



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 (\leq) 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

CA675 Cloud Technologies Assignment 1

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.



CA675 Cloud Technologies Assignment 1

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.CSExcelStorage 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 diffi
231767    11512  What is the diffi
477816    10894  What is the corr
348170    10045  How do I undo c
5767325  9877   How can I remove
6591213  9747   How do I rename
1642028  9539   What is the -- c
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
```



CA675 Cloud Technologies Assignment 1

The result of the query is as follows:

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.

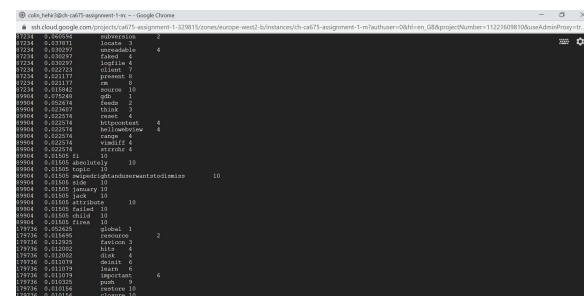
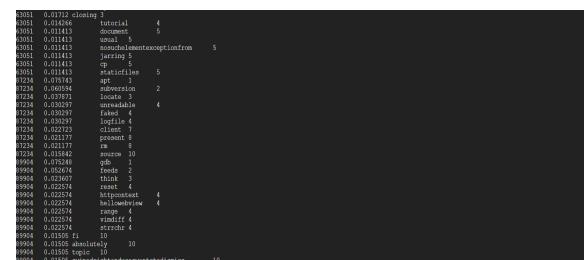
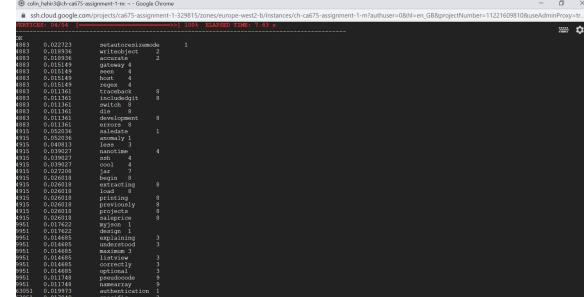
For each row in the dataset, RANK() was used to form a position of hierarchy for each user's TF-IDF values. The PARTITION BY clause was used to define the subset of data in a partition, which in this case was the user's ID and then ORDER BY to rank the highest value of results in descending order.

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

The WHERE clause was (\leq) 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 tfidfrank
DISTRIBUTE BY userId SORT BY OwnerUserId
desc) a WHERE rank <= 10 ORDER BY
OwnerUserId, tfidfrank;
```



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.



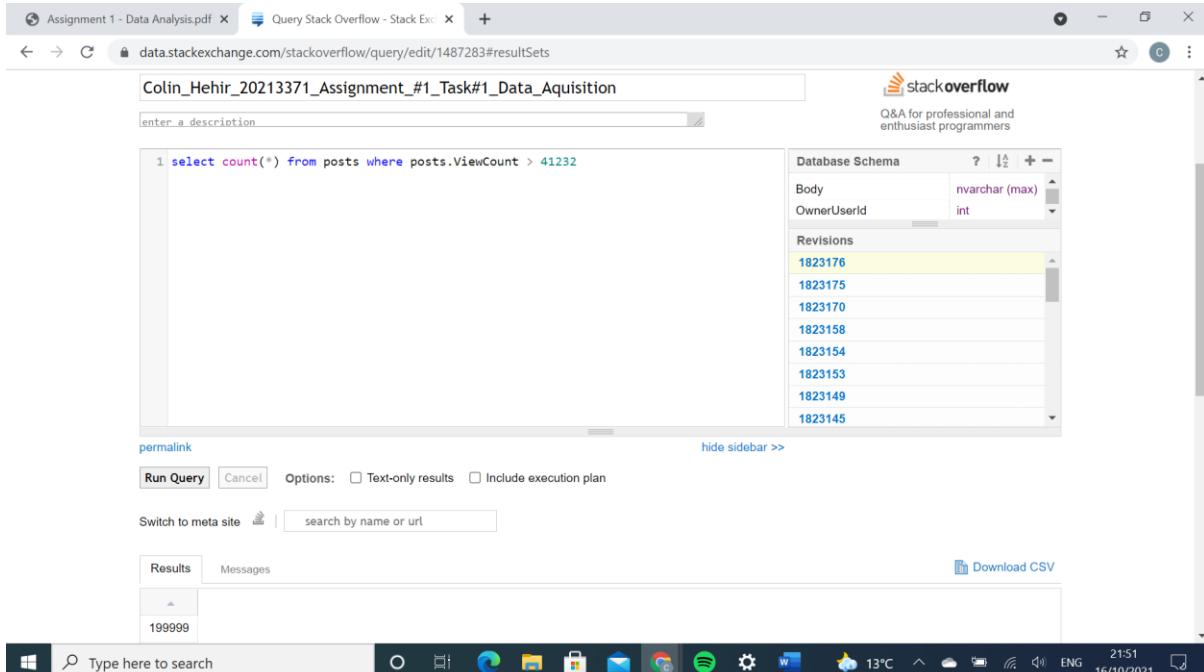
REFERENCES

- 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>

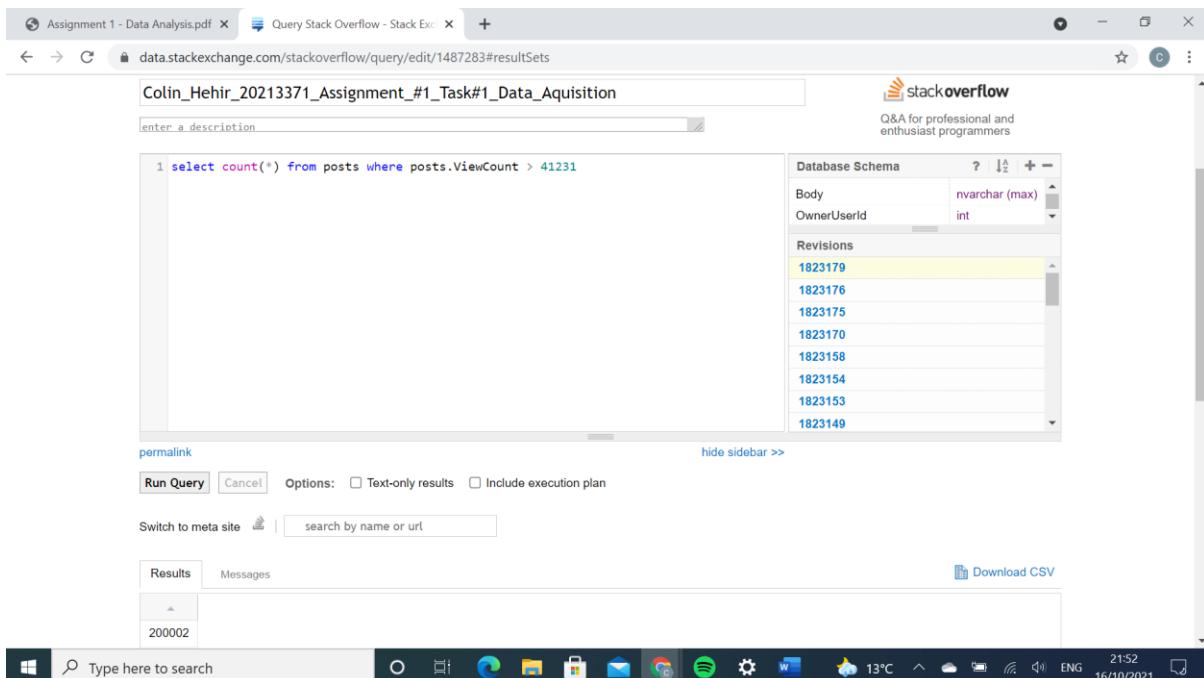


Screenshot of a Stack Overflow query results page. The query is:

```
1 select count(*) from posts where posts.ViewCount > 41232
```

The results table shows 199999 rows. The database schema on the right includes:

Body	nvarchar (max)
OwnerId	int
1823176	
1823175	
1823170	
1823158	
1823154	
1823153	
1823149	
1823145	



Screenshot of a Stack Overflow query results page. The query is:

```
1 select count(*) from posts where posts.ViewCount > 41231
```

The results table shows 200002 rows. The database schema on the right includes:

Body	nvarchar (max)
OwnerId	int
1823179	
1823176	
1823175	
1823170	
1823158	
1823154	
1823153	
1823149	



CA675 Cloud Technologies Assignment 1

A screenshot of a Windows desktop showing a browser window. The title bar says "Assignment 1 - Data Analysis.pdf" and "Query Stack Overflow - Stack Exchange". The URL is "data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets". The page displays a SQL query:

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

The results table has columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, Body. One row is shown:

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

Below the table are "Run Query" and "Cancel" buttons, and "Options" checkboxes for "Text-only results" and "Include execution plan". There's also a "Switch to meta site" link and a search bar. The status bar at the bottom shows "Type here to search" and a taskbar with various icons.

A screenshot of a Windows desktop showing a browser window. The title bar says "Assignment 1 - Data Analysis.pdf" and "Query Stack Overflow - Stack Exchange". The URL is "data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets". The page displays a large dataset of posts. The results table has columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, Body. The table contains many rows, with the last few visible:

Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount	Body
927358	1	927386		2009-05-29 18:09:14		23274	10001585	<p>I accidentally committed the wro...
2003505	1	2003515		2010-01-05 01:12:15		18451	9233751	<p>I want to delete a branch both locally and ...
5767325	1	5767357		2011-04-23 22:17:18		9877	8860159	<p>I have an array of numbers and I'm using ...
16956810	1	16957078		2013-06-08 08:06:45		6270	8177226	<p>I'm trying to find a way to scan my entire ...
2906582	1	2906586		2010-05-25 16:39:47		2294	7760798	<p>I would like to create an HTML button that...
4114095	1	4114122		2010-11-06 16:58:14		7621	7565960	<p>How do I revert from my current state to a...
503093	1	506004		2009-02-02 12:54:16		7710	6918255	<p>How can I redirect the user from one pag...
1789945	1	1789952		2009-11-24 13:04:29		7418	6836325	<p>Usually I would expect a <code>String.co...
5585779	1	5585800		2011-04-07 18:27:54		3275	6383401	<p>How can I convert a <code>String</code>...
1783405	1	1783426		2009-11-23 14:23:46		7521	6217855	<p>Somebody pushed a branch called <code>...
1125968	1	8888015		2009-07-14 14:58:15		8258	6073747	<p>How do I force an overwrite of local files o...
3207219	1			2010-07-08 19:31:22		3466	5983823	<p>How can I list all files of a directory in Pyt...
4366730	1	4366748		2010-12-06 13:14:05		2660	5828599	<p>Consider:</p> <pre><code>\$a = 'How ar...
3437059	1	3437070		2010-08-09 02:52:50		3596	5254761	<p>I'm looking for a <code>string.contains</code>
20035101	1	20035319		2013-11-17 19:29:06		2877	5102090	<blockquote> <p>Mod note...
1200621	1	1200646		2009-07-29 14:22:27		2277	5013476	<p>How do I declare and initialize an array in...

At the bottom, it says "50000 rows returned in 35072 ms". Below the table is a search bar and a taskbar.



CA675 Cloud Technologies Assignment 1

A screenshot of a Microsoft Edge browser window. The title bar shows "Assignment 1 - Data Analysis.pdf" and "Query Stack Overflow - Stack Exchange". The address bar shows "data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets". The main content area displays a table titled "Results" with columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, and Body. The table contains 50,000 rows, as indicated by the status bar at the bottom which says "50000 rows returned in 35072 ms".

Id	PostTypeId	AcceptedAnswerId	ParentId	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...

A screenshot of a Microsoft Edge browser window. The title bar shows "Assignment 1 - Data Analysis.pdf" and "Query Stack Overflow - Stack Exchange". The address bar shows "data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets". The main content area has a "edit description" section and a "Q&A for professional and enthusiast programmers" sidebar. Below is a code editor with the following SQL query:

```
1 select top 50000 * from posts where posts.ViewCount <= 127078 order by posts.ViewCount DESC
```

Below the code editor are buttons for "Run Query" and "Cancel", and checkboxes for "Text-only results" and "Include execution plan". There is also a "permalink" link and a "switch to meta site" link. At the bottom, there is a search bar and a table titled "Results" with columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, and Body. The table contains 2 rows, matching the results shown in the first screenshot.

Id	PostTypeId	AcceptedAnswerId	ParentId	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...
2794317	1	2794366		2010-05-08 12:59:11		58	127078	<p>I'm writing the JS for a chat application I...



CA675 Cloud Technologies Assignment 1

A screenshot of a Microsoft Edge browser window. The address bar shows 'data.stackexchange.com/stackoverflow/query/edit/1487283#resultSets'. The main content area displays a table with 20 rows of data from a Stack Overflow query. The columns are: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, and Body. The 'Body' column contains snippets of code or text. A message at the bottom right of the table says '50000 rows returned in 114101 ms'.

ID	PostTypeID	AcceptedAnswerID	ParentID	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...
4150782	1			2010-11-11 01:50:37		91	127077	<p>Are variables within YAML files possible? ...
43191294	1	43295074		2017-04-03 17:57:26		38	127068	<p>I really need help. I searched in all the que...
13054451	1	13089688		2012-10-24 17:24:39		48	127067	<p>I have a problem with this <code>CMake...
6466031	1	6466060		2011-06-24 09:39:37		88	127066	<p>I'm new to ajax and callback functions, pl...
19501636	1	25197251		2013-10-21 18:07:45		13	127065	<p>I want to include a small horizontal space ...
5438567	1	5438653		2011-03-25 21:58:26		251	127063	<p>I noticed that if I style my buttons with CS...
2193307	1	2193427		2010-02-03 16:02:57		113	127062	<p>In a LaTeX document I'm writing, I get an ...
4192277	1			2010-11-16 08:30:08		77	127059	<p>I'm having an issue where horizontal scro...
39173992	1	39174024		2016-08-26 20:09:10		61	127048	<p>I would like to drop all data in a pandas d...
9353822	1	9353830		2012-02-19 22:17:44		69	127045	<p>I know this might be really a simple questi...
7628311	1	7628347		2011-10-02 18:22:32		11	127044	<p>I have list of strings</p> <pre><code>a = ...
6017987	1	6018043		2011-05-16 13:15:27		346	127042	<p>I know Git stores information of when files...
14711956	1	14711978		2013-02-05 16:17:27		43	127039	<p>The typical way of creating a Javascript o...
12188509	1	12188551		2012-08-30 00:58:30		109	127033	<p>In R, I have an operation which creates s...

50000 rows returned in 114101 ms



A screenshot of a Microsoft Edge browser window, identical to the one above but with a different set of 20 rows of data from a Stack Overflow query. The columns and structure are the same, showing IDs, Post Types, Accepted Answers, Parents, Creation Dates, Deletion Dates, Scores, View Counts, and Bodies. A message at the bottom right of the table says '50000 rows returned in 114101 ms'.

ID	PostTypeID	AcceptedAnswerID	ParentID	CreationDate	DeletionDate	Score	ViewCount	Body
30987143	1	49316925		2015-06-17 09:11:41		62	74427	<p>When I deploy Apache Mesos on Ubuntu...
46832394	1			2017-10-19 14:31:26		29	74427	<p>I want to access the first 100 rows of a sp...
11312525	1	35134329		2012-07-03 13:56:48		86	74426	<p>How do I catch a <kbd>Ctrl</kbd>+<kbd>...
7918571	1	7918720		2011-10-27 15:44:51		49	74425	<p>I have read really a lot of posts about this ...
25902288	1	25916838		2014-09-18 00:13:11		134	74425	<p>After updating to Xcode 6.1 beta 2 when I...
8816212	1	8819742		2012-01-11 08:23:27		34	74424	<p>I'm trying to implement a composite comp...
8323760	1	8324127		2011-11-30 09:35:51		19	74424	<p>I've little knowledge of Java. I need to con...
10254180	1	10254238		2012-04-20 22:04:21		142	74424	<p>PHPUnit contains an <a href="https://ph...
42839074	1			2017-03-18 15:59:02		59	74424	<p>Hi I am trying to extract the id part of the ...
18570424	1	18574593		2013-09-02 09:59:08		36	74424	<p>I'm trying to use FontAwesome in a web ...
4332982	1	4333037		2010-12-02 08:17:01		86	74423	<p>What happens if a clustered index is not ...
45777232	1	45777277		2017-08-19 23:45:03		27	74423	<p>In AngularJS is possible to style tooltips i...
3846847	1			2010-10-02 17:44:09		8	74422	<p><pre><code><i...
602937	1			2009-03-02 16:21:22		8	74422	<p>I am trying to execute my first "Hello...
10651349	1	10651376		2012-05-18 10:57:09		54	74421	<p>How can I check for empty values of (<co...
6707657	1	6707677		2011-07-15 13:24:14		38	74420	<p>Is there any universal method to detect st...

50000 rows returned in 114101 ms





CA675 Cloud Technologies Assignment 1

Screenshot of a web browser showing a query editor on Stack Overflow. The query is:

```
1 select top 50000 * from posts where posts.ViewCount <= 74420 order by posts.ViewCount DESC
```

The results table has columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, Body. The first row is:

Id	PostTypeId	AcceptedAnswerId	ParentId	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...

Below the table, it says "50000 rows returned in 115777 ms".



CA675 Cloud Technologies Assignment 1

Screenshot of a web browser showing a query results page from Stack Overflow. The results table has columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, and Body. The body column contains truncated XML snippets. A note at the bottom says "50000 rows returned in 115777 ms".

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



Screenshot of a web browser showing a query results page from Stack Overflow. The results table has columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, and Body. The body column contains truncated text. A note at the bottom says "50000 rows returned in 115777 ms".

Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount	Body
24671249	1			2014-07-10 07:57:30		43	53077	<p>I'm trying to parse a json but I have some...



CA675 Cloud Technologies Assignment 1

Screenshot of a web browser showing a table of results from Stack Overflow. The table has columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, and Body. The Body column contains truncated text snippets. A 'Download CSV' button is at the top right. The status bar at the bottom says '50000 rows returned in 111716 ms'.

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



Screenshot of a web browser showing a table of results from Stack Overflow. The table has columns: Id, PostTypeId, AcceptedAnswerId, ParentId, CreationDate, DeletionDate, Score, ViewCount, and Body. The Body column contains truncated text snippets. A 'Download CSV' button is at the top right. The status bar at the bottom says '50000 rows returned in 111716 ms'.

Id	PostTypeId	AcceptedAnswerId	ParentId	CreationDate	DeletionDate	Score	ViewCount	Body
14370554	1		14400242	2013-01-17 00:57:59		14	41237	<p>I have a <code>TabPane</code> with se...
5513499	1			2011-04-01 12:41:14		20	41237	<p>I have a .net app that I just opened on in ...
36850485	1		36873474	2016-04-25 20:30:56		30	41237	<p>It looks like the GitHubPullRequestBuilder...
5364270	1		5364845	2011-03-19 19:12:38		4	41237	<p>Can someone please explain the function...
30332165	1			2015-05-19 17:21:38		10	41236	<p>I have a samsung note pro 12.2 and was ...
93999365	1		9399907	2012-02-22 16:59:04		43	41235	<p>I am struggling with deep copies of object...
44698296	1		45152845	2017-06-22 11:36:01		38	41235	<p><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	<p>Is it possible to read and write csv files us...
3853749	1			2010-10-04 08:44:47		88	41235	<p>What is the difference between a MVC M...
8058793	1			2011-11-09 00:04:44		41	41235	<p>I seem to remember seeing a single line i...
11806570	1		11815229	2012-08-04 06:46:17		34	41235	<p>In EF projects, Is there any best practice f...
5865353	1			2011-05-03 05:19:22		34	41235	<p>What is <code>FacesContext</code> us...
15258594	1		15259359	2013-03-06 21:38:46		5	41234	<p>I'm trying to start project with Hibernate a...
16107965	1		16107998	2013-04-19 15:18:43		1	41234	<p>My code looks like below,</p> <pre><cod...
38522931	1		38523848	2016-07-22 09:29:16		4	41234	<p>I am trying to read data from Excel sheets...





CA675 Cloud Technologies Assignment 1

A screenshot of a Windows desktop showing a web browser window. The browser has two tabs: "Assignment 1 - Data Analysis.pdf" and "Query Stack Overflow - Stack Exchange". The main content area shows a SQL query in a code editor:

```
1 select count(*) from posts where posts.ViewCount <= 41234 and posts.ViewCount > 41231
```

Below the code editor are buttons for "Run Query", "Cancel", and "Options: Text-only results Include execution plan". There is also a "Results" tab and a "Messages" tab. The status bar at the bottom shows the date and time: 16/10/2021, 23:05.

A screenshot of a Windows desktop showing a web browser window. The browser has two tabs: "Assignment 1 - Data Analysis.pdf" and "Query Stack Overflow - Stack Exchange". The main content area shows a SQL query in a code editor:

```
1 select top 50000 * from posts where posts.ViewCount <= 41234 and posts.ViewCount > 41231 order by posts.ViewCount DESC
```

Below the code editor are buttons for "Run Query", "Cancel", and "Options: Text-only results Include execution plan". There is also a "Results" tab and a "Messages" tab. The status bar at the bottom shows the date and time: 16/10/2021, 23:07.



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>

```
colin_hehir@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...
Connected, host fingerprint: ssh-rsa 0 82:5A:1B:F2:56:FI:10:C5:16:2A:57:EF:31:BD
+0:9:1:BB:2:A3:C:E1:58:E1:17:C6:B7:33:0:5:E9:0:0:9B:CL
Linux ch-ca675-assignment-1-m 5.10.0-0.bpo.8-amd64 #1 SMP Debian 5.10.46-4-bpo1
+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.
colin_hehir@ch-ca675-assignment-1-m:~$ whoami
colin_hehir
colin_hehir@ch-ca675-assignment-1-m:~$ hadoop fs -mkdir -p /user/colin_hehir
colin_hehir@ch-ca675-assignment-1-m:~$ java -version
openjdk version "1.8.0_292"
OpenJDK Runtime Environment (AdoptOpenJDK) (build 1.8.0_292-b10)
OpenJDK 64-Bit Server VM (AdoptOpenJDK) (build 25.292-b10, mixed mode)
colin_hehir@ch-ca675-assignment-1-m:~$ export PATH=$JAVA_HOME/bin:$PATH
colin_hehir@ch-ca675-assignment-1-m:~$ export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
colin_hehir@ch-ca675-assignment-1-m:~$ hadoop fs -ls
colin_hehir@ch-ca675-assignment-1-m:~$ en
-bash: en: command not found
colin_hehir@ch-ca675-assignment-1-m:~$ env
SHELL=/bin/bash
DATAPROC_STARTUP_SCRIPT=/usr/local/share/google/dataproc/startup-script.sh
DATAPROC_MASTER_SERVICES=hadoop-hdfs-namenode hive-metastore hive-server2 solr-server hadoop-yarn-resourcemanager
DATAPROC_DIR=/usr/local/share/google/dataproc
CONDA_EXE=/opt/conda/miniconda3/bin/conda
_CE_M=
DATAPROC_VERSION=2.0
SPARK_LOG_DIR=/var/log/spark
DATAPROC_IMAGE_BUILD=20210917-180200-RC01-2_0_deb10_20210908_132200-RC01
JAVA_HOME=/usr/lib/jvm/adoptopenjdk-8-hotspot-amd64
SSH_AUTH_SOCK=/tmp/ssh-jk4qak58IR/agent.7466
CONDA_HOME=/opt/conda/default
PWD=/home/colin_hehir
LOGNAME=colin_hehir
XDG_SESSION_TYPE=tty
HADOOP_CLASSPATH=/usr/lib/jvm/adoptopenjdk-8-hotspot-amd64/lib/tools.jar
DATAPROC_IMAGE_TYPE=standard
```

The screenshot shows the Google Cloud Platform Compute Engine interface. On the left, there's a sidebar with 'Compute Engine' selected under 'Virtual machines'. The main area displays a table of VM instances:

Status	Name	Zone	Recomm.	Connect
✓	ch-ca675-assignment-1-m	europe-west2-b		SSH
✗	ch-ca675-assignment-1-w-0	europe-west2-b		SSH
✗	ch-ca675-assignment-1-w-1	europe-west2-b		SSH
✗	ch-ca675-assignment-1-w-2	europe-west2-b		SSH

To the right of the table, a detailed view for the instance 'ch-ca675-assignment-1-m' is shown. It includes sections for 'PERMISSIONS', 'LABELS', and 'MONITORING'. Under 'PERMISSIONS', there's a list of principals and roles:

- Dataproc Service Agent (1)
- Editor (2)
- Owner (1)

A blue button labeled 'ADD PRINCIPAL' is visible. Below the permissions, there's a section for 'Role/Principal' inheritance.



CA675 Cloud Technologies Assignment 1

The screenshot shows the Google Cloud Platform Cloud Storage interface. The left sidebar has 'Cloud Storage' selected. The main area shows a bucket named 'dataproc-staging-europe-west2-11221609810-kalheazk'. It lists objects in the 'OBJECTS' tab, including several Excel files ('Stack_Exchange_Query_Results_1.xlsx' through 'Stack_Exchange_Query_Results_5.xlsx') and a folder 'google-cloud-dataproc-metainfo/'. The table includes columns for Name, Size, Type, Created, Storage class, Last modified, and Public access.

Name	Size	Type	Created	Storage class	Last modified	Public access
Stack_Exchange_Query_Results_1.xlsx	52.3 MB	application/vnd.ms-excel	22 Oct 20...	Standard	22 Oct 202...	Not public
Stack_Exchange_Query_Results_2.xlsx	59.4 MB	application/vnd.ms-excel	22 Oct 20...	Standard	22 Oct 202...	Not public
Stack_Exchange_Query_Results_3.xlsx	62.3 MB	application/vnd.ms-excel	22 Oct 20...	Standard	22 Oct 202...	Not public
Stack_Exchange_Query_Results_4.xlsx	65.5 MB	application/vnd.ms-excel	22 Oct 20...	Standard	22 Oct 202...	Not public
Stack_Exchange_Query_Results_5.xlsx	16.7 KB	application/vnd.ms-excel	22 Oct 20...	Standard	22 Oct 202...	Not public
google-cloud-dataproc-metainfo/	—	Folder	—	—	—	—

The screenshot shows an SSH session on a terminal window. The command entered is '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=true'. A 'File Transfer' dialog box is open, showing the message 'Stack_Exchange_Query_Results_1.... Finished'. Below the terminal, it says '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=true
colin_hehir3@ch-ca675-assignment-1-m:~$ 
```

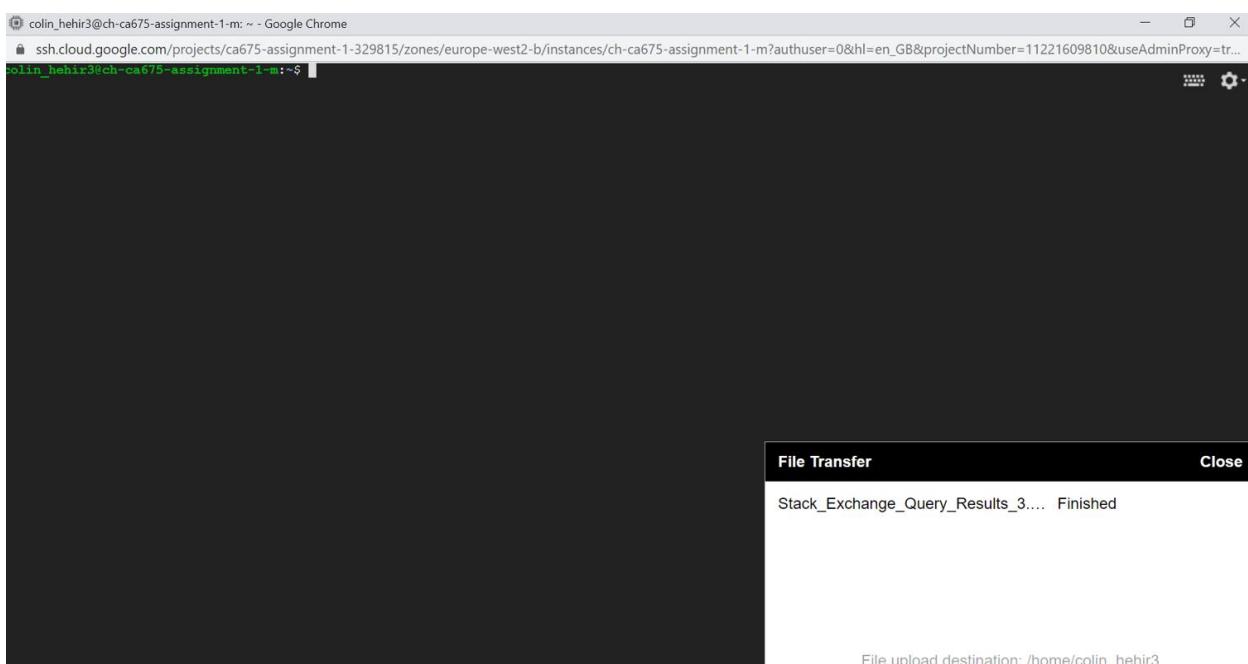
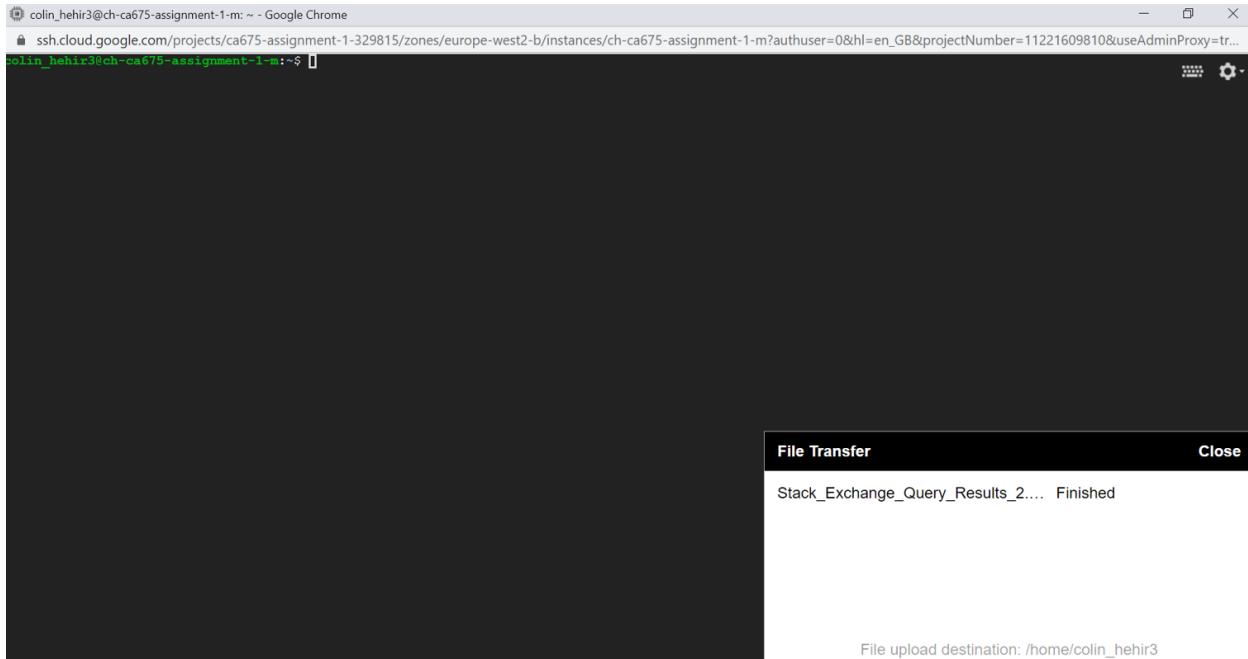
File Transfer

Stack_Exchange_Query_Results_1.... Finished

File upload destination: /home/colin_hehir3

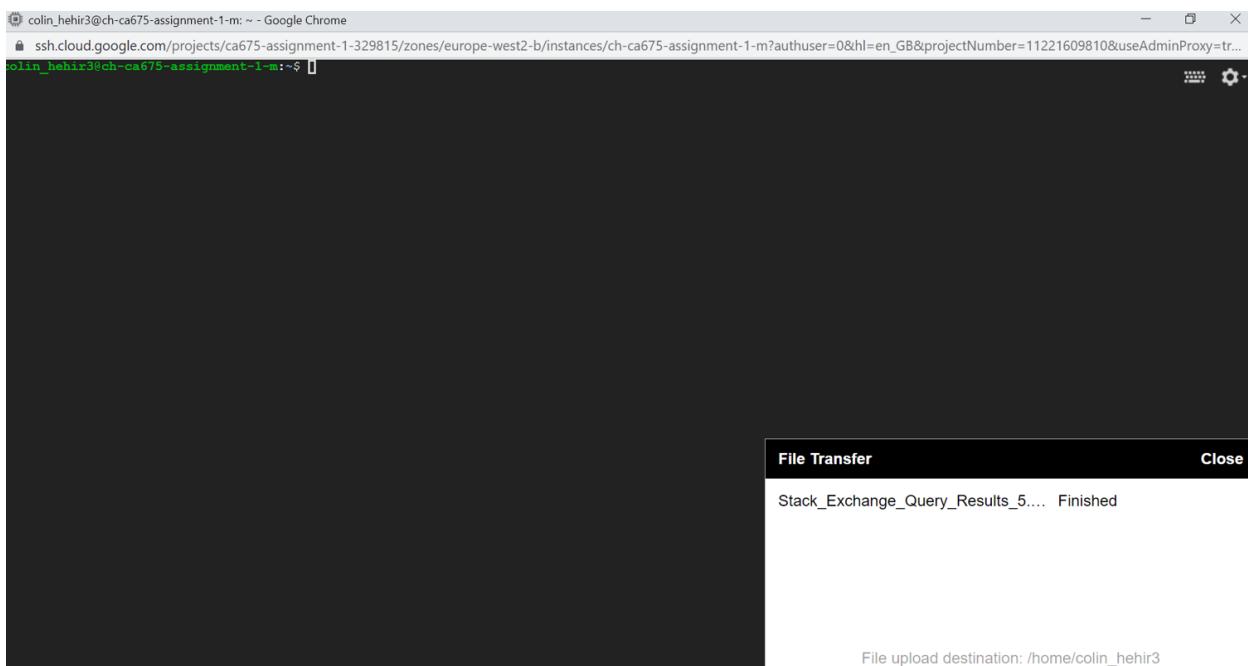
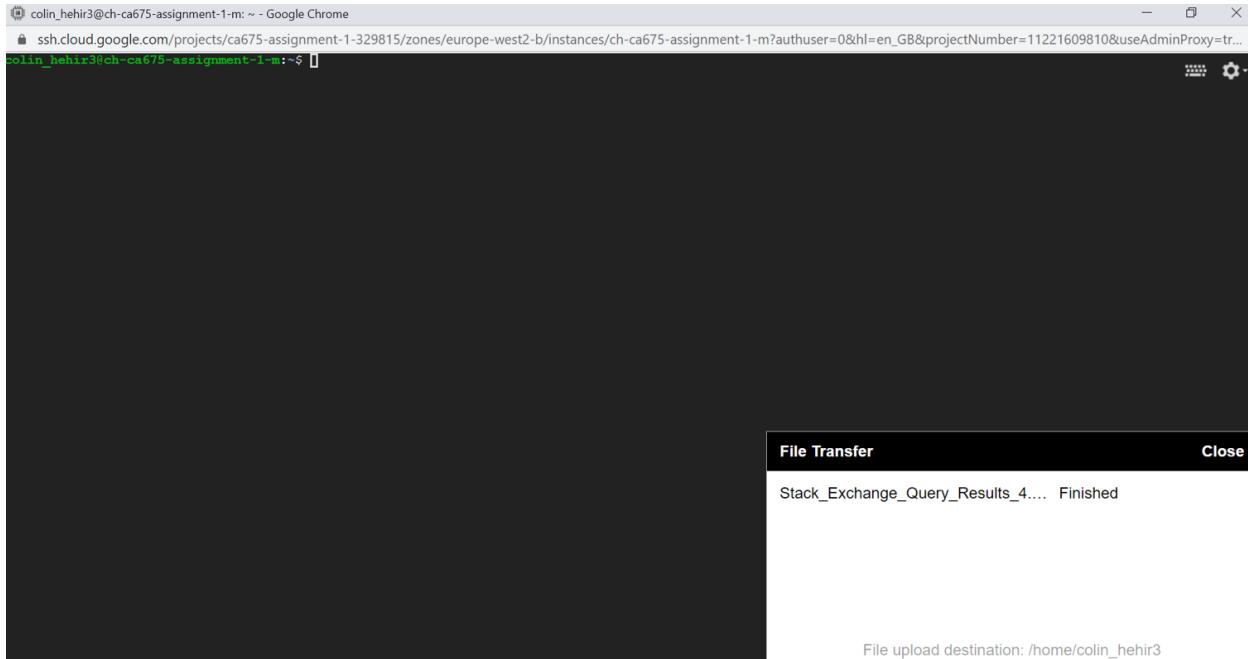


CA675 Cloud Technologies Assignment 1





CA675 Cloud Technologies Assignment 1





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>

```
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=true
colin_hehir3@ch-ca675-assignment-1-m:~$ ls
Stack_Exchange_Query_Results_1.csv Stack_Exchange_Query_Results_3.csv Stack_Exchange_Query_Results_5.csv
Stack_Exchange_Query_Results_2.csv Stack_Exchange_Query_Results_4.csv
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -put /home/colin_hehir3/Stack_Ex
change_Query_Results_1.csv /Data_Acquisition_Collection
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -put /home/colin_hehir3/Stack_Ex
-bash: $: command not found
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -put /home/colin_hehir3/Stack_Ex
put: '/home/colin_hehir3/Stack_Ex': No such file or directory
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -put /home/colin_hehir3/Stack_Exchange_Query_Results_2.csv /Data_Acquisition_Collection
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -put /home/colin_hehir3/Stack_Exchange_Query_Results_3.csv /Data_Acquisition_Collection
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -put /home/colin_hehir3/Stack_Exchange_Query_Results_4.csv /Data_Acquisition_Collection
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -put /home/colin_hehir3/Stack_Exchange_Query_Results_5.csv /Data_Acquisition_Collection
colin_hehir3@ch-ca675-assignment-1-m:~$
```

```
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=true
colin_hehir3@ch-ca675-assignment-1-m:~$ hadoop fs -ls /Data_Acquisition_Collection
Found 5 items
-rw-r--r--  2 colin_hehir3 hadoop  54837285 2021-10-22 21:53 /Data_Acquisition_Collection/Stack_Exchange_Query_Results_1.csv
-rw-r--r--  2 colin_hehir3 hadoop  62270983 2021-10-22 21:56 /Data_Acquisition_Collection/Stack_Exchange_Query_Results_2.csv
-rw-r--r--  2 colin_hehir3 hadoop  65368755 2021-10-22 21:57 /Data_Acquisition_Collection/Stack_Exchange_Query_Results_3.csv
-rw-r--r--  2 colin_hehir3 hadoop  68685253 2021-10-22 21:58 /Data_Acquisition_Collection/Stack_Exchange_Query_Results_4.csv
-rw-r--r--  2 colin_hehir3 hadoop   17085 2021-10-22 21:58 /Data_Acquisition_Collection/Stack_Exchange_Query_Results_5.csv
colin_hehir3@ch-ca675-assignment-1-m:~$
```



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>

```
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=true
Connected, host fingerprint: ssh-rsa 0:23:5a:1:b5:f2:5e:f1:10:0:5:16:2:a:57:ef:3:1:bd
:69:91:0:8:2a:c3:e1:59:3:17:0:6:b7:3:3:c5:85:9:0:98:c1
Linux ch-ca675-assignment-1-m 5.10.0-8-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: Sat Oct 23 17:41:10 2021 from 35.235.244.114
colin_hehir3@ch-ca675-assignment-1-m: ~
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2021-10-23 17:55:55,495 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-10-23 17:55:55,496 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2021-10-23 17:55:55,496 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-23 17:55:55,528 [main] INFO org.apache.pig.Main - Apache Pig version 0.18.0-SNAPSHOT (r: unknown) compiled Dec 21 1969, 06:05:27
2021-10-23 17:55:55,528 [main] INFO org.apache.pig.Main - Logging error messages to: /home/colin_hehir3/pig_163501175523.log
2021-10-23 17:55:55,537 [main] INFO org.apache.pig.impl.Utils - Default bootup file /home/colin_hehir3/pigbootup not found
2021-10-23 17:55:55,749 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-10-23 17:55:55,749 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://ch-ca675-assignmen
+1-m
2021-10-23 17:55:56,333 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-6b365653-24f1-46b6-a178-98d36d2c923
2021-10-23 17:55:56,457 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: ch-ca675-assignment-1-m:8188
2021-10-23 17:55:56,664 [main] INFO org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2021-10-23 17:55:56,681 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead
, use yarn.system-metrics-publisher.enabled
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));
2021-10-23 17:56:05,758 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThre
hold = 489580128, usageThreshold = 489580128
grunt> StackExchangeData_Cleaning_1 = FOREACH StackExchangeData_Total_Distinct_ID GENERATE Id,Score,ViewCount, REPLACE(Body, '\n', '') as Body,OwnerUserId,OwnerDisp
layName,REPLACE>Title, '\n', '' as Title;
grunt>
```

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>

```
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=true...
grunt>
grunt> StackExchangeData_Cleaning_2 = FOREACH StackExchangeData_Cleaning_1 GENERATE Id,Score,ViewCount, REPLACE(Body,'</p>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '</p>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_3 = FOREACH StackExchangeData_Cleaning_2 GENERATE Id,Score,ViewCount, REPLACE(Body,'<p>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<p>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_4 = FOREACH StackExchangeData_Cleaning_3 GENERATE Id,Score,ViewCount, REPLACE(Body,'</code>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '</code>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_5 = FOREACH StackExchangeData_Cleaning_4 GENERATE Id,Score,ViewCount, REPLACE(Body,'<code>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<code>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_6 = FOREACH StackExchangeData_Cleaning_5 GENERATE Id,Score,ViewCount, REPLACE(Body,'<pre>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<pre>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_7 = FOREACH StackExchangeData_Cleaning_6 GENERATE Id,Score,ViewCount, REPLACE(Body,'<pre>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<pre>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_8 = FOREACH StackExchangeData_Cleaning_7 GENERATE Id,Score,ViewCount, REPLACE(Body,'<blockquote>', '') as Body,OwnerUserId,OwnerDis
playName,REPLACE>Title, '<blockquote>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_9 = FOREACH StackExchangeData_Cleaning_8 GENERATE Id,Score,ViewCount, REPLACE(Body,'<code>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<code>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_10 = FOREACH StackExchangeData_Cleaning_9 GENERATE Id,Score,ViewCount, REPLACE(Body,'<code>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<code>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_11 = FOREACH StackExchangeData_Cleaning_10 GENERATE Id,Score,ViewCount, REPLACE(Body,'<r>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<r>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_12 = FOREACH StackExchangeData_Cleaning_11 GENERATE Id,Score,ViewCount, REPLACE(Body,'<t>', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '<t>', '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_13 = FOREACH StackExchangeData_Cleaning_12 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\,,', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '\\,,' '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_14 = FOREACH StackExchangeData_Cleaning_13 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\\\,,', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '\\\\,,' '' as Title;
grunt>
grunt> StackExchangeData_Cleaning_15 = FOREACH StackExchangeData_Cleaning_14 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\\\,,', '') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title, '\\\\\\,,' '' as Title;
```



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>

```
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=true
JobId Maps Reduces MaxMapTime MinMapTime AvgMapTime MedianMapTime MaxReduceTime MinReduceTime AvgReduceTime MedianReduceTime
Feature Outputs
Job 1634991989209_0015 2 1 30 26 28 28 43 43 43 StackExchangeData Cleaning 1,StackExchangeData Cleaning 10,StackExchangeData Cleaning 11,StackExchangeData Cleaning 12,StackExchangeData Cleaning 13,StackExchangeData Cleaning 14,StackExchangeData Cleaning 15,StackExchangeData Cleaning 16,StackExchangeData Cleaning 17,StackExchangeData Cleaning 18,StackExchangeData Cleaning 19,StackExchangeData Cleaning 2,StackExchangeData Cleaning 20,StackExchangeData Cleaning 21,StackExchangeData Cleaning 3,StackExchangeData Cleaning 4,StackExchangeData Cleaning 5,StackExchangeData Cleaning 6,StackExchangeData Cleaning 7,StackExchangeData Cleaning 8,StackExchangeData Cleaning 9,StackExchangeData Total,StackExchangeData Total Column_Filter,StackExchangeData Total Distinct_D,StackExchangeData Total_Grouped GROUP_BY,COMBINER /Final_Dataset

Input(s):
Successfully read 200014 records (251180543 bytes) from: "/Data_Acquisition_Collection/*.csv"

Output(s):
Successfully stored 200003 records (187913709 bytes) in: "/Final_Dataset"

Counters:
Total records written : 200003
Total bytes written : 187913709
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1634991989209_0015

2021-10-23 18:35:26,110 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-23 18:35:26,110 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-23 18:35:26,111 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-23 18:35:26,126 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-23 18:35:26,126 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-23 18:35:26,127 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-23 18:35:26,138 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ch-ca675-assignment-1-m/10.154.0.11:8032
2021-10-23 18:35:26,139 [main] INFO org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:10200
2021-10-23 18:35:26,140 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-10-23 18:35:26,152 [main] WARN org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION FAILED 12 time(s).
2021-10-23 18:35:26,152 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
runtrnct: ■
```

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>

```
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=true
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: Sat Oct 23 19:16:28 2021 from 35.235.242.1
colin_hehir3@ch-ca675-assignment-1-m:~$ hive
Hive Session ID = cb9bf2f49-79aa-4608-9e0a-bfa29481585a

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Hive Session ID = c743c2f7-b568-40b8-8476-0294d8bb5ade
hive> CREATE TABLE FINALDATASET (Id int, Score int, ViewCount int, Body string, OwnerUserId int, OwnerDisplayName string, Title string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' TBLPROPERTIES('skip.header.line.count'=1);
OK
Time taken: 1.72 seconds
hive> LOAD DATA INPUTPATH '/Final_Dataset/part-r-00000' INTO TABLE FINALDATASET;
Loading data to table default.finaledataset
OK
Time taken: 0.333 seconds
hive> SELECT COUNT(*) FROM FINALDATASET
> 1
Query ID = colin_hehir3_20211023194243_3d6c190f-5219-4db3-832f-5dd0001accb8
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1634991989209_0018)

-----  

      VERTICES    MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container    SUCCEEDED     1      1      0      0      0      0  

Reducer 2 ..... container    SUCCEEDED     1      1      0      0      0      0  

-----  

VERTICES: 02/02  ==> 100% ELAPSED TIME: 6.72 s  

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



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=true
ossible as possible. I would also like it so there aren't any extra characters or parameters in the URL. How can I achieve this? Based on the answers posted so far I am currently doing this: <input type="submit" value="Continue"/> But this adds a question mark character to the end of the URL. I need to find a solution that doesn't add any characters to the end of the URL. There are two other solutions to do this: Using JavaScript or styling a link to look like a button. Using JavaScript <a href="#" onclick="window.location.href='/page2'; Continue</a>; But this obviously requires JavaScript and for that reason it is less accessible to screen readers. The point of a link is to go to another page. So trying to make a button act like a link is the wrong solution. My suggestion is that you should use a link and style it to look like a button. <a href="/link/to/page2"; Continue</a> 48523 How to create an HTML button that acts like a link
Time taken: 9.815 seconds, Fetched: 5 row(s)
hive> set hive.cli.print.header=true;
hive> select Id, Score, Title from DATASET sort by Score desc limit 10;
Query ID = colin_hehir3_20211023202758_9634638a-b5e0-4e3a-ba42-0852eal3c736
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1634991989209_0022)

-----  
 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: 11.63 s |  |  |  |  |  |  |  |  |
| OK |  |  |  |  |  |  |  |  |
| id | score | title |  |  |  |  |  |  |
| 11227809 | 25893 | Why is processing a sorted array faster than processing an unsorted array? |  |  |  |  |  |  |
| 927358 | 23274 | How do I undo the most recent local commits in Git? |  |  |  |  |  |  |
| 2003505 | 18451 | How do I delete a Git branch locally and remotely? |  |  |  |  |  |  |
| 929357 | 12796 | What is the difference between 'git pull' and 'git fetch'? |  |  |  |  |  |  |
| 231767 | 11512 | What does the yield keyword do? |  |  |  |  |  |  |
| 477816 | 10894 | What is the correct JSON content type? |  |  |  |  |  |  |
| 348170 | 10045 | How do I undo 'git add' before commit? |  |  |  |  |  |  |
| 5767325 | 9877 | How can I remove a specific item from an array? |  |  |  |  |  |  |
| 6591213 | 9747 | How do I rename a local Git branch? |  |  |  |  |  |  |
| 1642028 | 9539 | What is the -- operator in C/C++? |  |  |  |  |  |  |
| Time taken: 18.117 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=true
6591213 9747 How do I rename a local Git branch?
1642028 9539 What is the -- operator in C/C++
Time taken: 18.117 seconds, Fetched: 10 row(s)
hive> Select owneruserid, ownerdisplayname,
    > SUM(Score) AS Aggregate_score from F
    > GROUP BY owneruserid, ownerdisplayname
    > ;
FAILED: SemanticException [Error 10001]: Line 2:35 Table not found 'F'
hive> select OwnerUserId, SUM(Score) as TotalUserScore from DATASET group by OwnerUserId orderby TotalUserScore desc limit 10;
FAILED: ParseException line 1:83 missing EOF at 'orderby' near 'OwnerUserId'
hive> select OwnerUserId, SUM(Score) as TotalUserScore from DATASET group by OwnerUserId order by TotalUserScore desc limit 10;
Query ID = colin_hehir3_20211023203918_a4f15235-94a5-4534-8fc8-e5cedebf1f42
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
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: 10.52 s |  |  |  |  |  |  |  |  |
| OK |  |  |  |  |  |  |  |  |
| owneruserid | totaluserscore |  |  |  |  |  |  |  |
| NULL | 1033287 |  |  |  |  |  |  |  |
| 87234 | 37606 |  |  |  |  |  |  |  |
| 4883 | 28155 |  |  |  |  |  |  |  |
| 9951 | 26728 |  |  |  |  |  |  |  |
| 6068 | 25860 |  |  |  |  |  |  |  |
| 99904 | 23949 |  |  |  |  |  |  |  |
| 51816 | 23632 |  |  |  |  |  |  |  |
| 179736 | 19415 |  |  |  |  |  |  |  |
| 95592 | 19413 |  |  |  |  |  |  |  |
| 63051 | 18738 |  |  |  |  |  |  |  |
| Time taken: 17.219 seconds, Fetched: 10 row(s) |  |  |  |  |  |  |  |  |
| hive> select OwnerUserId, SUM(Score) as TotalUserScore from DATASET where OwnerUserId is not null group by OwnerUserId order by TotalUserScore desc limit 10; |  |  |  |  |  |  |  |  |

```



CA675 Cloud Technologies Assignment 1

```
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=true
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: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_2.Vertex 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]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]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]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
99904      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=true...
Connected, host fingerprint: 8e0-7e:0 0:2:5A:1B:52:66:FI:10:C5:16:2A:57:EF:31:BD
0:91:0:1B:2A:13:FI:8:f5:f1:17:D6:B7:33:5:E9:f1:0:9B: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 = 779ef11-a912-477f-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>
```



CA675 Cloud Technologies Assignment 1

```
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=true
Connected, host fingerprint: ssh-rsa 0 02:5a:1b:f2:56:f1:01c5:16:2a:57:ef:31:bd
:69:91:0b:2a:c3:e1:58:d3:17:c6:b7:33:c5:e9: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 19:32:16 2021 from 35.235.244.114
colin_hehir3@ch-ca675-assignment-1-m:~$ hive
Hive Session ID = c8822576-972b-4753-92b7-7943871bc7ee

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Hive Session ID = a27laab3-d798-4550-b719-9860d67db0ae
hive> set hive.cli.print.header=true;
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(Title) like LOWER('% cloud %') or LOWER>Title) like LOWER('% cloud %') or LOWER>Title) like LOWER(' cloud %') or LOWER>Title) like LOWER('cloud %');
Query ID = colin_hehir3_20211028195025_1beaae96-0cd8-4bff-97ff-eedcc6a4a94ca
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1635449054866_0002)

-----  

VERTICES      MODE      STATUS    TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED  

Map 1 ..... container  SUCCEEDED   1      1        0        0        0        0        0  

Reducer 2 ..... container  SUCCEEDED   8      8        0        0        0        0        0  

Reducer 3 ..... container  SUCCEEDED   1      1        0        0        0        0        0  

-----  

VERTICES: 03/03 [=] ----->1 100% ELAPSED TIME: 19.87 s  

OK
_c0
290
Time taken: 24.048 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=true...
colin_hehir3@ch-ca675-assignment-1-m:~$ pig
WARNING: HADOOP_HOME has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2021-10-28 20:21:20,079 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-10-28 20:21:20,080 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2021-10-28 20:21:20,080 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-28 20:21:20,115 [main] INFO org.apache.pig.Main - Apache Pig version 0.18.0-SNAPSHOT (r: unknown) compiled Dec 21 1969, 06:05:27
2021-10-28 20:21:20,115 [main] INFO org.apache.pig.Main - Logging error messages to: /home/colin_hehir3/pig_1635452480112.log
2021-10-28 20:21:20,136 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/colin_hehir3/.pigbootup not found
2021-10-28 20:21:20,344 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-10-28 20:21:20,344 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://ch-ca675-assignmen
t-1:m
2021-10-28 20:21:20,917 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-7eb4e8b4-4910-4d19-a6ee-1ddff5ff93cf1
2021-10-28 20:21:21,039 [main] INFO org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: ch-ca675-assignment-1-m:8188
2021-10-28 20:21:21,223 [main] INFO org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2021-10-28 20:21:21,240 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead
, use yarn.system-metrics-publisher.enabled
grunt> StackExchangeData Total_Task4 = LOAD '/Data_Aquisition_Collection/*.*' USING org.apache.pig.piggybank.storage.CSVExcelStorage('','YES MULTILINE','NOCHANG
E','SKIP_INPUT_HEADER') AS(Id:int, PostTypeId:int, AcceptedAnswerId:int, ParentId:int, CreationDate:datetime, DeletionDate:datetime, Score:int, ViewCount:int, Body
:chararray, OwnerUserId:int, OwnerDisplayName:chararray, LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:datetime, LastActivityDate:datetime, Tit
le:chararray, Tags:chararray, AnswerCount:int, CommentCount:int, FavoriteCount:int, ClosedDate:datetime, CommunityOwnedDate:datetime, ContentLicense:chararray);
2021-10-28 20:21:25,993 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead
, use yarn.system-metrics-publisher.enabled
grunt> StackExchangeData Total_Column_Filter = FOREACH StackExchangeData Total_Task4 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);
2021-10-28 20:22:20,378 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 699400192 to monitor. collectionUsageThre
shold = 489580128, usageThreshold = 489580128
grunt> StackExchangeData Cleaning_1 = FOREACH StackExchangeData_Total_Distinct_ID GENERATE Id,Score,ViewCount, REPLACE(Body,'<.*?>','') as Body,OwnerUserId,OwnerDi
splayName,REPLACE>Title,'<.*?>','') as Title;
grunt> StackExchangeData Cleaning_2 = FOREACH StackExchangeData_Cleaning_1 GENERATE Id,Score,ViewCount, REPLACE(Body,'\n',' ') as Body,OwnerUserId,OwnerDisplayName,
REPLACE>Title,'n','') as Title;
grunt> StackExchangeData Cleaning_3 = FOREACH StackExchangeData_Cleaning_2 GENERATE Id,Score,ViewCount, REPLACE(Body,'<*>','') as Body,OwnerUserId,OwnerDisplayName,
REPLACE>Title,'<>','') as Title;
grunt> StackExchangeData Cleaning_4 = FOREACH StackExchangeData_Cleaning_3 GENERATE Id,Score,ViewCount, REPLACE(Body,'</p>','') as Body,OwnerUserId,OwnerDisplayName
,REPLACE>Title,'<p>','') as Title;
grunt> StackExchangeData Cleaning_5 = FOREACH StackExchangeData_Cleaning_4 GENERATE Id,Score,ViewCount, REPLACE(Body,'<p>','') as Body,OwnerUserId,OwnerDisplayName
,REPLACE>Title,'<p>','') as Title;
grunt> StackExchangeData Cleaning_6 = FOREACH StackExchangeData_Cleaning_5 GENERATE Id,Score,ViewCount, REPLACE(Body,'</code>','') as Body,OwnerUserId,OwnerDisplayNa
me,REPLACE>Title,'<code>','') as Title;
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>
```



CA675 Cloud Technologies Assignment 1

```
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=true
grunt> StackExchangeData_Cleaning_7 = FOREACH StackExchangeData_Cleaning_6 GENERATE Id,Score,ViewCount, REPLACE(Body,<code>'') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,</code>'') as Title;
grunt> StackExchangeData_Cleaning_8 = FOREACH StackExchangeData_Cleaning_7 GENERATE Id,Score,ViewCount, REPLACE(Body,<pre>'') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,</pre>'') as Title;
grunt> StackExchangeData_Cleaning_9 = FOREACH StackExchangeData_Cleaning_8 GENERATE Id,Score,ViewCount, REPLACE(Body,</pre>'') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,</pre>'') as Title;
grunt> StackExchangeData_Cleaning_10 = FOREACH StackExchangeData_Cleaning_9 GENERATE Id,Score,ViewCount, REPLACE(Body,</blockquote>'') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,</blockquote>'') as Title;
grunt> StackExchangeData_Cleaning_11 = FOREACH StackExchangeData_Cleaning_10 GENERATE Id,Score,ViewCount, REPLACE(Body,</ul>'') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,</ul>'') as Title;
grunt> StackExchangeData_Cleaning_12 = FOREACH StackExchangeData_Cleaning_11 GENERATE Id,Score,ViewCount, REPLACE(Body,</li>'') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,</li>'') as Title;
grunt> StackExchangeData_Cleaning_13 = FOREACH StackExchangeData_Cleaning_12 GENERATE Id,Score,ViewCount, REPLACE(Body,'r') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<r>'') as Title;
grunt> StackExchangeData_Cleaning_14 = FOREACH StackExchangeData_Cleaning_13 GENERATE Id,Score,ViewCount, REPLACE(Body,'t') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<t>'') as Title;
grunt> StackExchangeData_Cleaning_15 = FOREACH StackExchangeData_Cleaning_14 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<\\>'') as Title;
grunt> StackExchangeData_Cleaning_16 = FOREACH StackExchangeData_Cleaning_15 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\\\') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<\\\\>'') as Title;
grunt> StackExchangeData_Cleaning_17 = FOREACH StackExchangeData_Cleaning_16 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\\\') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<\\\\>'') as Title;
grunt> StackExchangeData_Cleaning_18 = FOREACH StackExchangeData_Cleaning_17 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\\\') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<\\\\>'') as Title;
grunt> StackExchangeData_Cleaning_19 = FOREACH StackExchangeData_Cleaning_18 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\*p') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<\\*p>'') as Title;
grunt> StackExchangeData_Cleaning_20 = FOREACH StackExchangeData_Cleaning_19 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\*') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<\\*>'') as Title;
grunt> StackExchangeData_Cleaning_21 = FOREACH StackExchangeData_Cleaning_20 GENERATE Id,Score,ViewCount, REPLACE(Body,'\\>') as Body,OwnerId,OwnerDisplayName,REPLACE>Title,<\\>'') as Title;
grunt> StackExchangeData_Total_Task4_Cleaned = FILTER StackExchangeData_Cleaning_21 BY (OwnerId == 87234 or OwnerUserId == 4883 or OwnerUserId == 9951 or OwnerUserId == 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.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:~$
```





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-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_output
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 mapred.FileInputFormat: 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=true
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/streamjob6746467992186234996.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_output/part-0000* -output /user/colin_hehir3/ch_output
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%
```



CA675 Cloud Technologies Assignment 1

```
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=true

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/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/streaming-3.2.2.jar] /tmp/streamjob4877655264200457079.jar tmpDir=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/10.154.0.11:8020
2021-10-25 15:17:14,933 INFO client.AHSProxy: Connecting to Application History server at ch-ca675-assignment-1-m/10.154.0.11:8020
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=true

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)=443317888

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/streaming.jar -file /home/colin_hehir3/mapper4.py -mapper "python mapper4.py" -input /user/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/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/10.154.0.11:8020
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/10.154.0.11:8020
2021-10-25 15:19:12,359 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/colin_hehir3/.staging/job_1635169827887_0009
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.hehir@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=true
VERTICES: 04/04 [----->>>] 100% ELAPSED TIME: 7.63 s

OK
4983 0.022723 setautoresizemode 1
4983 0.018936 writeobject 2
4983 0.018936 accurate 2
4983 0.015149 gateway 4
4983 0.015149 seem 4
4983 0.015149 host 4
4983 0.015149 regex 4
4983 0.011361 traceback 8
4983 0.011361 includedigit 8
4983 0.011361 switch 8
4983 0.011361 die 8
4983 0.011361 development 8
4983 0.011361 errors 8
4915 0.052036 saledate 1
4915 0.052036 anomaly 1
4915 0.040813 less 3
4915 0.039027 nanotime 4
4915 0.039027 ssh 4
4915 0.039027 cool 4
4915 0.027208 jar 7
4915 0.026018 begin 8
4915 0.026018 extracting 8
4915 0.026018 load 8
4915 0.026018 printing 8
4915 0.026018 previously 8
4915 0.026018 projects 8
4915 0.026018 saleprice 8
9951 0.017622 myjson 1
9951 0.017622 design 1
9951 0.014685 explaining 3
9951 0.014685 understood 3
9951 0.014685 maximum 3
9951 0.014685 listview 3
9951 0.014685 correctly 3
9951 0.014685 optional 3
9951 0.011748 pseudocode 9
9951 0.011748 namearray 9
63051 0.019973 authentication 1
63051 0.017949 specific 2

63051 0.01712 closing 3
63051 0.014266 tutorial 4
63051 0.011413 document 5
63051 0.011413 usual 5
63051 0.011413 nosuchelementexceptionfrom 5
63051 0.011413 jarring 5
63051 0.011413 cp 5
63051 0.011413 staticfiles 5
87234 0.075743 apt 1
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 helloworldview 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
```



CA675 Cloud Technologies Assignment 1

```
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    delint 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
```