27		2277	5013476	How do I declare and initialize an array in…

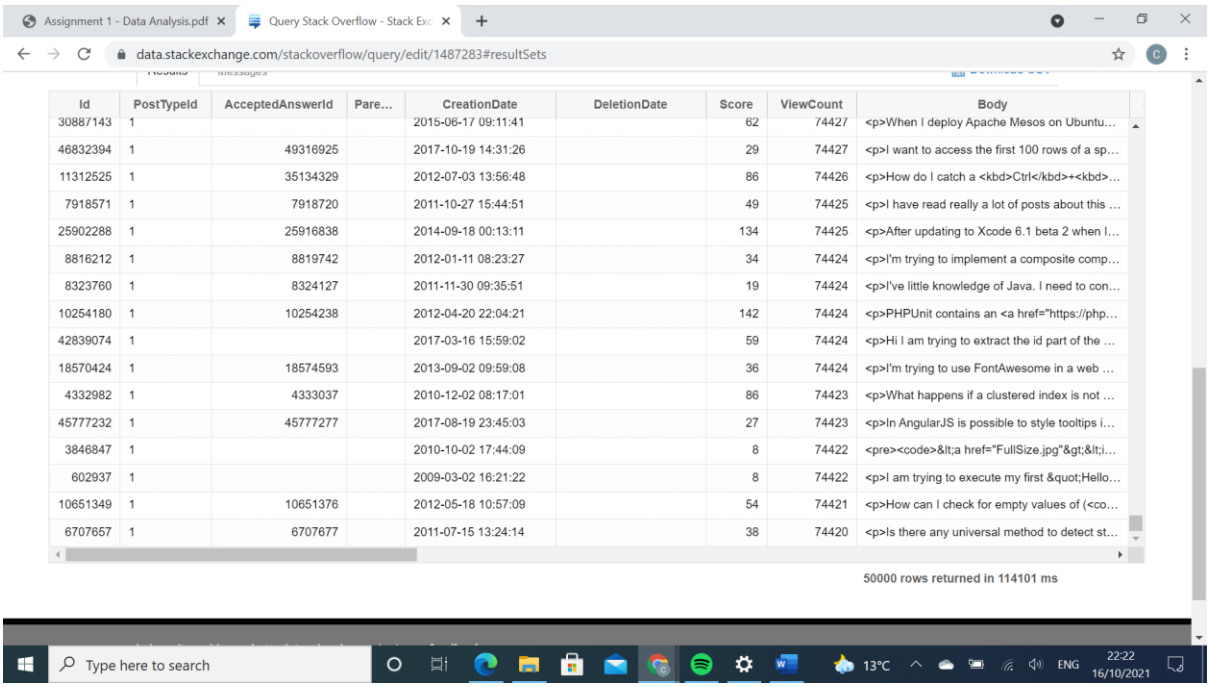
50000 rows returned in 35072 ms

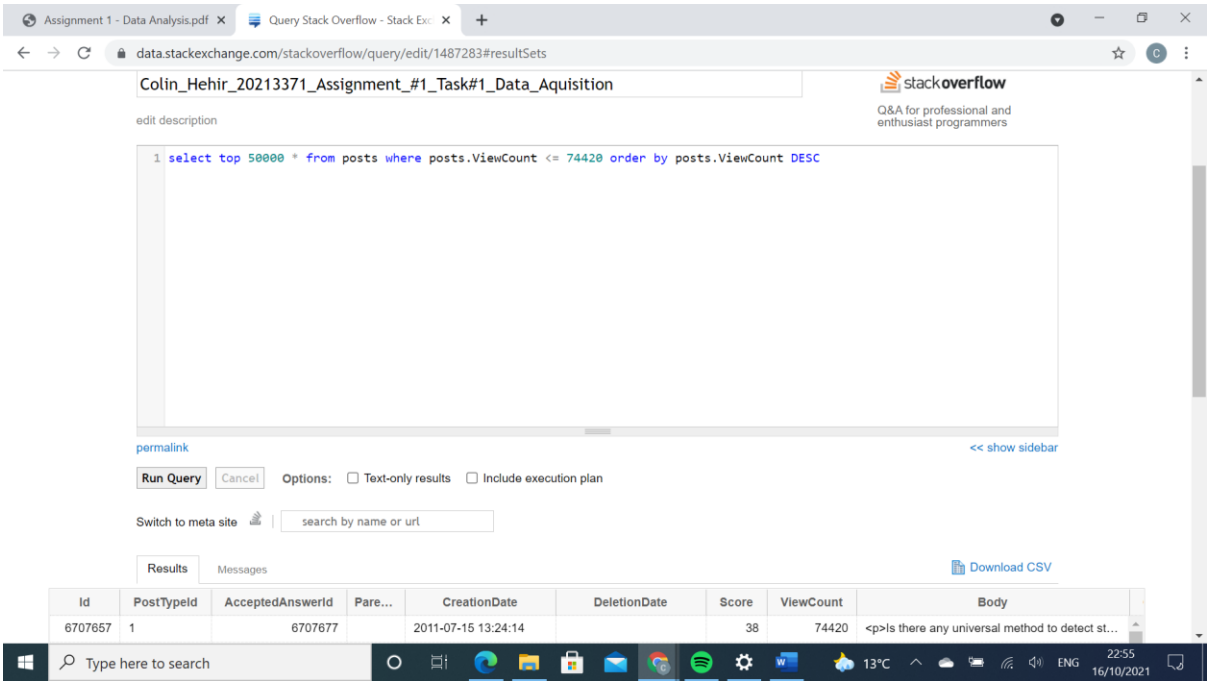
Type here to search 13°C ENG 22:05 16/10/2021

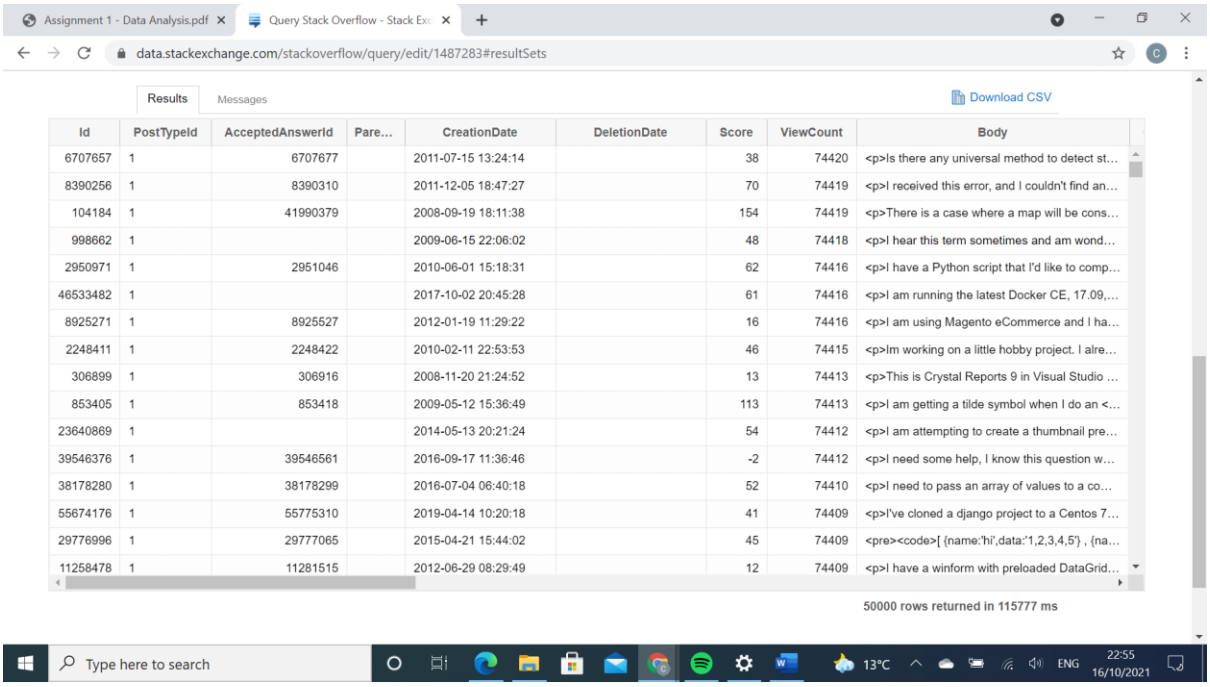

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Results
Messages
Download CSV
Id | PostTypeId | AcceptedAnswerId | Pare... | CreationDate | DeletionDate | Score | ViewCount | Body
3862310 | 1 | 3862957 | | 2010-10-05 09:17:35 | | 269 | 127097 | <p>I need a working approach of getting all cl...
6390388 | 1 | 6390518 | | 2011-06-17 18:49:09 | | 14 | 127094 | <p>I have a directory in all_directories, but I n...
25767777 | 1 | 25849014 | | 2014-09-10 14:16:35 | | 134 | 127091 | <p>I want to delete one of my app builds fro...
2477261 | 1 | 2477316 | | 2010-03-19 12:38:28 | | 54 | 127090 | <p>I have a generic <code>Collection</code>...
9153262 | 1 | | | 2012-02-05 21:37:17 | | 290 | 127089 | <p>Many posters have problems debugging t...
21700364 | 1 | 21700383 | | 2014-02-11 11:17:30 | | 71 | 127088 | <p>I have a list view for delete id. I'd like to a...
4758770 | 1 | 4760745 | | 2011-01-21 12:26:58 | | 56 | 127088 | <p>I need to get the access token from <cod...
11374059 | 1 | 17133804 | | 2012-07-07 10:05:16 | | 159 | 127087 | <p>I have an SVG object in my HTML page a...
6594085 | 1 | 6596481 | | 2011-07-06 09:17:17 | | 86 | 127084 | <p>I need to calculate md5sum of one string ...
247948 | 1 | 251644 | | 2008-10-29 18:46:24 | | 173 | 127084 | <p>Is there a better way than the following to ...
1180115 | 1 | 2180841 | | 2009-07-24 20:58:31 | | 130 | 127081 | <p>I need to add some extra text to an existi...
3468250 | 1 | 26792844 | | 2010-08-12 13:40:40 | | 110 | 127081 | <p>In C#, say that you want to pull a value of...
1402390 | 1 | 1402445 | | 2009-09-09 22:06:26 | | 128 | 127079 | <p>I'm just learning Git and there is somethin...
25829143 | 1 | 25829178 | | 2014-09-14 00:54:46 | | 28 | 127079 | <p>I know there are several ways to do this i...
29368837 | 1 | | | 2015-03-31 12:35:49 | | 262 | 127079 | <p>I have a private repository on GitHub that ...
8147220 | 1 | 8147392 | | 2011-11-16 05:40:52 | | 102 | 127078 | <p>My XML looks like this and the filename i...
50000 rows returned in 35072 ms
Type here to search
13°C
ENG
22:05
16/10/2021

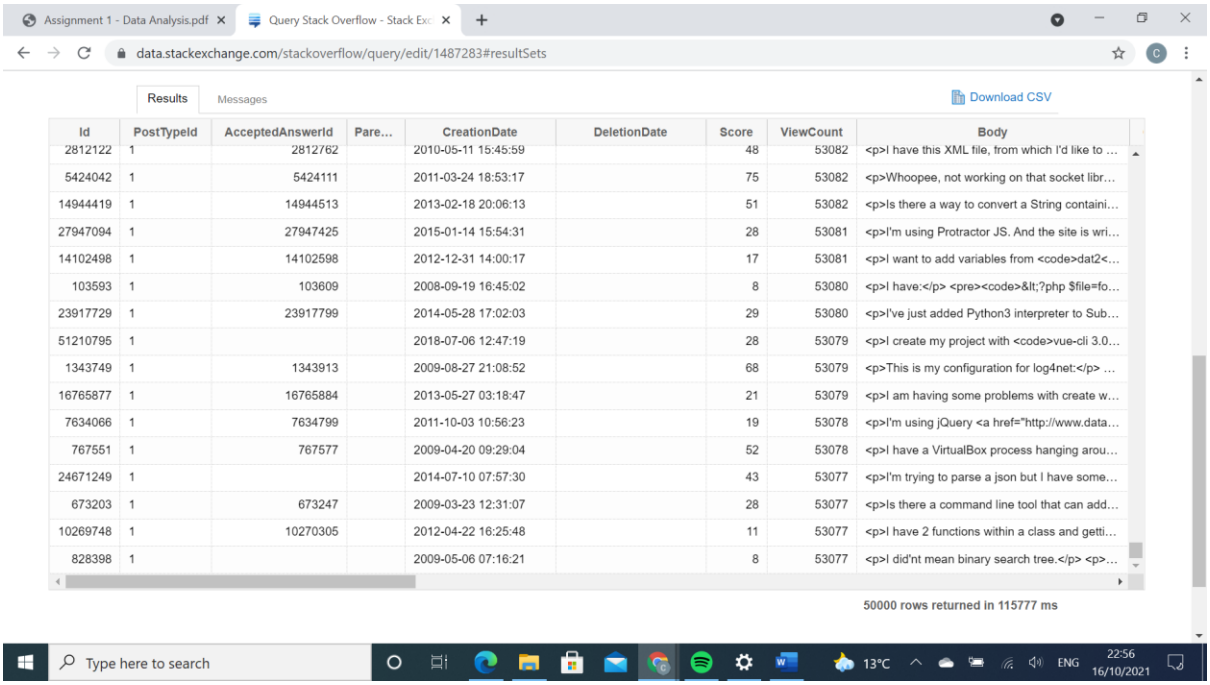

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
edit description
Q&A for professional and enthusiast programmers
1 select top 50000 * from posts where posts.ViewCount <= 127078 order by posts.ViewCount DESC
permalink
<< show sidebar
Run Query
Cancel
Options:
Text-only results
Include execution plan
Switch to meta site
search by name or url
Results
Messages
Download CSV
Id | PostTypeId | AcceptedAnswerId | Pare... | CreationDate | DeletionDate | Score | ViewCount | Body
8147220 | 1 | 8147392 | | 2011-11-16 05:40:52 | | 102 | 127078 | <p>My XML looks like this and the filename i...
2794137 | 1 | 2794366 | | 2010-05-08 12:59:11 | | 58 | 127078 | <p>I'm writing the JS for a chat application I'...
Type here to search
13°C
ENG
22:21
16/10/2021

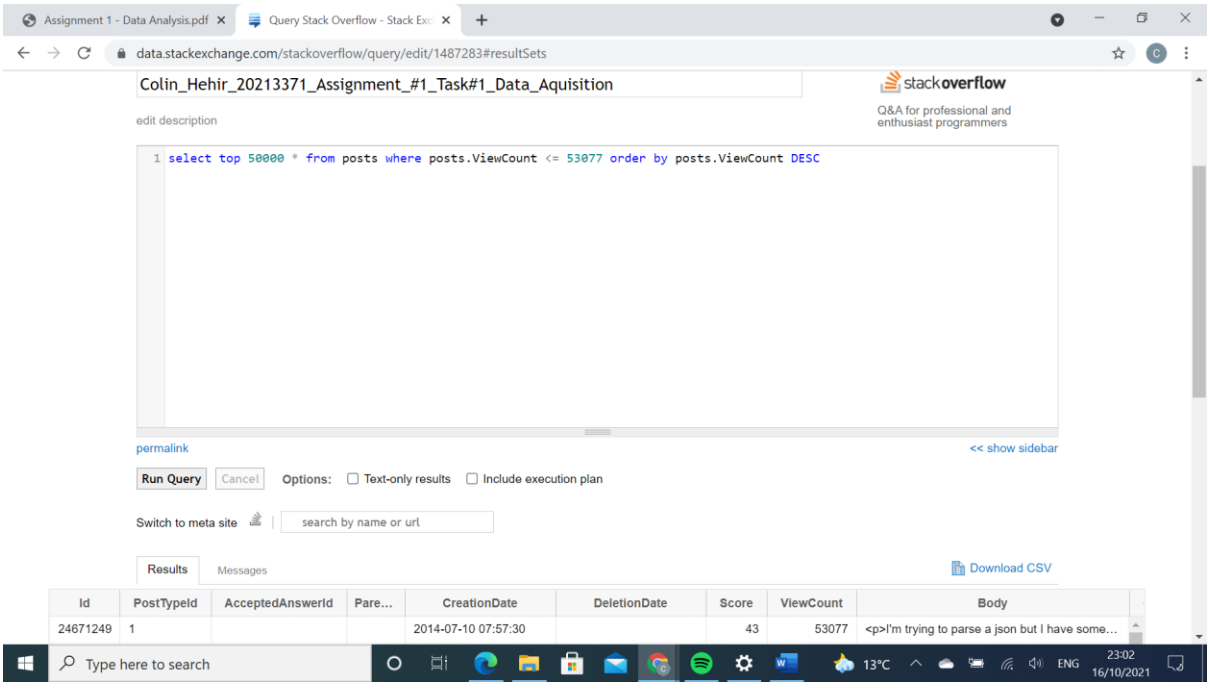

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Id | PostTypeId | AcceptedAnswerId | Pare... | CreationDate | DeletionDate | Score | ViewCount | Body
8147220 | 1 | 8147392 | | 2011-11-16 05:40:52 | | 102 | 127078 | My XML looks like this and the filename i...
2794137 | 1 | 2794366 | | 2010-05-08 12:59:11 | | 58 | 127078 | I'm writing the JS for a chat application I'...
4150782 | 1 | | | 2010-11-11 01:50:37 | | 91 | 127077 | Are variables within YAML files possible? ...
43191294 | 1 | 43295074 | | 2017-04-03 17:57:26 | | 38 | 127068 | I really need help I searched in all the que...
13054451 | 1 | 13089688 | | 2012-10-24 17:24:39 | | 48 | 127067 | I have a problem with this <code>CMake...
6466031 | 1 | 6466060 | | 2011-06-24 09:39:37 | | 88 | 127066 | I'm new to ajax and callback functions, pl...
19501636 | 1 | 25197251 | | 2013-10-21 18:07:45 | | 13 | 127065 | I want to include a small horizontal space ...
5438567 | 1 | 5438653 | | 2011-03-25 21:58:26 | | 251 | 127063 | I noticed that if I style my buttons with CS...
2193307 | 1 | 2193427 | | 2010-02-03 16:02:57 | | 113 | 127062 | In a LaTeX document I'm writing, I get an ...
4192277 | 1 | | | 2010-11-16 08:30:08 | | 77 | 127059 | I'm having an issue where horizontal scro...
39173992 | 1 | 39174024 | | 2016-08-26 20:09:10 | | 61 | 127048 | I would like to drop all data in a pandas d...
9353822 | 1 | 9353830 | | 2012-02-19 22:17:44 | | 69 | 127045 | I know this might be really a simple questi...
7628311 | 1 | 7628347 | | 2011-10-02 18:22:32 | | 11 | 127044 | I have list of strings <pre><code>a = ...
6017987 | 1 | 6018043 | | 2011-05-16 13:15:27 | | 346 | 127042 | I know Git stores information of when files...
14711956 | 1 | 14711978 | | 2013-02-05 16:17:27 | | 43 | 127039 | The typical way of creating a Javascript o...
12188509 | 1 | 12188551 | | 2012-08-30 00:58:30 | | 109 | 127033 | In R, I have an operation which creates s...
50000 rows returned in 114101 ms
Type here to search
13°C
ENG
22:22
16/10/2021

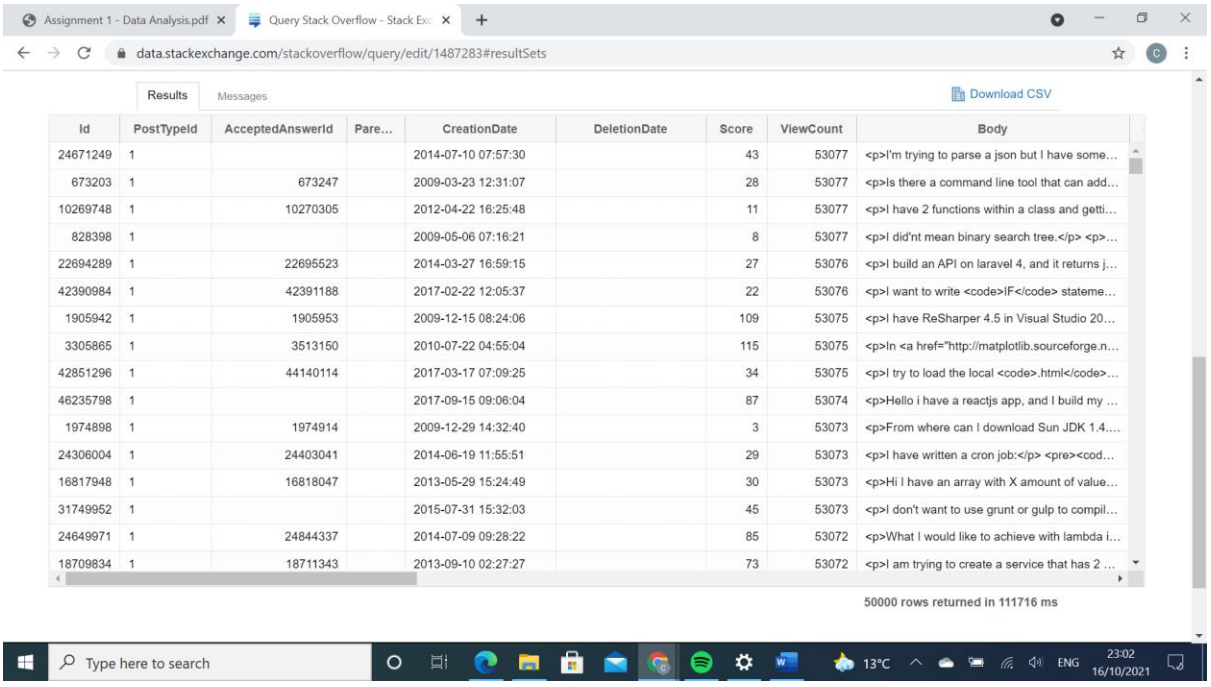

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Id | PostTypeId | AcceptedAnswerId | Pare... | CreationDate | DeletionDate | Score | ViewCount | Body
30887143 | 1 | | | 2015-06-17 09:11:41 | | 62 | 74427 | When I deploy Apache Mesos on Ubuntu...
46832394 | 1 | 49316925 | | 2017-10-19 14:31:26 | | 29 | 74427 | I want to access the first 100 rows of a sp...
11312525 | 1 | 35134329 | | 2012-07-03 13:56:48 | | 86 | 74426 | How do I catch a <kbd>Ctrl</kbd>+<kbd>...
7918571 | 1 | 7918720 | | 2011-10-27 15:44:51 | | 49 | 74425 | I have read really a lot of posts about this ...
25902288 | 1 | 25916838 | | 2014-09-18 00:13:11 | | 134 | 74425 | After updating to Xcode 6.1 beta 2 when I...
8816212 | 1 | 8819742 | | 2012-01-11 08:23:27 | | 34 | 74424 | I'm trying to implement a composite comp...
8323760 | 1 | 8324127 | | 2011-11-30 09:35:51 | | 19 | 74424 | I've little knowledge of Java. I need to con...
10254180 | 1 | 10254238 | | 2012-04-20 22:04:21 | | 142 | 74424 | PHPUnit contains an <a href="https://php...
42839074 | 1 | | | 2017-03-16 15:59:02 | | 59 | 74424 | Hi I am trying to extract the id part of the ...
18570424 | 1 | 18574593 | | 2013-09-02 09:59:08 | | 36 | 74424 | I'm trying to use FontAwesome in a web ...
4332982 | 1 | 4333037 | | 2010-12-02 08:17:01 | | 86 | 74423 | What happens if a clustered index is not ...
45777232 | 1 | 45777277 | | 2017-08-19 23:45:03 | | 27 | 74423 | In AngularJS is possible to style tooltips i...
3846847 | 1 | | | 2010-10-02 17:44:09 | | 8 | 74422 | <pre><code><a href="FullSize.jpg"><i...
602937 | 1 | | | 2009-03-02 16:21:22 | | 8 | 74422 | I am trying to execute my first "Hello...
10651349 | 1 | 10651376 | | 2012-05-18 10:57:09 | | 54 | 74421 | How can I check for empty values of (<co...
6707657 | 1 | 6707677 | | 2011-07-15 13:24:14 | | 38 | 74420 | Is there any universal method to detect st...
50000 rows returned in 114101 ms
Type here to search
13°C
ENG
22:22
16/10/2021

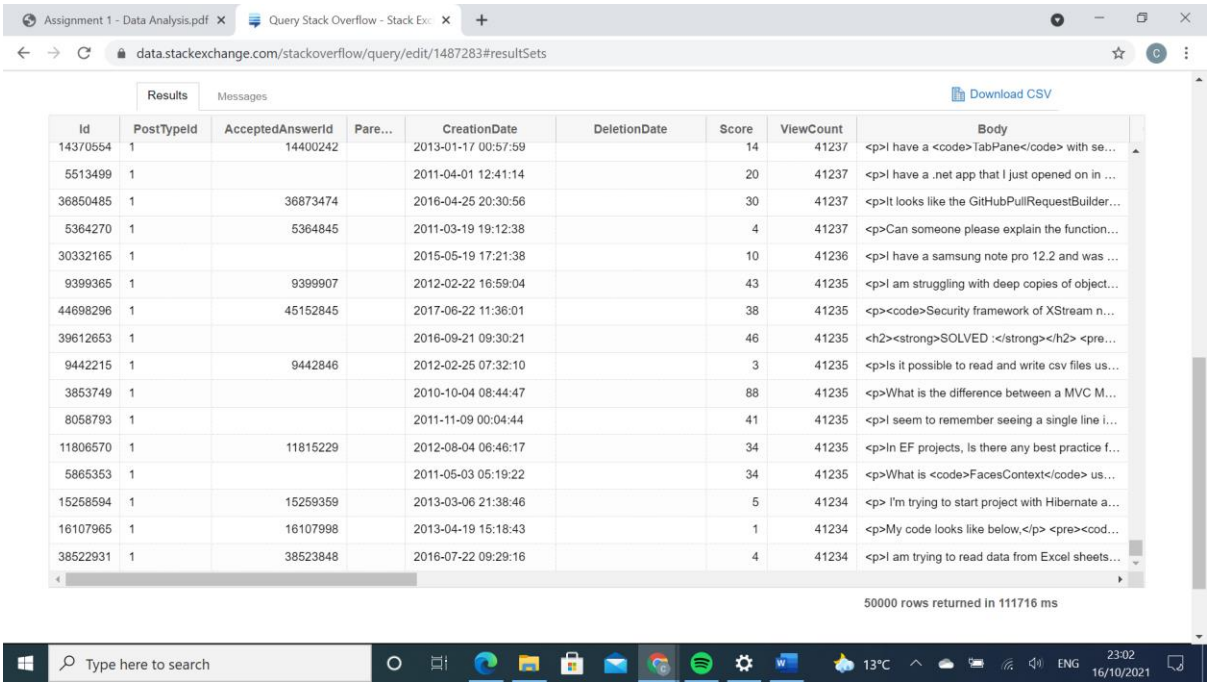

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Colin_Hehir_20213371_Assignment_#1_Task#1_Data_Aquisition
edit description
stackoverflow
Q&A for professional and enthusiast programmers
1 select top 50000 * from posts where posts.ViewCount <= 74420 order by posts.ViewCount DESC
permalink
<< show sidebar
Run Query
Cancel
Options:
Text-only results
Include execution plan
Switch to meta site
search by name or url
Results
Messages
Download CSV
Id PostTypeId AcceptedAnswerId Pare... CreationDate DeletionDate Score ViewCount Body
6707657 1 6707677 2011-07-15 13:24:14 38 74420 <p>Is there any universal method to detect st...
Type here to search
13°C
ENG
22:55
16/10/2021

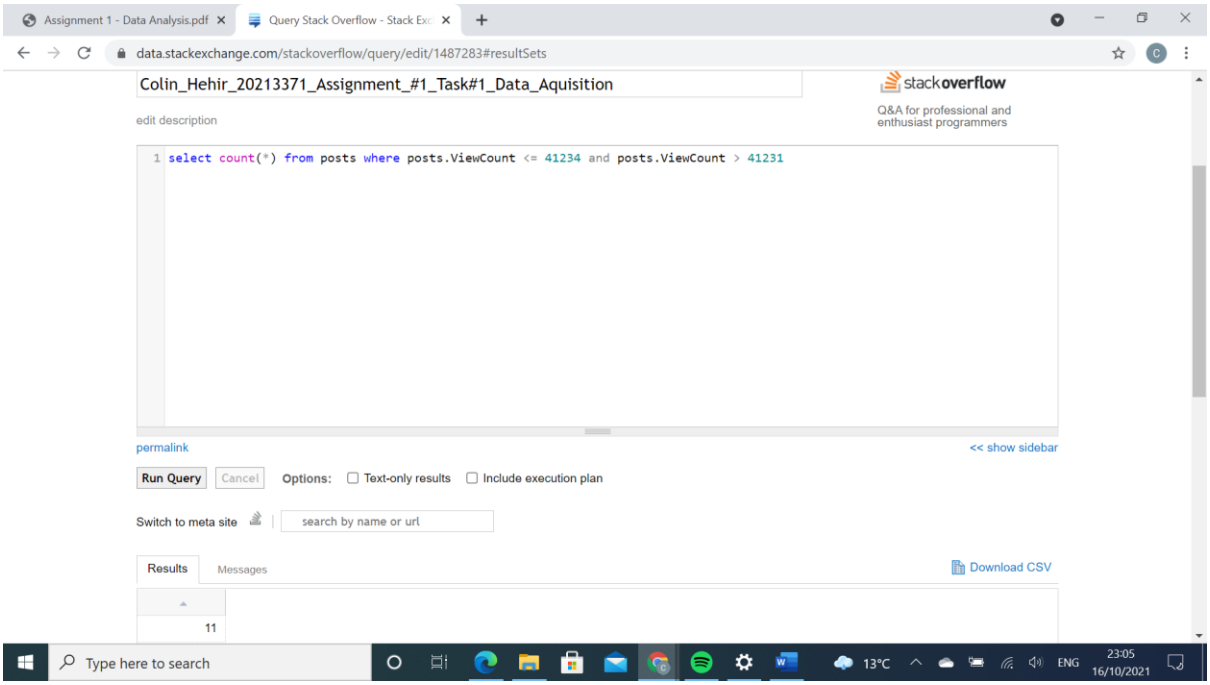

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Results
Messages
Download CSV
Id PostTypeId AcceptedAnswerId Pare... CreationDate DeletionDate Score ViewCount Body
6707657 1 6707677 2011-07-15 13:24:14 38 74420 <p>Is there any universal method to detect st...
8390256 1 8390310 2011-12-05 18:47:27 70 74419 <p>I received this error, and I couldn't find an...
104184 1 41990379 2008-09-19 18:11:38 154 74419 <p>There is a case where a map will be cons...
998662 1 2009-06-15 22:06:02 48 74418 <p>I hear this term sometimes and am wond...
2950971 1 2951046 2010-06-01 15:18:31 62 74416 <p>I have a Python script that I'd like to comp...
46533482 1 2017-10-02 20:45:28 61 74416 <p>I am running the latest Docker CE, 17.09,...
8925271 1 8925527 2012-01-19 11:29:22 16 74416 <p>I am using Magento eCommerce and I ha...
2248411 1 2248422 2010-02-11 22:53:53 46 74415 <p>Im working on a little hobby project. I alre...
306899 1 306916 2008-11-20 21:24:52 13 74413 <p>This is Crystal Reports 9 in Visual Studio ...
853405 1 853418 2009-05-12 15:36:49 113 74413 <p>I am getting a tilde symbol when I do an <...
23640869 1 2014-05-13 20:21:24 54 74412 <p>I am attempting to create a thumbnail pre...
39546376 1 39546561 2016-09-17 11:36:46 -2 74412 <p>I need some help, I know this question w...
38178280 1 38178299 2016-07-04 06:40:18 52 74410 <p>I need to pass an array of values to a co...
55674176 1 55775310 2019-04-14 10:20:18 41 74409 <p>I've cloned a django project to a Centos 7...
29776996 1 29777065 2015-04-21 15:44:02 45 74409 <pre><code>[{name:'hi',data:'1,2,3,4,5'} , {na...
11258478 1 11281515 2012-06-29 08:29:49 12 74409 <p>I have a winform with preloaded DataGrid...
50000 rows returned in 115777 ms
Type here to search
13°C
ENG
22:55
16/10/2021

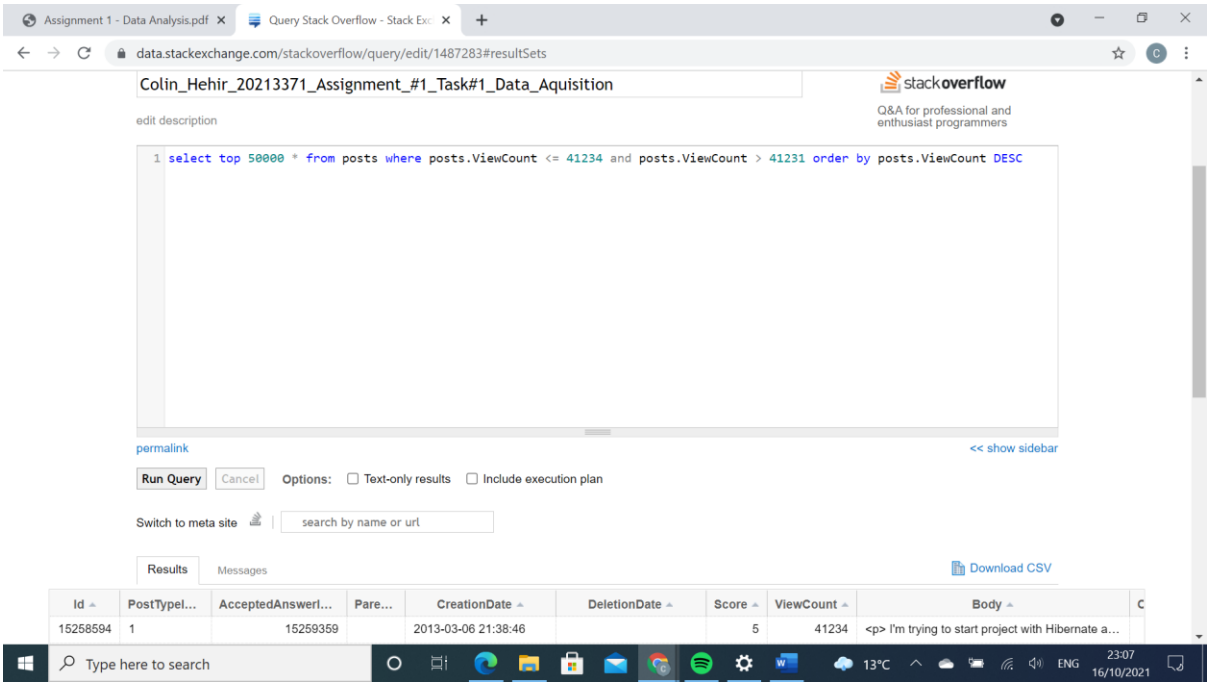

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Results
Messages
Download CSV
Id | PostTypeId | AcceptedAnswerId | Pare... | CreationDate | DeletionDate | Score | ViewCount | Body
2812122 | 1 | 2812762 | | 2010-05-11 15:45:59 | | 48 | 53082 | I have this XML file, from which I'd like to ...
5424042 | 1 | 5424111 | | 2011-03-24 18:53:17 | | 75 | 53082 | Whoopee, not working on that socket libr...
14944419 | 1 | 14944513 | | 2013-02-18 20:06:13 | | 51 | 53082 | Is there a way to convert a String containi...
27947094 | 1 | 27947425 | | 2015-01-14 15:54:31 | | 28 | 53081 | I'm using Protractor JS. And the site is wri...
14102498 | 1 | 14102598 | | 2012-12-31 14:00:17 | | 17 | 53081 | I want to add variables from <code>dat2<...
103593 | 1 | 103609 | | 2008-09-19 16:45:02 | | 8 | 53080 | I have: <pre><code><?php $file=fo...
23917729 | 1 | 23917799 | | 2014-05-28 17:02:03 | | 29 | 53080 | I've just added Python3 interpreter to Sub...
51210795 | 1 | | | 2018-07-06 12:47:19 | | 28 | 53079 | I create my project with <code>vue-cli 3.0...
1343749 | 1 | 1343913 | | 2009-08-27 21:08:52 | | 68 | 53079 | This is my configuration for log4net: ...
16765877 | 1 | 16765884 | | 2013-05-27 03:18:47 | | 21 | 53079 | I am having some problems with create w...
7634066 | 1 | 7634799 | | 2011-10-03 10:56:23 | | 19 | 53078 | I'm using jQuery <a href="http://www.data...
767551 | 1 | 767577 | | 2009-04-20 09:29:04 | | 52 | 53078 | I have a VirtualBox process hanging arou...
24671249 | 1 | | | 2014-07-10 07:57:30 | | 43 | 53077 | I'm trying to parse a json but I have some...
673203 | 1 | 673247 | | 2009-03-23 12:31:07 | | 28 | 53077 | Is there a command line tool that can add...
10269748 | 1 | 10270305 | | 2012-04-22 16:25:48 | | 11 | 53077 | I have 2 functions within a class and getti...
828398 | 1 | | | 2009-05-06 07:16:21 | | 8 | 53077 | I did'nt mean binary search tree. ...
50000 rows returned in 115777 ms
Type here to search
13°C
ENG
22:56
16/10/2021


Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Colin_Hehir_20213371_Assignment_#1_Task#1_Data_Aquisition
edit description
stackoverflow
Q&A for professional and enthusiast programmers
1 select top 50000 * from posts where posts.ViewCount <= 53077 order by posts.ViewCount DESC
permalink
<< show sidebar
Run Query
Cancel
Options:
Text-only results
Include execution plan
Switch to meta site
search by name or url
Results
Messages
Download CSV
Id | PostTypeId | AcceptedAnswerId | Pare... | CreationDate | DeletionDate | Score | ViewCount | Body
24671249 | 1 | | | 2014-07-10 07:57:30 | | 43 | 53077 | I'm trying to parse a json but I have some...
Type here to search
13°C
ENG
23:02
16/10/2021


Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Results
Messages
Download CSV
Id PostTypeId AcceptedAnswerId Pare... CreationDate DeletionDate Score ViewCount Body
24671249 1 2014-07-10 07:57:30 43 53077 I'm trying to parse a json but I have some...
673203 1 673247 2009-03-23 12:31:07 28 53077 Is there a command line tool that can add...
10269748 1 10270305 2012-04-22 16:25:48 11 53077 I have 2 functions within a class and getti...
828398 1 2009-05-06 07:16:21 8 53077 I did'nt mean binary search tree. ...
22694289 1 22695523 2014-03-27 16:59:15 27 53076 I build an API on laravel 4, and it returns j...
42390984 1 42391188 2017-02-22 12:05:37 22 53076 I want to write <code>IF</code> stateme...
1905942 1 1905953 2009-12-15 08:24:06 109 53075 I have ReSharper 4.5 in Visual Studio 20...
3305865 1 3513150 2010-07-22 04:55:04 115 53075 In <a href="http://matplotlib.sourceforge.n...
42851296 1 44140114 2017-03-17 07:09:25 34 53075 I try to load the local <code>.html</code>...
46235798 1 2017-09-15 09:06:04 87 53074 Hello i have a reactjs app, and I build my ...
1974898 1 1974914 2009-12-29 14:32:40 3 53073 From where can I download Sun JDK 1.4....
24306004 1 24403041 2014-06-19 11:55:51 29 53073 I have written a cron job: <pre><cod...
16817948 1 16818047 2013-05-29 15:24:49 30 53073 Hi I have an array with X amount of value...
31749952 1 2015-07-31 15:32:03 45 53073 I don't want to use grunt or gulp to compil...
24649971 1 24844337 2014-07-09 09:28:22 85 53072 What I would like to achieve with lambda i...
18709834 1 18711343 2013-09-10 02:27:27 73 53072 I am trying to create a service that has 2 ...
50000 rows returned in 111716 ms
Type here to search
13°C
ENG
23:02
16/10/2021

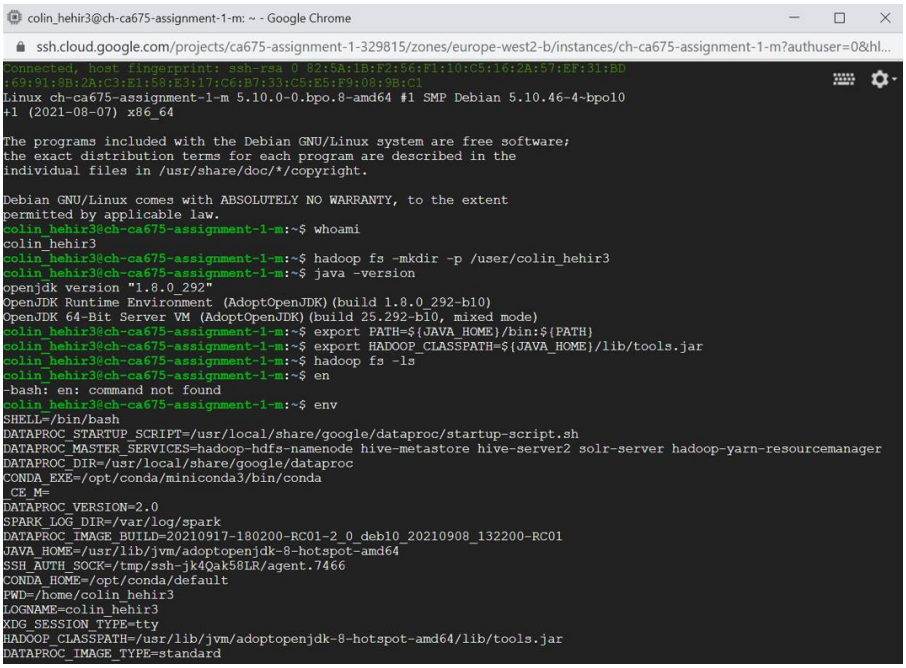

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Results
Messages
Download CSV
Id PostTypeId AcceptedAnswerId Pare... CreationDate DeletionDate Score ViewCount Body
14370554 1 14400242 2013-01-17 00:57:59 14 41237 I have a <code>TabPane</code> with se...
5513499 1 2011-04-01 12:41:14 20 41237 I have a .net app that I just opened on in ...
36850485 1 36873474 2016-04-25 20:30:56 30 41237 It looks like the GitHubPullRequestBuilder...
5364270 1 5364845 2011-03-19 19:12:38 4 41237 Can someone please explain the function...
30332165 1 2015-05-19 17:21:38 10 41236 I have a samsung note pro 12.2 and was ...
9399365 1 9399907 2012-02-22 16:59:04 43 41235 I am struggling with deep copies of object...
44698296 1 45152845 2017-06-22 11:36:01 38 41235 <code>Security framework of XStream n...
39612653 1 2016-09-21 09:30:21 46 41235 <h2>SOLVED :</h2> <pre...
9442215 1 9442846 2012-02-25 07:32:10 3 41235 Is it possible to read and write csv files us...
3853749 1 2010-10-04 08:44:47 88 41235 What is the difference between a MVC M...
8058793 1 2011-11-09 00:04:44 41 41235 I seem to remember seeing a single line i...
11806570 1 11815229 2012-08-04 06:46:17 34 41235 In EF projects, Is there any best practice f...
5865353 1 2011-05-03 05:19:22 34 41235 What is <code>FacesContext</code> us...
15258594 1 15259359 2013-03-06 21:38:46 5 41234 I'm trying to start project with Hibernate a...
16107965 1 16107998 2013-04-19 15:18:43 1 41234 My code looks like below, <pre><cod...
38522931 1 38523848 2016-07-22 09:29:16 4 41234 I am trying to read data from Excel sheets...
50000 rows returned in 111716 ms
Type here to search
13°C
ENG
23:02
16/10/2021

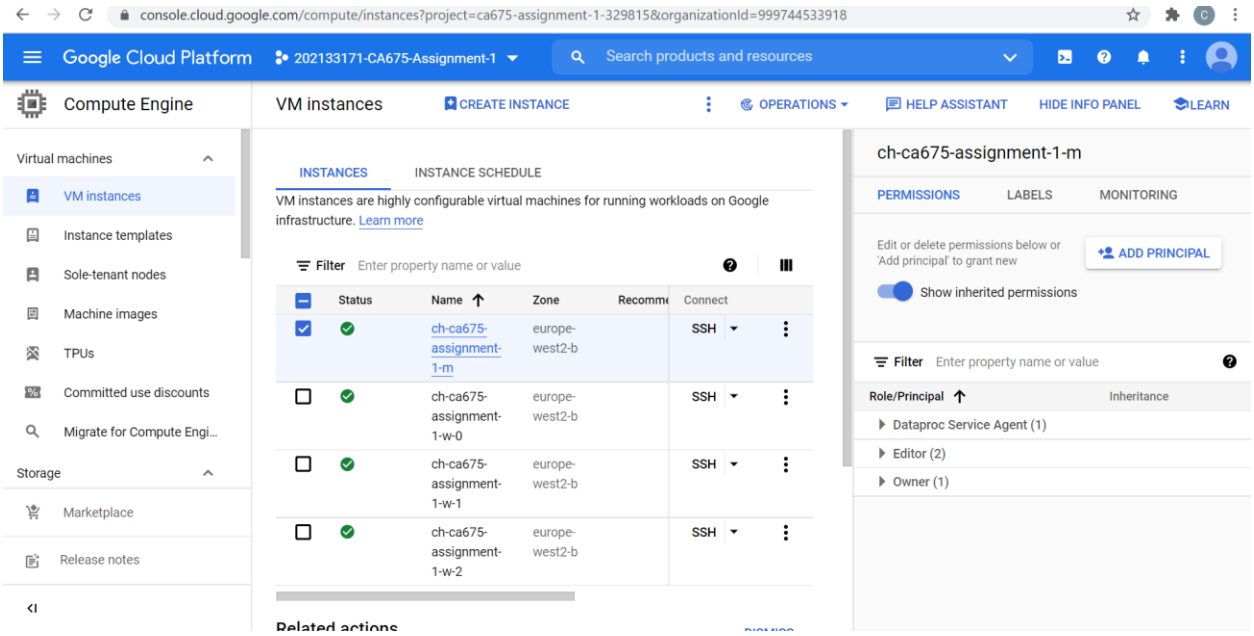

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Colin_Hehir_20213371_Assignment_#1_Task#1_Data_Aquisition
edit description
stackoverflow
Q&A for professional and enthusiast programmers
1 select count(*) from posts where posts.ViewCount <= 41234 and posts.ViewCount > 41231
permalink
<< show sidebar
Run Query
Cancel
Options:
Text-only results
Include execution plan
Switch to meta site
search by name or url
Results
Messages
Download CSV
11
Type here to search
13°C
ENG
23:05
16/10/2021

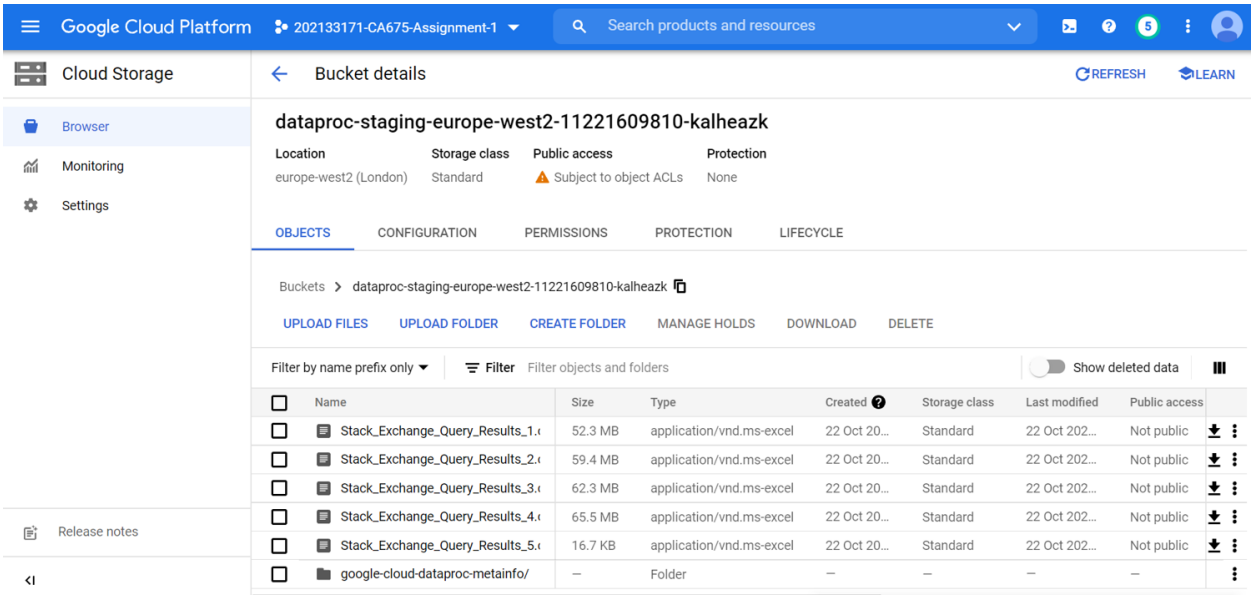

Assignment 1 - Data Analysis.pdf
Query Stack Overflow - Stack Exc
data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets
Colin_Hehir_20213371_Assignment_#1_Task#1_Data_Aquisition
edit description
stackoverflow
Q&A for professional and enthusiast programmers
1 select top 50000 * from posts where posts.ViewCount <= 41234 and posts.ViewCount > 41231 order by posts.ViewCount DESC
permalink
<< show sidebar
Run Query
Cancel
Options:
Text-only results
Include execution plan
Switch to meta site
search by name or url
Results
Messages
Download CSV
Id
PostTypeI...
AcceptedAnswerI...
Pare...
CreationDate
DeletionDate
Score
ViewCount
Body
15258594
1
15259359
2013-03-06 21:38:46
5
41234
<p> I'm trying to start project with Hibernate a...
Type here to search
13°C
ENG
23:07
16/10/2021

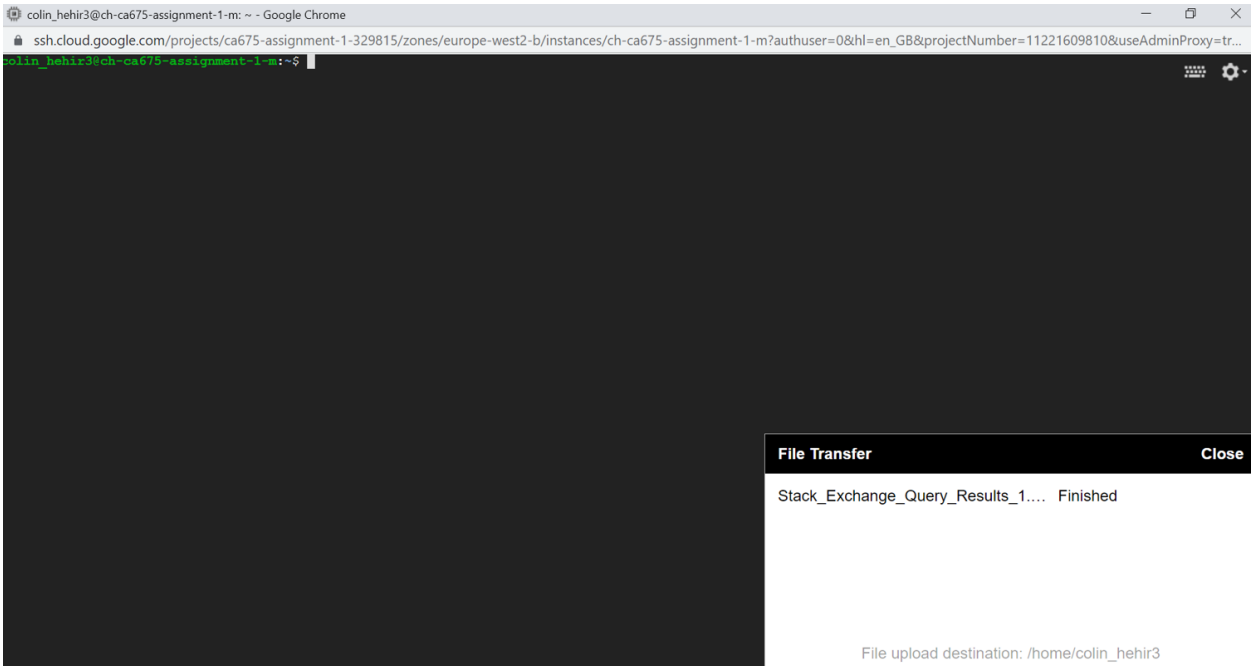
APPENDIX B: ENVIRONMENT VARIABLES SET UP IN GCP HADOOP CLUSTER

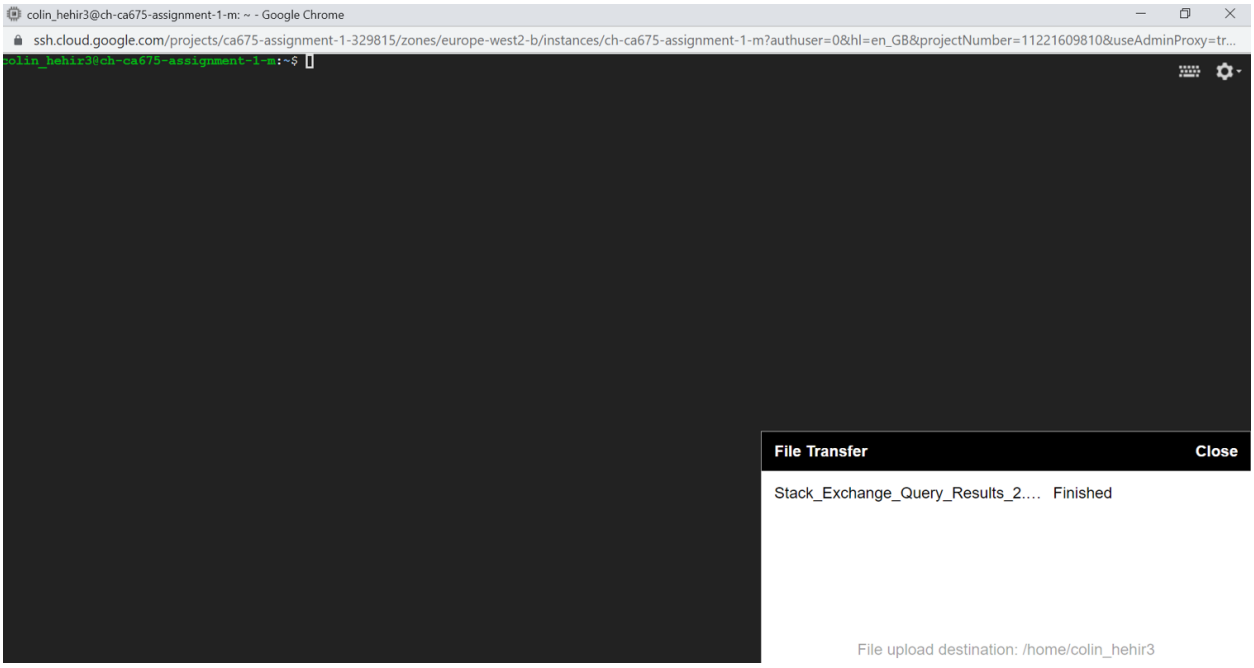
Please refer to the '1. Source Code - Environment Variables Set Up in GCP Hadoop Cluster.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis

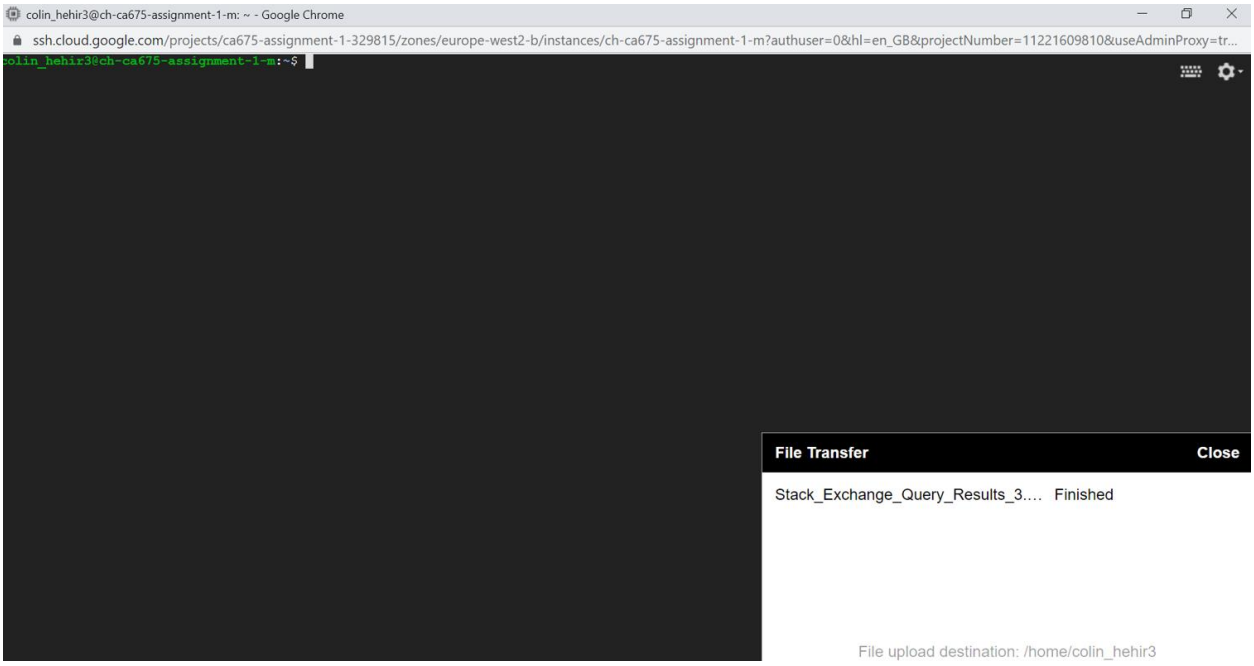


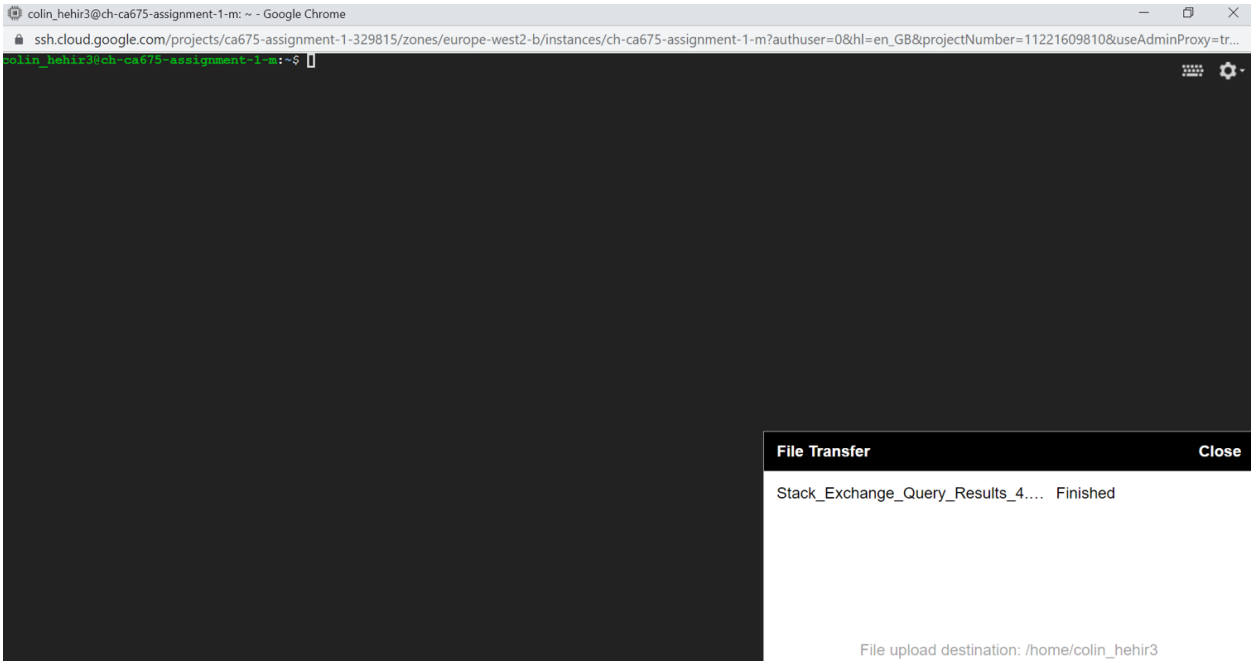


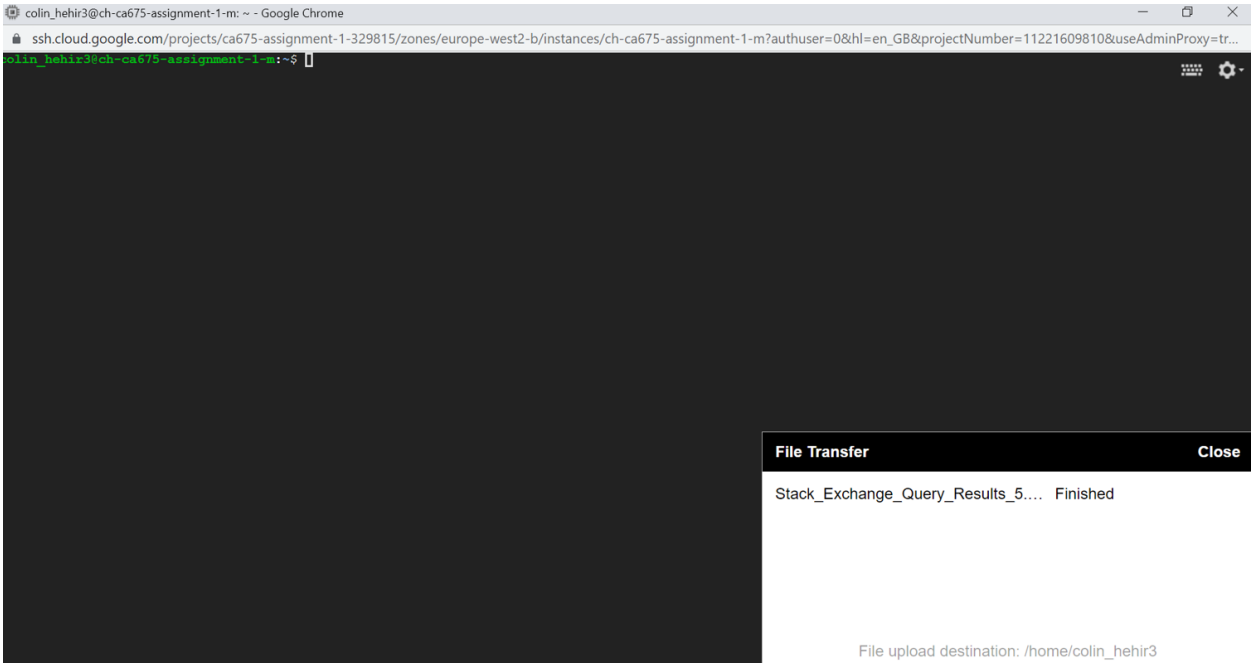

Google Cloud Platform
202133171-CA675-Assignment-1
Search products and resources
Cloud Storage
Bucket details
REFRESH
LEARN
Browser
Monitoring
Settings
dataproc-staging-europe-west2-11221609810-kalheazk
Location
europe-west2 (London)
Storage class
Standard
Public access
Subject to object ACLs
Protection
None
OBJECTS
CONFIGURATION
PERMISSIONS
PROTECTION
LIFECYCLE
Buckets
dataproc-staging-europe-west2-11221609810-kalheazk
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
Stack_Exchange_Query_Results_1.
52.3 MB
application/vnd.ms-excel
22 Oct 20...
Standard
22 Oct 202...
Not public
Stack_Exchange_Query_Results_2.
59.4 MB
Stack_Exchange_Query_Results_3.
62.3 MB
Stack_Exchange_Query_Results_4.
65.5 MB
Stack_Exchange_Query_Results_5.
16.7 KB
google-cloud-dataproc-metainfo/
Folder
Release notes


colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
colin_hehir3@ch-ca675-assignment-1-m:~$
File Transfer
Close
Stack_Exchange_Query_Results_1.... Finished
File upload destination: /home/colin_hehir3


colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
colin_hehir3@ch-ca675-assignment-1-m:~$
File Transfer
Close
Stack_Exchange_Query_Results_2.... Finished
File upload destination: /home/colin_hehir3

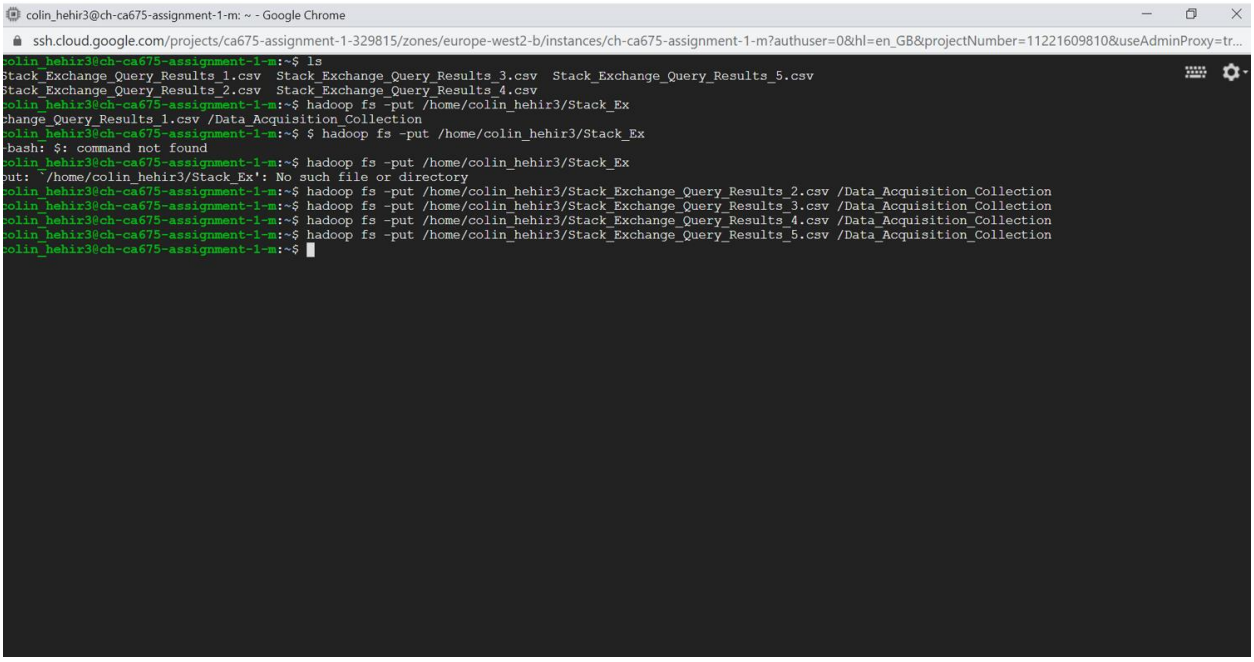

colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
colin_hehir3@ch-ca675-assignment-1-m:~$
File Transfer
Close
Stack_Exchange_Query_Results_3.... Finished
File upload destination: /home/colin_hehir3

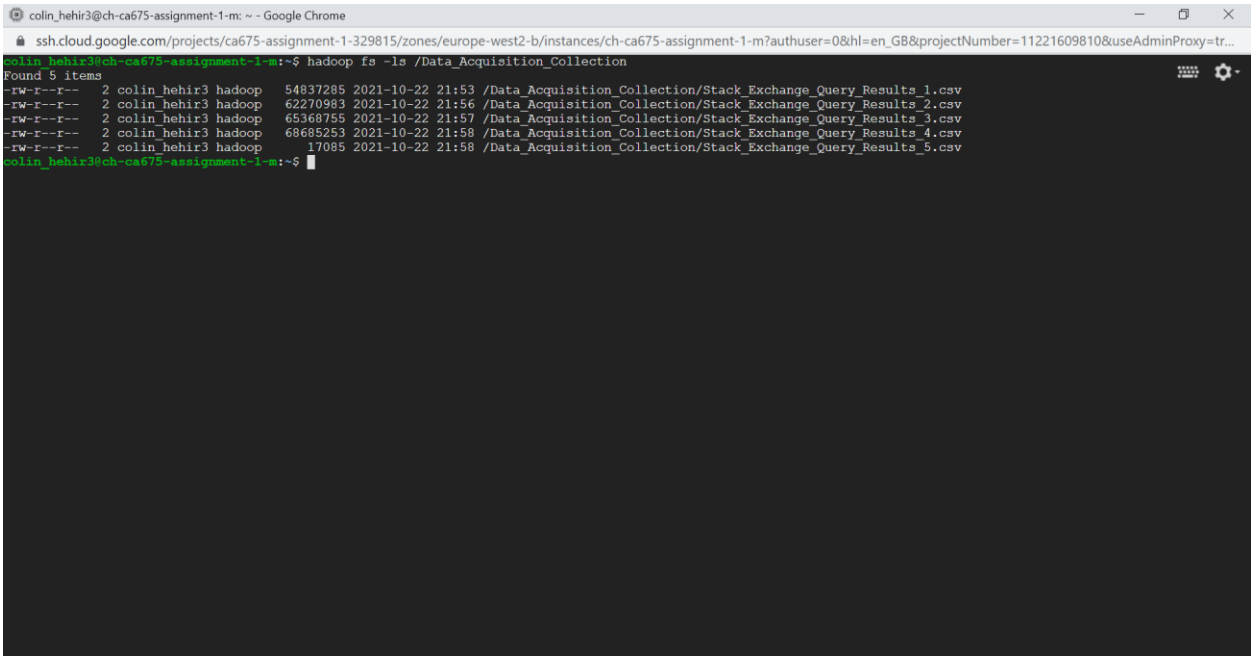

colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
colin_hehir3@ch-ca675-assignment-1-m:~$
File Transfer
Close
Stack_Exchange_Query_Results_4.... Finished
File upload destination: /home/colin_hehir3


colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
colin_hehir3@ch-ca675-assignment-1-m:~$
File Transfer
Close
Stack_Exchange_Query_Results_5.... Finished
File upload destination: /home/colin_hehir3

APPENDIX C: LOAD DATA FROM FILES INTO HDFS

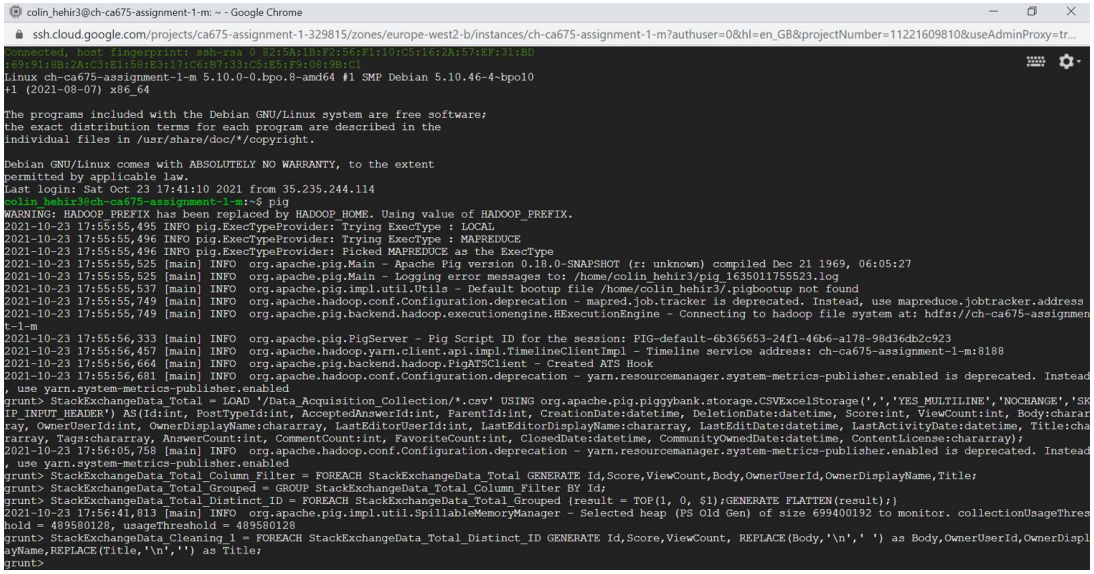
Please refer to the '2. Source Code - Load Data from Files into HDFS.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis





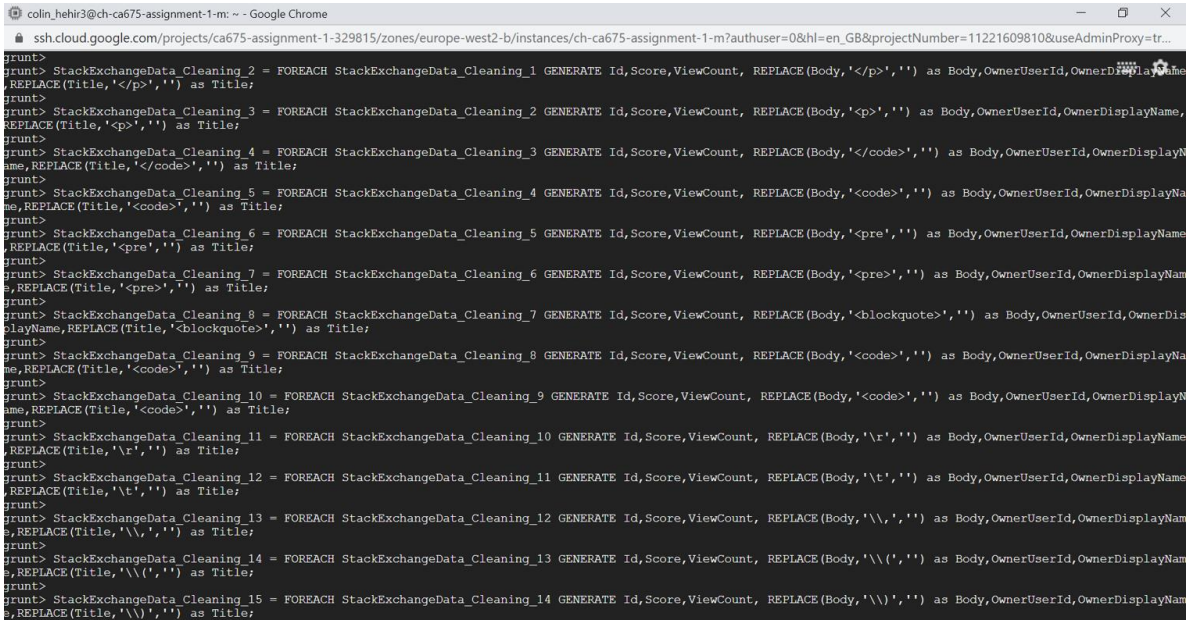
APPENDIX D: PIG ETL - LOAD, MERGE AND REMOVE DUPLICATES

Please refer to the '3. Source Code - Pig ETL - Load, Merge and Remove Duplicates.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis



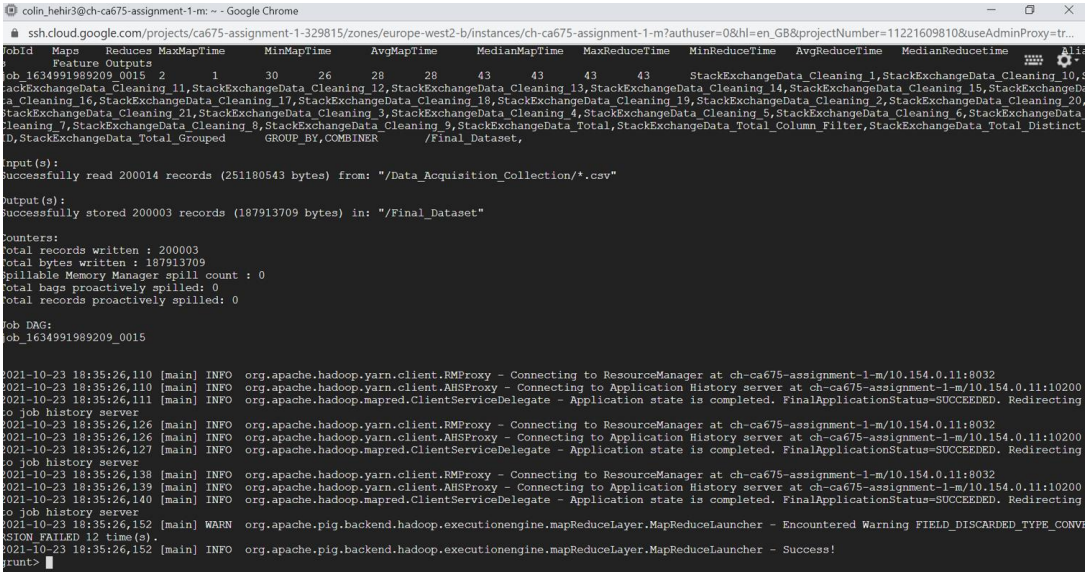
APPENDIX E: PIG ETL - DATA CLEANING - REMOVING DATA ARTEFACTS

Please refer to the '4. Source Code - Pig ETL - Data Cleaning - Removing Data Artefacts.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis



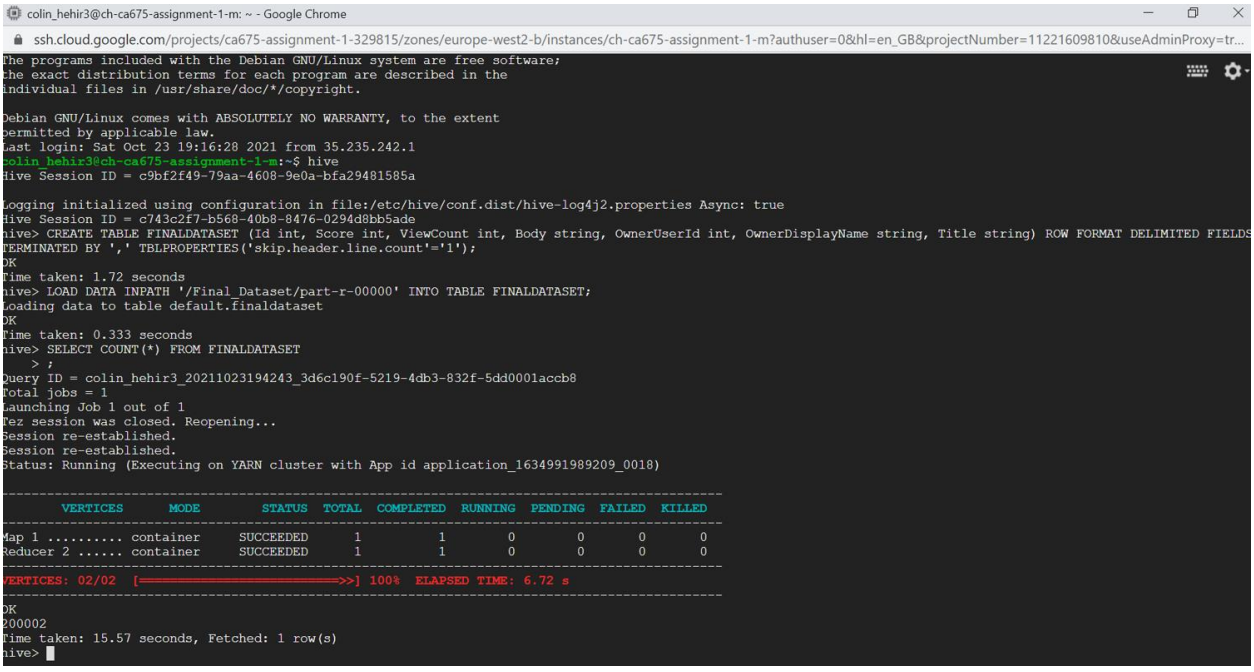
APPENDIX F: PIG ETL - STORE NEW DATASET

Please refer to the '5. Source Code - Pig ETL - Store New Dataset.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis



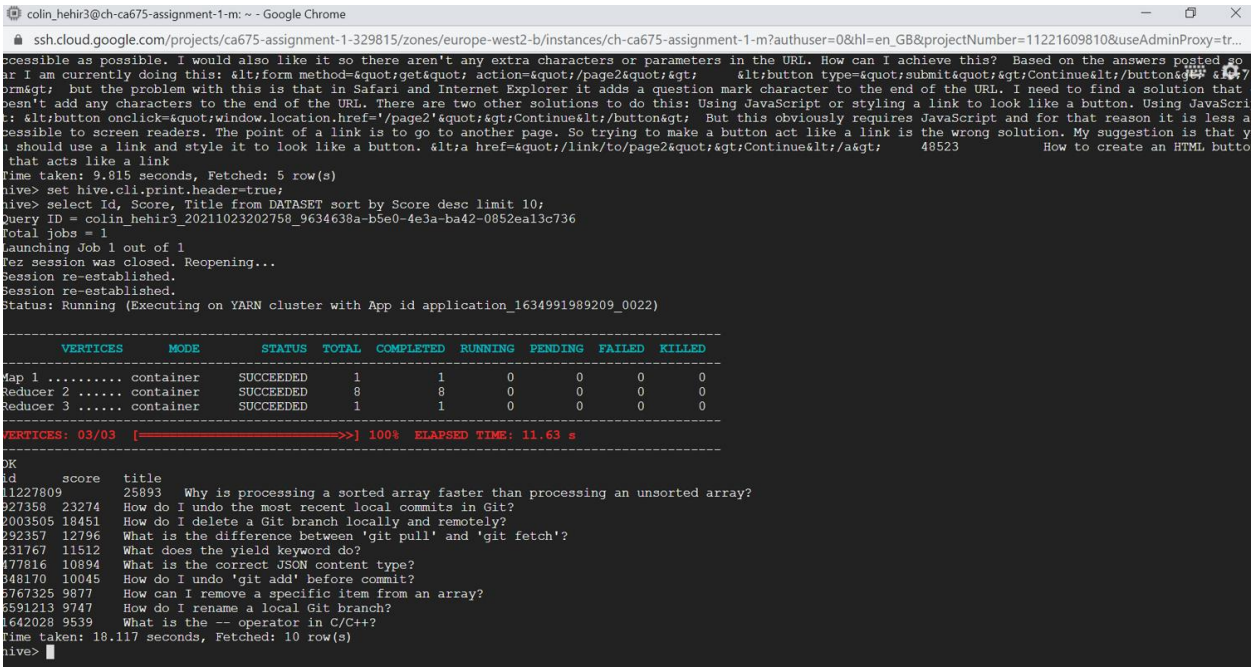
APPENDIX G: HIVE - TABLE CREATION AND DATA LOAD

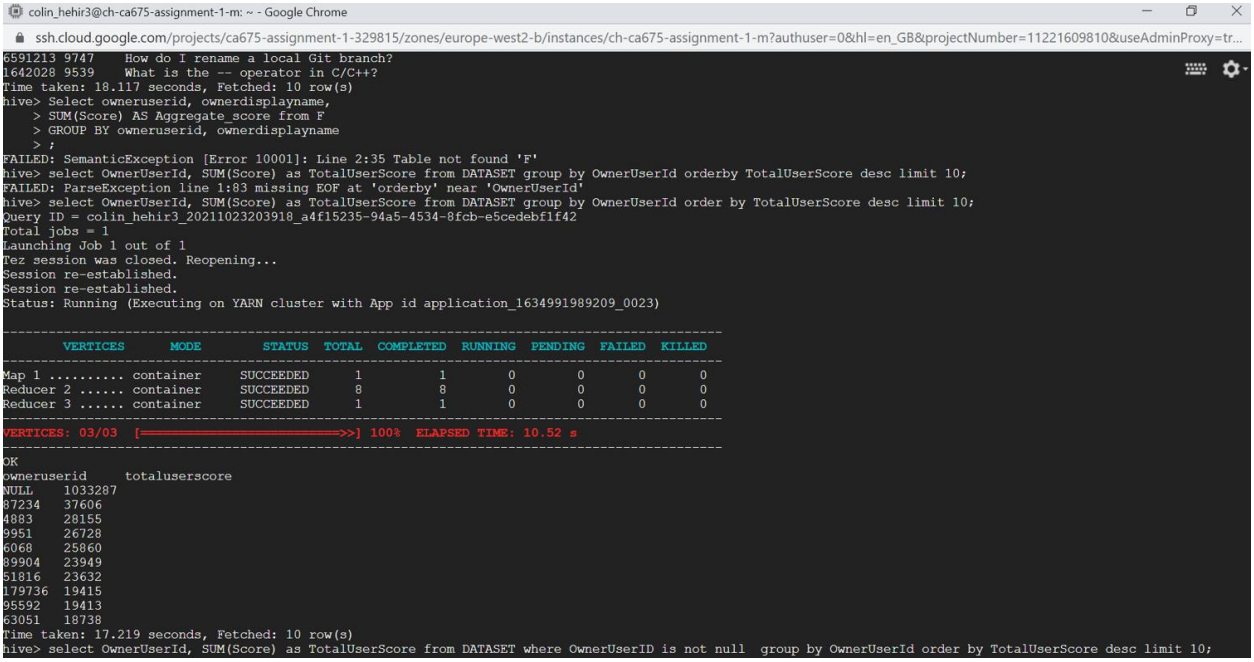
Please refer to the '6. Source Code - Hive - Table Creation and Data Load.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis

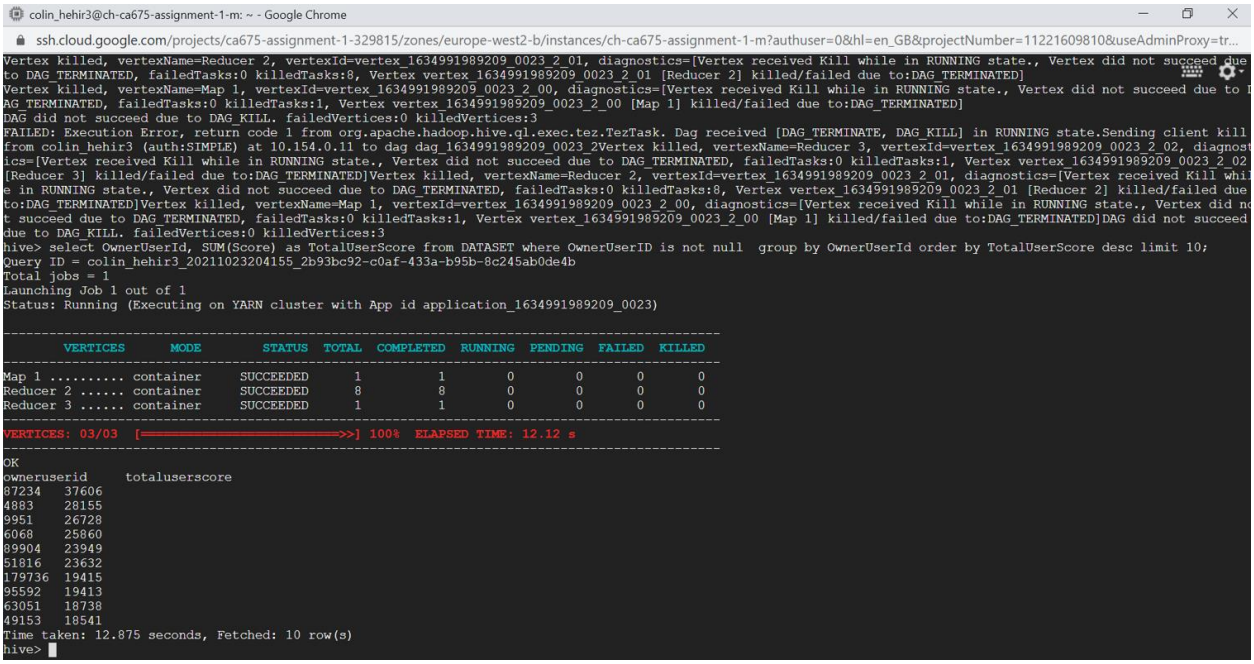


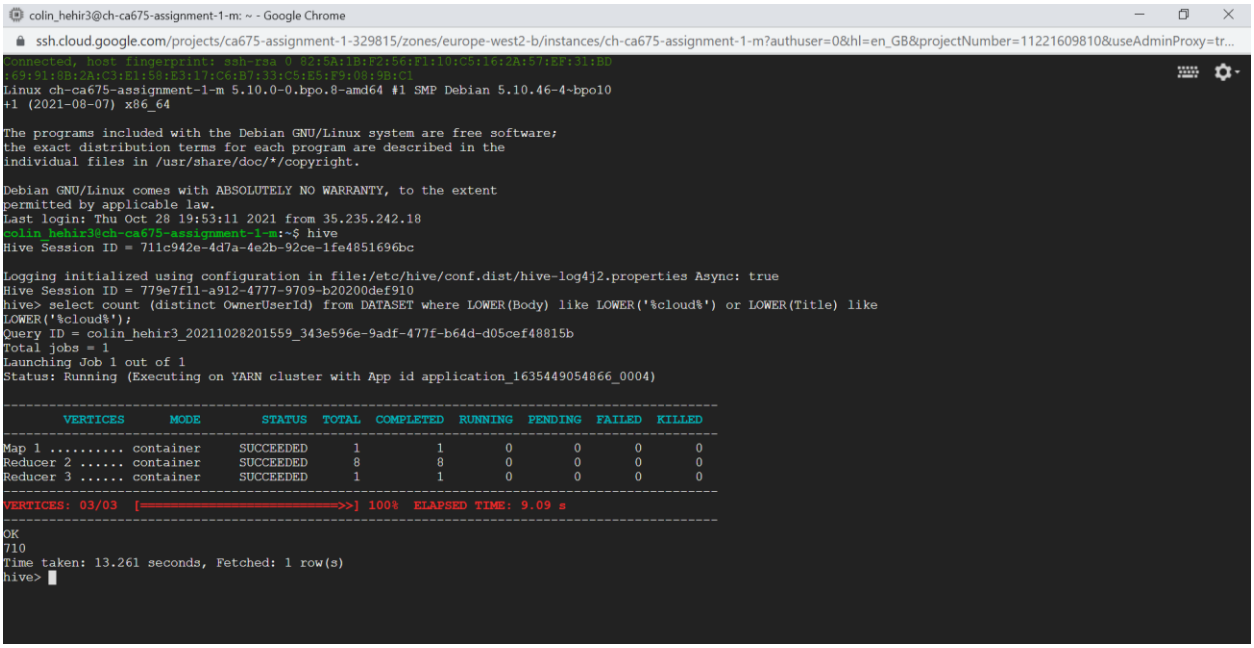
APPENDIX H: HIVE - TASK 3 QUERIES

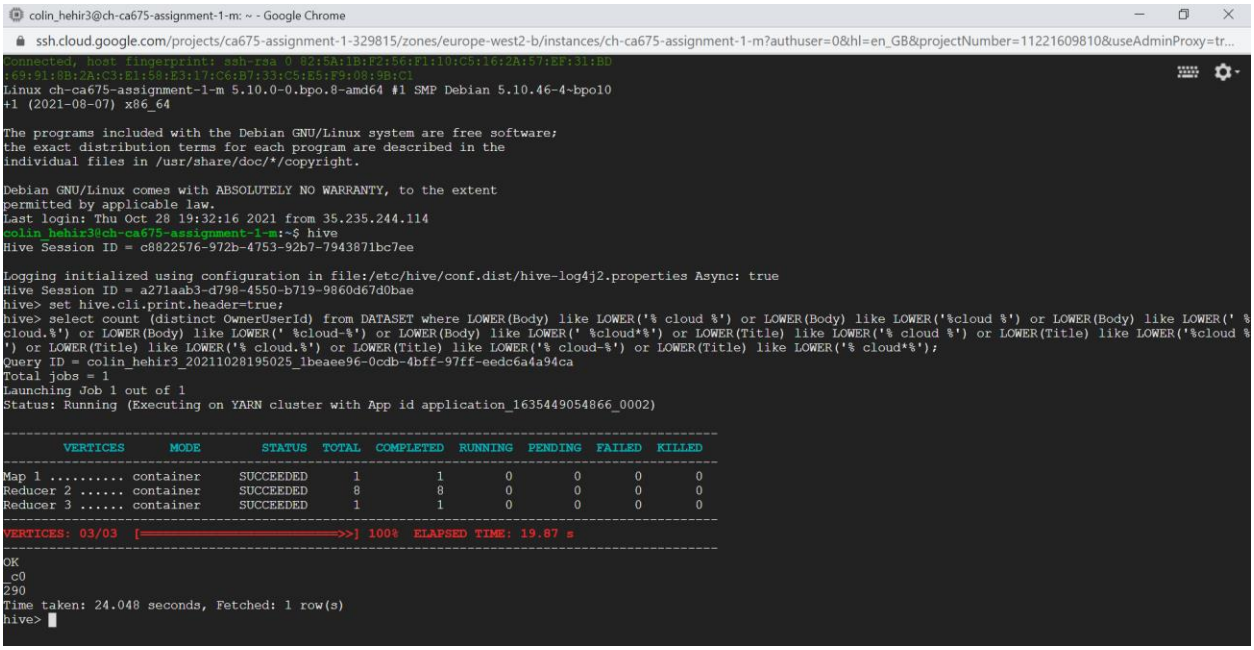
Please refer to the '7. Source Code - Hive - Task 3 Queries.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis





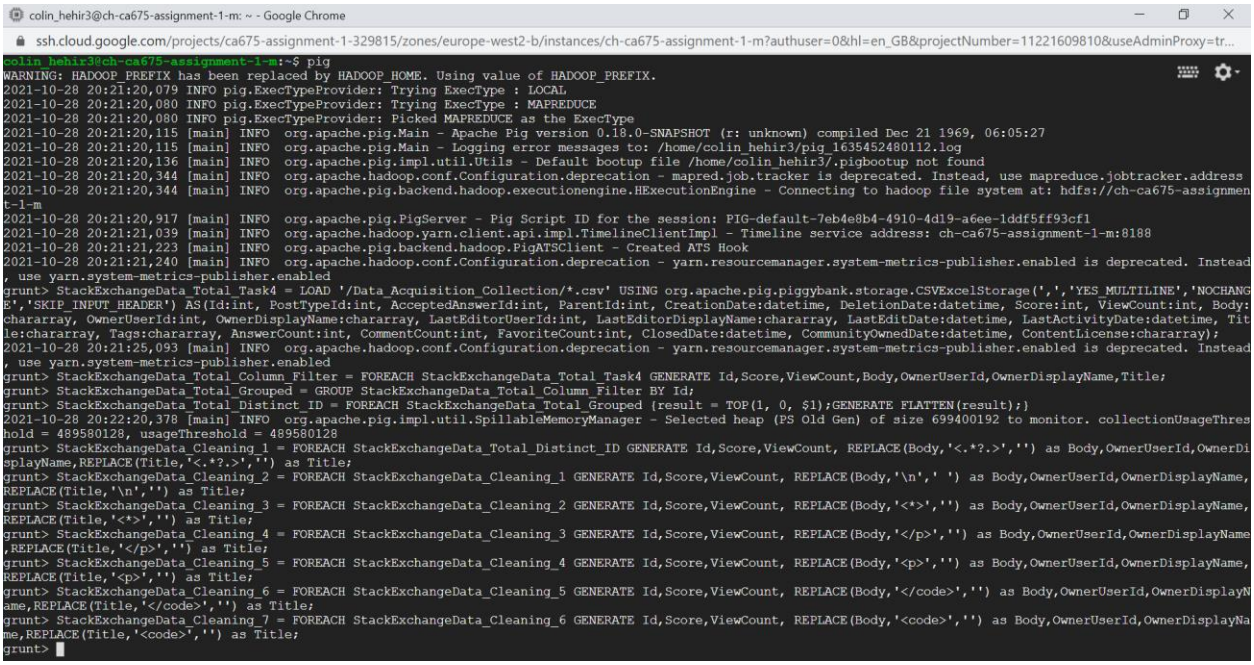

colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
Vertex killed, vertexName=Reducer 2, vertexId=vertex_1634991989209_0023_2_01, diagnostics=[Vertex received Kill while in RUNNING state., Vertex did not succeed due to DAG_TERMINATED, failedTasks:0 killedTasks:8, Vertex vertex_1634991989209_0023_2_01 [Reducer 2] killed/failed due to:DAG_TERMINATED]
Vertex killed, vertexName=Map 1, vertexId=vertex_1634991989209_0023_2_00, diagnostics=[Vertex received Kill while in RUNNING state., Vertex did not succeed due to DAG_TERMINATED, failedTasks:0 killedTasks:1, Vertex vertex_1634991989209_0023_2_00 [Map 1] killed/failed due to:DAG_TERMINATED]
DAG did not succeed due to DAG_KILL. failedVertices:0 killedVertices:3
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.tez.TezTask. Dag received [DAG_TERMINATE, DAG_KILL] in RUNNING state.Sending client kill from colin_hehir3 (auth:SIMPLE) at 10.154.0.11 to dag dag_1634991989209_0023_2Vertex killed, vertexName=Reducer 3, vertexId=vertex_1634991989209_0023_2_02, diagnostics=[Vertex received Kill while in RUNNING state., Vertex did not succeed due to DAG_TERMINATED, failedTasks:0 killedTasks:1, Vertex vertex_1634991989209_0023_2_02 [Reducer 3] killed/failed due to:DAG_TERMINATED]Vertex killed, vertexName=Reducer 2, vertexId=vertex_1634991989209_0023_2_01, diagnostics=[Vertex received Kill while in RUNNING state., Vertex did not succeed due to DAG_TERMINATED, failedTasks:0 killedTasks:8, Vertex vertex_1634991989209_0023_2_01 [Reducer 2] killed/failed due to:DAG_TERMINATED]Vertex killed, vertexName=Map 1, vertexId=vertex_1634991989209_0023_2_00, diagnostics=[Vertex received Kill while in RUNNING state., Vertex did not succeed due to DAG_TERMINATED, failedTasks:0 killedTasks:1, Vertex vertex_1634991989209_0023_2_00 [Map 1] killed/failed due to:DAG_TERMINATED]DAG did not succeed due to DAG_KILL. failedVertices:0 killedVertices:3
hive> select OwnerUserId, SUM(Score) as TotalUserScore from DATASET where OwnerUserID is not null group by OwnerUserId order by TotalUserScore desc limit 10;
Query ID = colin_hehir3_20211023204155_2b93bc92-c0af-433a-b95b-8c245ab0de4b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1634991989209_0023)
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 8 8 0 0 0 0
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 03/03 [==========>>] 100% ELAPSED TIME: 12.12 s
OK
owneruserid totaluserscore
87234 37606
4883 28155
9951 26728
6068 25860
89904 23949
51816 23632
179736 19415
95592 19413
63051 18738
49153 18541
Time taken: 12.875 seconds, Fetched: 10 row(s)
hive>


colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
Connected, host fingerprint: ssh-rsa 0 82:5A:1B:F2:56:F1:10:C5:16:2A:57:EF:31:BD
:69:91:8B:2A:C3:E1:58:E3:17:C6:B7:33:C5:E5:F9:08:9B:C1
Linux ch-ca675-assignment-1-m 5.10.0-0.bpo.8-amd64 #1 SMP Debian 5.10.46-4~bpo10
+1 (2021-08-07) x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Oct 28 19:53:11 2021 from 35.235.242.18
colin_hehir3@ch-ca675-assignment-1-m:~$ hive
Hive Session ID = 711c942e-4d7a-4e2b-92ce-1fe4851696bc
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Hive Session ID = 779e7f11-a912-4777-9709-b20200def910
hive> select count (distinct OwnerUserId) from DATASET where LOWER(Body) like LOWER('%cloud%') or LOWER(Title) like LOWER('%cloud%');
Query ID = colin_hehir3_20211028201559_343e596e-9adf-477f-b64d-d05cef48815b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635449054866_0004)
VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 8 8 0 0 0 0
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 03/03 [==========>>] 100% ELAPSED TIME: 9.09 s
OK
710
Time taken: 13.261 seconds, Fetched: 1 row(s)
hive>



APPENDIX I: TASK 4 - PIG ETL - CLEANING AND FILTER ON 3.2 TOP USERS

Please refer to the '8. Source Code - Task 4 - Pig ETL - Cleaning and Filter On 3.2 Top Users.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis



colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome

ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...

```
grunt> StackExchangeData_Cleaning_7 = FOREACH StackExchangeData_Cleaning_6 GENERATE Id,Score,ViewCount, REPLACE(Body,'<code>','') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE(Title,'<code>','') as Title;
grunt> StackExchangeData_Cleaning_8 = FOREACH StackExchangeData_Cleaning_7 GENERATE Id,Score,ViewCount, REPLACE(Body,'<pre','') as Body,OwnerUserId,OwnerDisplayName
,REPLACE(Title,'<pre','') as Title;
grunt> StackExchangeData_Cleaning_9 = FOREACH StackExchangeData_Cleaning_8 GENERATE Id,Score,ViewCount, REPLACE(Body,'</pre>','') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE(Title,'</pre>','') as Title;
grunt> StackExchangeData_Cleaning_10 = FOREACH StackExchangeData_Cleaning_9 GENERATE Id,Score,ViewCount, REPLACE(Body,'</blockquote>','') as Body,OwnerUserId,OwnerD
isplayName,REPLACE(Title,'</blockquote>','') as Title;
grunt> StackExchangeData_Cleaning_11 = FOREACH StackExchangeData_Cleaning_10 GENERATE Id,Score,ViewCount, REPLACE(Body,'</ul>','') as Body,OwnerUserId,OwnerDisplayN
ame,REPLACE(Title,'</ul>','') as Title;
grunt> StackExchangeData_Cleaning_12 = FOREACH StackExchangeData_Cleaning_11 GENERATE Id,Score,ViewCount, REPLACE(Body,'</li>','') as Body,OwnerUserId,OwnerDisplayN
ame,REPLACE(Title,'<li>','') as Title;
grunt> StackExchangeData_Cleaning_13 = FOREACH StackExchangeData_Cleaning_12 GENERATE Id,Score,ViewCount, REPLACE(Body,'\r','') as Body,OwnerUserId,OwnerDisplayName
,REPLACE(Title,'\r','') as Title;
grunt> StackExchangeData_Cleaning_14 = FOREACH StackExchangeData_Cleaning_13 GENERATE Id,Score,ViewCount, REPLACE(Body,'\t','') as Body,OwnerUserId,OwnerDisplayName
,REPLACE(Title,'\t','') as Title;
grunt> StackExchangeData_Cleaning_15 = FOREACH StackExchangeData_Cleaning_14 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\,','') as Body,OwnerUserId,OwnerDisplayNam
e,REPLACE(Title,'\\,','') as Title;
grunt> StackExchangeData_Cleaning_16 = FOREACH StackExchangeData_Cleaning_15 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\(','') as Body,OwnerUserId,OwnerDisplayNam
e,REPLACE(Title,'\\(','') as Title;
grunt> StackExchangeData_Cleaning_17 = FOREACH StackExchangeData_Cleaning_16 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\)','') as Body,OwnerUserId,OwnerDisplayNam
e,REPLACE(Title,'\\)','') as Title;
grunt> StackExchangeData_Cleaning_18 = FOREACH StackExchangeData_Cleaning_17 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\"','') as Body,OwnerUserId,OwnerDisplayNam
e,REPLACE(Title,'\\"','') as Title;
grunt> StackExchangeData_Cleaning_19 = FOREACH StackExchangeData_Cleaning_18 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\<*p>','') as Body,OwnerUserId,OwnerDisplay
Name,REPLACE(Title,'\\<*p>','') as Title;
grunt> StackExchangeData_Cleaning_20 = FOREACH StackExchangeData_Cleaning_19 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\<','') as Body,OwnerUserId,OwnerDisplayNam
e,REPLACE(Title,'\\<','') as Title;
grunt> StackExchangeData_Cleaning_21 = FOREACH StackExchangeData_Cleaning_20 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\>','') as Body,OwnerUserId,OwnerDisplayNam
e,REPLACE(Title,'\\>','') as Title;
grunt> StackExchangeData_Total_Task4_Cleaned = FILTER StackExchangeData_Cleaning_21 BY (OwnerUserId == 87234 or OwnerUserId == 4883 or OwnerUserId == 9951 or OwnerU
serId == 6068 or OwnerUserId == 89904 or OwnerUserId == 51816 or OwnerUserId == 179736 or OwnerUserId == 95592 or OwnerUserId == 63051 or OwnerUserId == 4915) ;
grunt> StackExchangeData_Total_Task4_Title_Body = FOREACH StackExchangeData_Total_Task4_Cleaned GENERATE OwnerUserId, CONCAT(Body,' ',Title) AS Body_Title;
grunt> StackExchangeData_Total_Task4_Final = FOREACH StackExchangeData_Total_Task4_Title_Body GENERATE OwnerUserId, LOWER(TRIM(Body_Title)) AS Body_Title;
grunt> STORE StackExchangeData_Total_Task4_Final INTO '/user/colin_hehir3/StackExchangeData_Total_Task4_Final_Output' USING org.apache.pig.piggybank.storage.CSVExce
lStorage(',','NO_MULTILINE','NOCHANGE','SKIP_OUTPUT_HEADER');
2021-10-28 20:32:22,175 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead
, use yarn.system-metrics-publisher.enabled
2021-10-28 20:32:22,211 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.outp
ut.textoutputformat.separator
2021-10-28 20:32:22,242 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY,FILTER
2021-10-28 20:32:22,261 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead
```

colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome

ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProx...

```
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -ls /user/colin_hehir3/StackExchangeData_Total_Task4_Final_Output
Found 2 items
-rw-r--r--   2 colin_hehir3 hadoop          0 2021-10-28 20:33 /user/colin_hehir3/StackExchangeData_Total_Task4_Final_Output/_SUCCESS
-rw-r--r--   2 colin_hehir3 hadoop     133713 2021-10-28 20:33 /user/colin_hehir3/StackExchangeData_Total_Task4_Final_Output/part-r-00000
colin_hehir3@ch-ca675-assignment-1-m:~$
```

Type here to search

12°C ENG 21:40 28/10/2021

APPENDIX J: TASK 4 - MAPREDUCE - TFIDF IMPLEMENTATION CREATION AND OUTPUT

Please refer to the '9. Source Code - Task 4 - MapReduce - TFIDF Implementation Creation and Output.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis

```
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file /home/colin_hehir3/mapper1.py /home/colin_hehir3/reducer1.py -mapper "
python mapper1.py" -reducer "python reducer1.py" -input /user/colin_hehir3/StackExchangeData_Total_Task4_Output/part-m-*/ -output /user/colin_hehir3/ch_output1
2021-10-25 15:11:47,042 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/colin_hehir3/mapper1.py, /home/colin_hehir3/reducer1.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob6044837599962810376.jar tm
pDir=null
2021-10-25 15:11:47,677 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:11:47,866 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:11:48,227 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:11:48,228 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:11:48,365 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/colin_hehir3/.staging/job_1635169827887_0005
2021-10-25 15:11:48,677 INFO mapred.FileInputFormat: Total input files to process : 2
2021-10-25 15:11:48,740 INFO mapreduce.JobSubmitter: number of splits:34
2021-10-25 15:11:48,847 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635169827887_0005
2021-10-25 15:11:48,848 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-25 15:11:49,017 INFO conf.Configuration: resource-types.xml not found
2021-10-25 15:11:49,018 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-25 15:11:49,065 INFO impl.YarnClientImpl: Submitted application application_1635169827887_0005
2021-10-25 15:11:49,113 INFO mapreduce.Job: The url to track the job: http://ch-ca675-assignment-1-m:8088/proxy/application_1635169827887_0005/
2021-10-25 15:11:49,115 INFO mapreduce.Job: Running job: job_1635169827887_0005
2021-10-25 15:11:56,288 INFO mapreduce.Job: Job job_1635169827887_0005 running in uber mode : false
2021-10-25 15:11:56,288 INFO mapreduce.Job:  map 0% reduce 0%
2021-10-25 15:12:05,387 INFO mapreduce.Job:  map 12% reduce 0%
```

```
colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...
2021-10-25 15:14:20,053 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/colin_hehir3/mapper2.py, /home/colin_hehir3/reducer2.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob6746467992186234[illegible].jar tm
pDir=null
2021-10-25 15:14:20,674 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:14:20,857 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:14:21,213 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:14:21,213 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:14:21,339 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/colin_hehir3/.staging/job_1635169827887_0006
2021-10-25 15:14:21,593 INFO mapreduce.JobSubmitter: Cleaning up the staging area /tmp/hadoop-yarn/staging/colin_hehir3/.staging/job_1635169827887_0006
2021-10-25 15:14:21,593 WARN concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #0,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
2021-10-25 15:14:21,600 ERROR streaming.StreamJob: Error Launching job : Input Pattern /user/colin_hehir3/choutput1/part-0000* matches 0 files
Streaming Command Failed!
colin_hehir3@ch-ca675-assignment-1-m:~$  hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file /home/colin_hehir3/mapper2.py /home/colin_hehir3/reducer2.py -mapper
"python mapper2.py" -reducer "python reducer2.py" -input /user/colin_hehir3/ch_output1/part-0000* -output /user/colin_hehir3/ch_output2
2021-10-25 15:15:08,863 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/colin_hehir3/mapper2.py, /home/colin_hehir3/reducer2.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob3844858372756429927.jar tm
pDir=null
2021-10-25 15:15:09,489 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:15:09,682 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:15:10,036 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:15:10,036 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:15:10,163 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/colin_hehir3/.staging/job_1635169827887_0007
2021-10-25 15:15:10,449 INFO mapred.FileInputFormat: Total input files to process : 10
2021-10-25 15:15:10,525 INFO mapreduce.JobSubmitter: number of splits:39
2021-10-25 15:15:10,636 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635169827887_0007
2021-10-25 15:15:10,638 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-25 15:15:10,810 INFO conf.Configuration: resource-types.xml not found
2021-10-25 15:15:10,810 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-25 15:15:10,856 INFO impl.YarnClientImpl: Submitted application application_1635169827887_0007
2021-10-25 15:15:10,883 INFO mapreduce.Job: The url to track the job: http://ch-ca675-assignment-1-m:8088/proxy/application_1635169827887_0007/
2021-10-25 15:15:10,885 INFO mapreduce.Job: Running job: job_1635169827887_0007
2021-10-25 15:15:18,005 INFO mapreduce.Job: Job job_1635169827887_0007 running in uber mode : false
2021-10-25 15:15:18,006 INFO mapreduce.Job:  map 0% reduce 0%
```

colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome

ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...

```
                CPU time spent (ms)=36500
                Physical memory (bytes) snapshot=22892367872
                Virtual memory (bytes) snapshot=221407260672
                Total committed heap usage (bytes)=22412787712
                Peak Map Physical memory (bytes)=557903872
                Peak Map Virtual memory (bytes)=4431421440
                Peak Reduce Physical memory (bytes)=282693632
                Peak Reduce Virtual memory (bytes)=4438765568
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=139251
        File Output Format Counters
                Bytes Written=85600
2021-10-25 15:16:00,341 INFO streaming.StreamJob: Output directory: /user/colin_hehir3/ch_output2
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file /home/colin_hehir3/mapper3.py /home/colin_hehir3/reducer3.py -mapper "
python mapper3.py" -reducer "python reducer3.py" -input /user/colin_hehir3/ch_output2/part-0000* -output /user/colin_hehir3/ch_output3
2021-10-25 15:17:13,806 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/colin_hehir3/mapper3.py, /home/colin_hehir3/reducer3.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob4877655264200457079.jar tm
pDir=null
2021-10-25 15:17:14,418 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:17:14,605 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:17:14,933 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:17:14,934 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:17:15,064 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/colin_hehir3/.staging/job_1635169827887_0008
2021-10-25 15:17:15,349 INFO mapred.FileInputFormat: Total input files to process : 10
2021-10-25 15:17:15,413 INFO mapreduce.JobSubmitter: number of splits:41
2021-10-25 15:17:15,521 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635169827887_0008
2021-10-25 15:17:15,523 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-25 15:17:15,697 INFO conf.Configuration: resource-types.xml not found
2021-10-25 15:17:15,697 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-25 15:17:15,738 INFO impl.YarnClientImpl: Submitted application application_1635169827887_0008
2021-10-25 15:17:15,763 INFO mapreduce.Job: The url to track the job: http://ch-ca675-assignment-1-m:8088/proxy/application_1635169827887_0008/
2021-10-25 15:17:15,764 INFO mapreduce.Job: Running job: job_1635169827887_0008
2021-10-25 15:17:22,871 INFO mapreduce.Job: Job job_1635169827887_0008 running in uber mode : false
2021-10-25 15:17:22,872 INFO mapreduce.Job:  map 0% reduce 0%
```

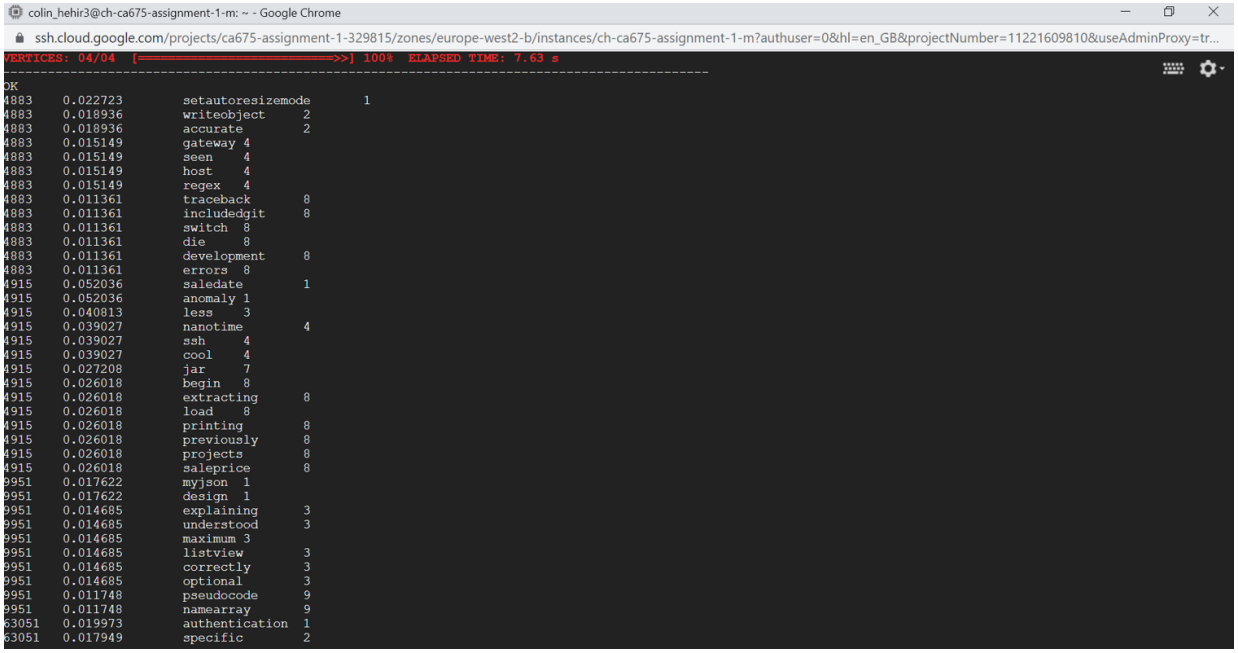
colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome

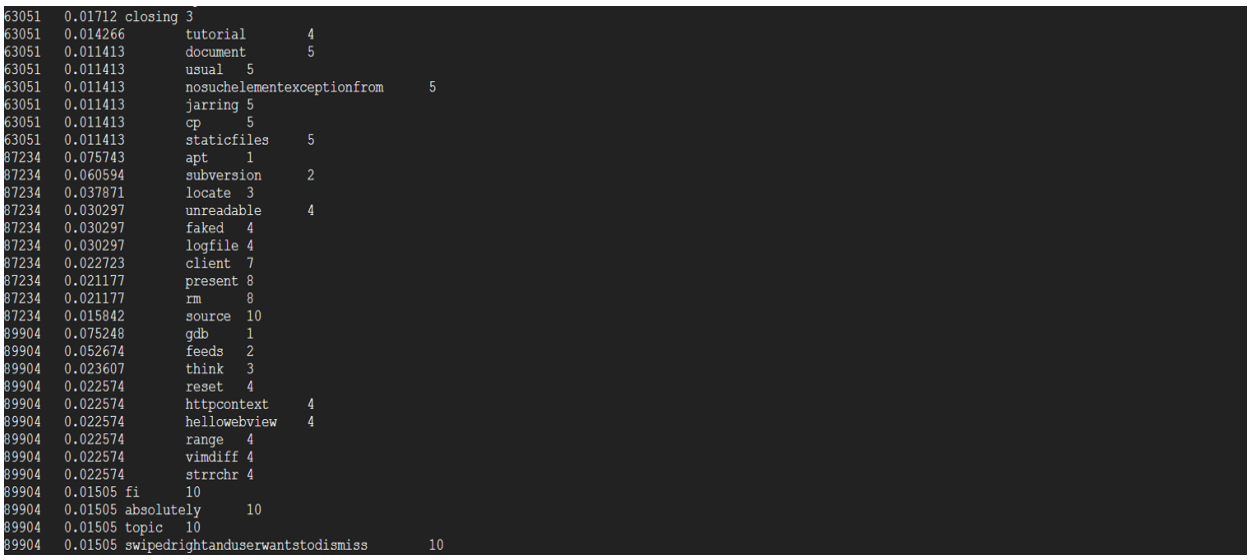
ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...

```
                Failed Shuffles=0
                Merged Map outputs=451
                GC time elapsed (ms)=7816
                CPU time spent (ms)=38210
                Physical memory (bytes) snapshot=24246767616
                Virtual memory (bytes) snapshot=230250393600
                Total committed heap usage (bytes)=23904387072
                Peak Map Physical memory (bytes)=553611264
                Peak Map Virtual memory (bytes)=4428316672
                Peak Reduce Physical memory (bytes)=298684416
                Peak Reduce Virtual memory (bytes)=4433317888
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=152692
        File Output Format Counters
                Bytes Written=40270
2021-10-25 15:18:05,217 INFO streaming.StreamJob: Output directory: /user/colin_hehir3/ch_output3
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop jar /usr/lib/hadoop/hadoop-streaming.jar -file /home/colin_hehir3/mapper4.py -mapper "python mapper4.py" -input /use
/colin_hehir3/ch_output3/part-0000* -output /user/colin_hehir3/ch_output4
2021-10-25 15:19:11,075 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [/home/colin_hehir3/mapper4.py] [/usr/lib/hadoop/hadoop-streaming-3.2.2.jar] /tmp/streamjob6263174121410055354.jar tmpDir=null
2021-10-25 15:19:11,713 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:19:11,900 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:19:12,238 INFO client.RMProxy: Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-25 15:19:12,239 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-25 15:19:12,359 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/colin_hehir3/.staging/job_1635169827887_000
2021-10-25 15:19:12,608 INFO mapred.FileInputFormat: Total input files to process : 10
2021-10-25 15:19:12,675 INFO mapreduce.JobSubmitter: number of splits:39
2021-10-25 15:19:12,783 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1635169827887_0009
2021-10-25 15:19:12,784 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-10-25 15:19:12,978 INFO conf.Configuration: resource-types.xml not found
2021-10-25 15:19:12,978 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-10-25 15:19:13,023 INFO impl.YarnClientImpl: Submitted application application_1635169827887_0009
2021-10-25 15:19:13,047 INFO mapreduce.Job: The url to track the job: http://ch-ca675-assignment-1-m:8088/proxy/application_1635169827887_0009/
2021-10-25 15:19:13,048 INFO mapreduce.Job: Running job: job_1635169827887_0009
```

APPENDIX K: TASK 4 - HIVE QUERY

Please refer to the '10. Source Code - Task 4 - Hive Query.txt' file in the GitHub Repository for the relevant source code: https://github.com/colin-hehir/Cloud-Technologies-Data-Analysis





colin_hehir3@ch-ca675-assignment-1-m: ~ - Google Chrome

ssh.cloud.google.com/projects/ca675-assignment-1-329815/zones/europe-west2-b/instances/ch-ca675-assignment-1-m?authuser=0&hl=en_GB&projectNumber=11221609810&useAdminProxy=tr...

```
87234   0.060594        subversion      2
87234   0.037871        locate  3
87234   0.030297        unreadable      4
87234   0.030297        faked   4
87234   0.030297        logfile 4
87234   0.022723        client  7
87234   0.021177        present 8
87234   0.021177        rm      8
87234   0.015842        source  10
89904   0.075248        gdb     1
89904   0.052674        feeds   2
89904   0.023607        think   3
89904   0.022574        reset   4
89904   0.022574        httpcontext     4
89904   0.022574        hellowebview    4
89904   0.022574        range   4
89904   0.022574        vimdiff 4
89904   0.022574        strrchr 4
89904   0.01505 fi      10
89904   0.01505 absolutely      10
89904   0.01505 topic   10
89904   0.01505 swipedrightanduserwantstodismiss        10
89904   0.01505 side    10
89904   0.01505 january 10
89904   0.01505 jack    10
89904   0.01505 attribute       10
89904   0.01505 failed  10
89904   0.01505 child   10
89904   0.01505 fires   10
179736  0.052625        global  1
179736  0.015695        resource        2
179736  0.012925        favicon 3
179736  0.012002        hits    4
179736  0.012002        disk    4
179736  0.011079        deinit  6
179736  0.011079        learn   6
179736  0.011079        important       6
179736  0.010325        push    9
179736  0.010156        restore 10
179736  0.010156        closure 10
```