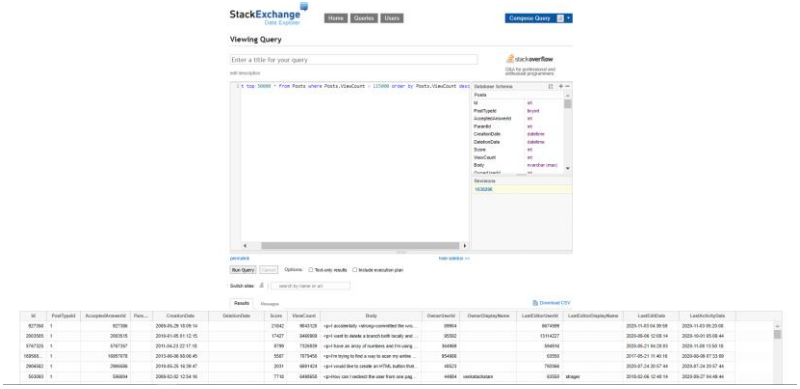
Assignment 1- Data Analysis

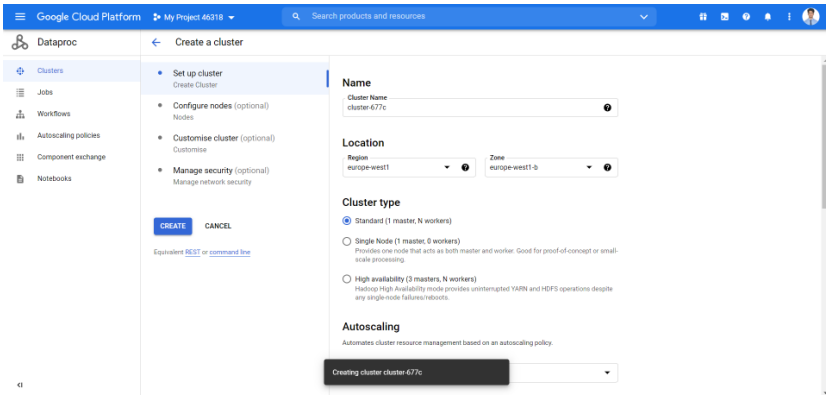
Srijith Unni
DCU Student ID: 20211114
MSc in Computing – Data Analytics
Cloud Technologies CA675

Documentation of steps undertaken for transformation and query of raw data:

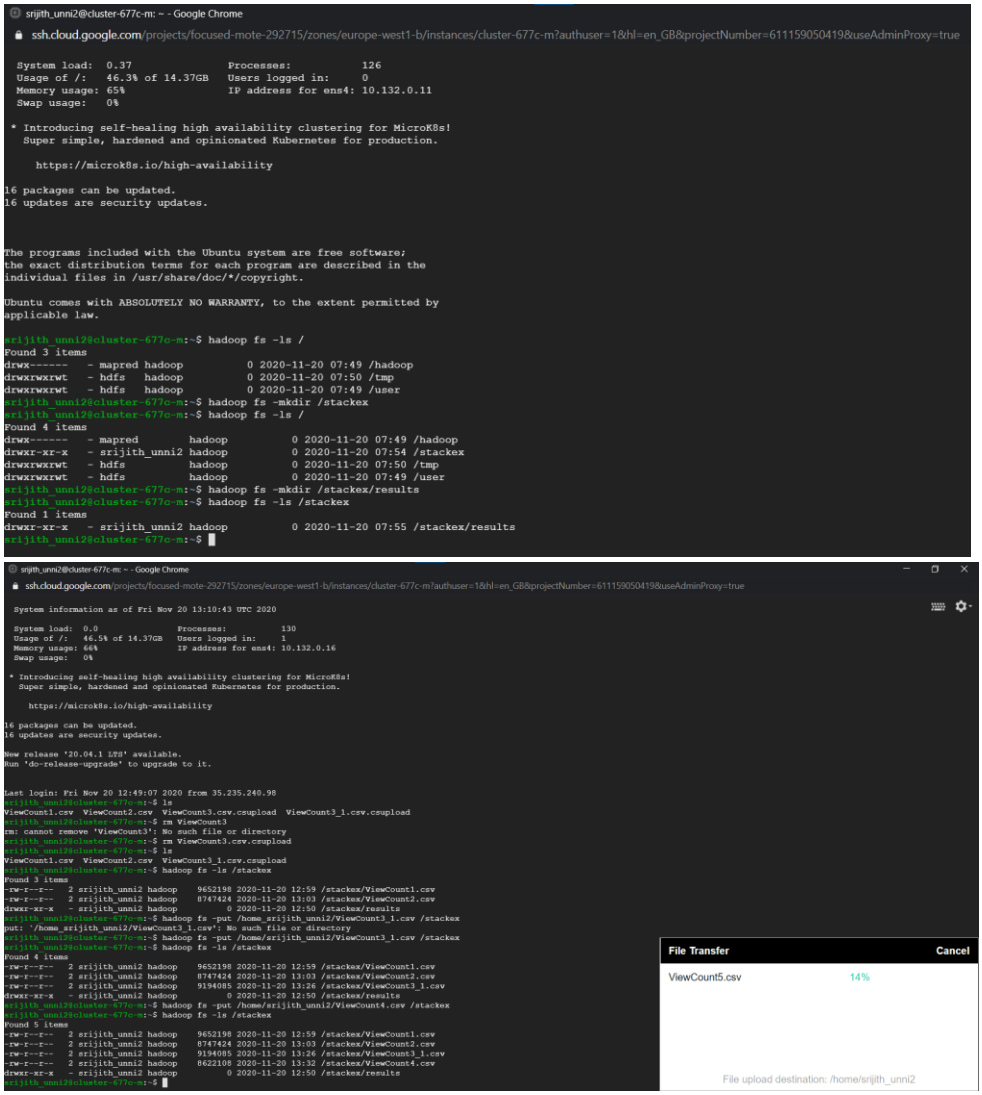
1. Extract input data from Stack Exchange using query provided in Source Code.



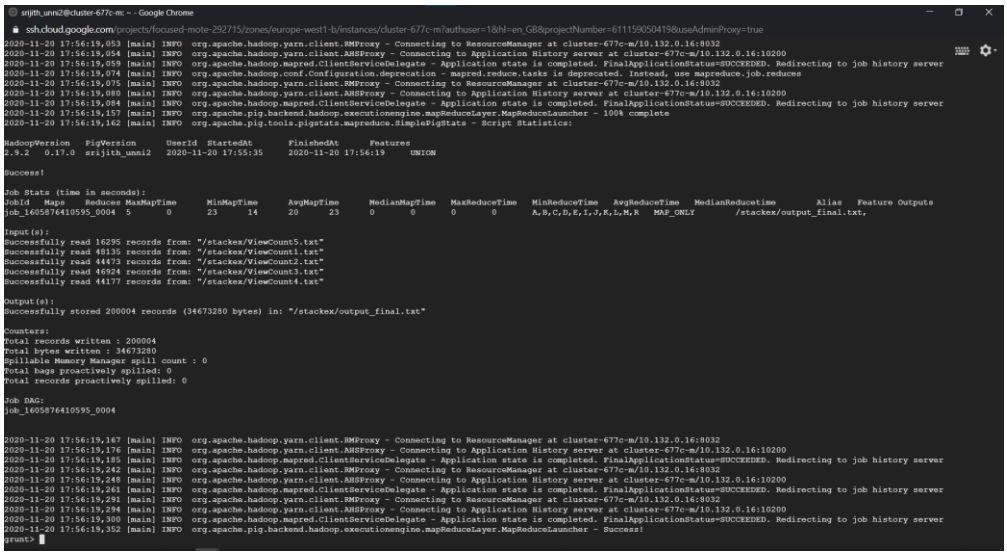
2. Create cluster in Google Cloud Platform and enable Component Gateway (would help in moving files by accessing HDFS NameNode)

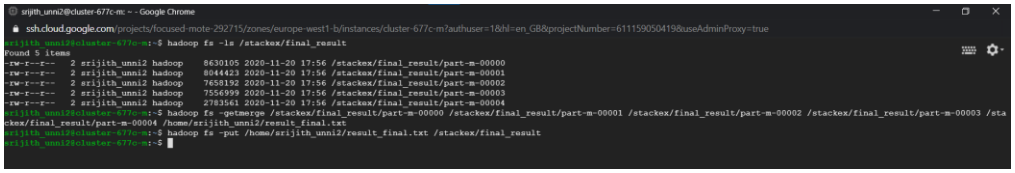


3. Create required folders and move the input files (in tab delimited) into those folders.

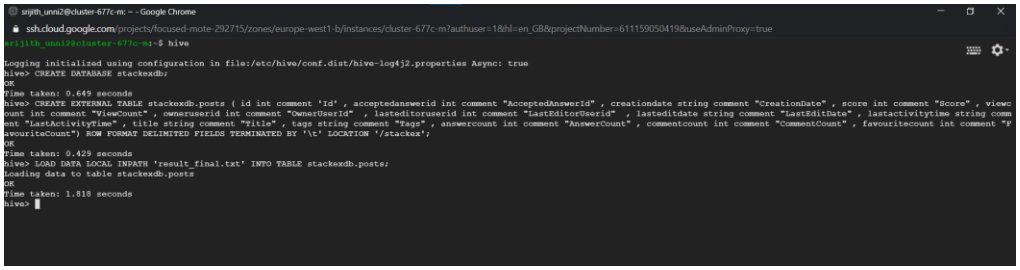


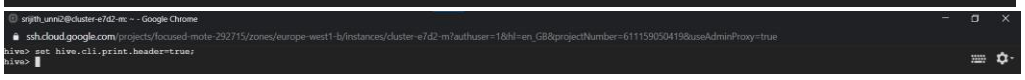
4. Perform ETL actions on the input data in Pig and merge the output files into final clean data file. This output file can then be downloaded from the HDFS NameNode.

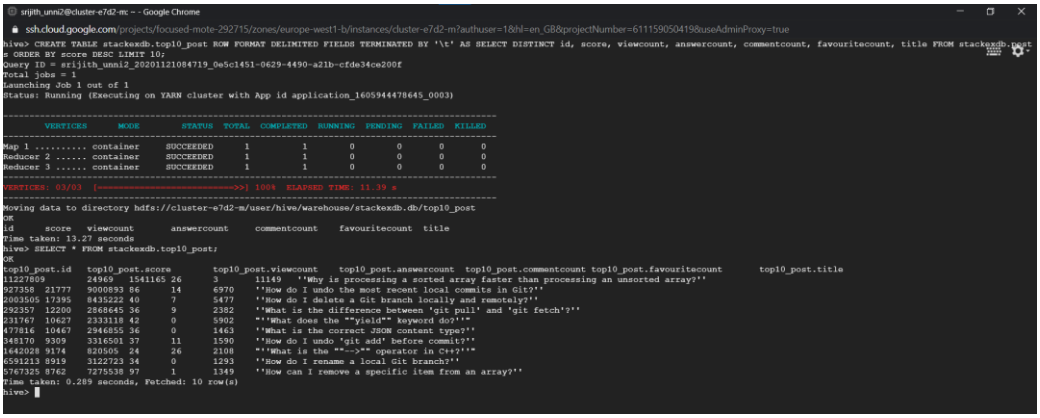


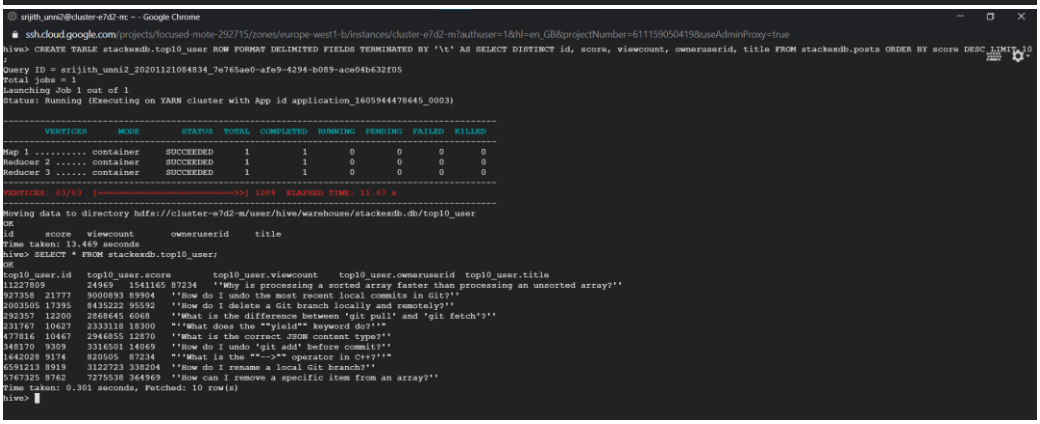


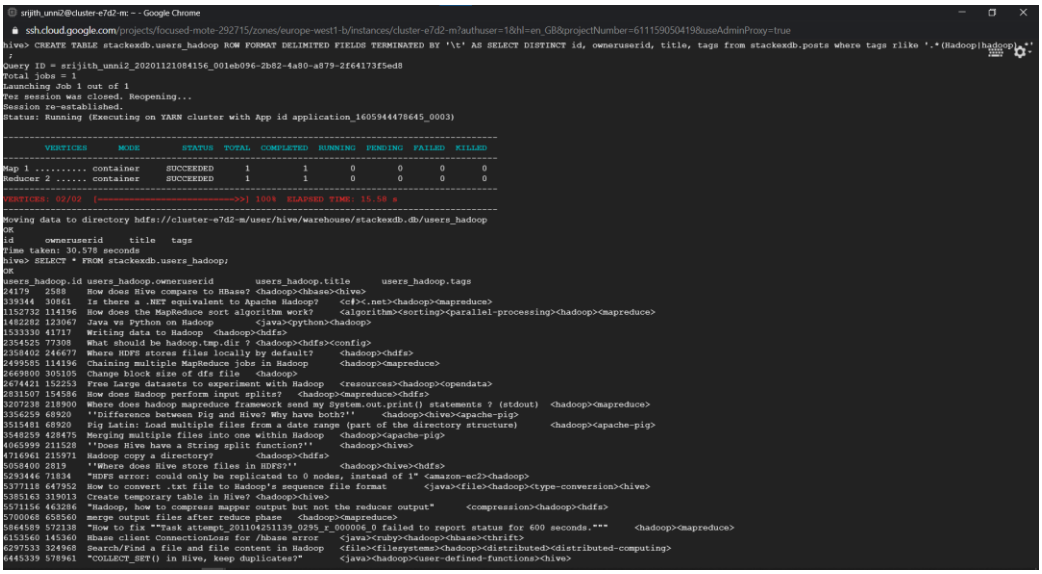
5. Download this file to check the result. On completion of ETL steps, we login to Hive and load the data in a table create in a database.
   Then query as required (10 records each for top 10 posts and users by score, and 295 distinct users with Hadoop in their posts)

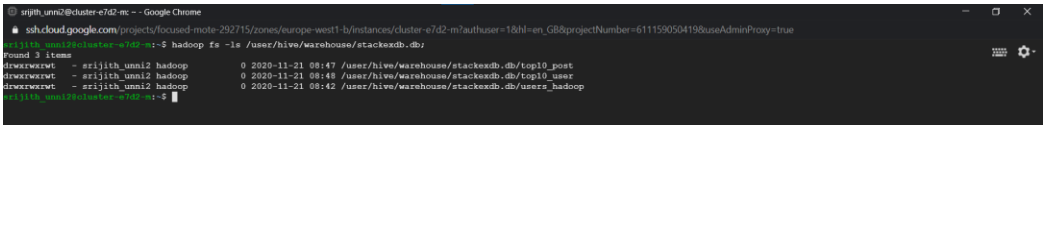




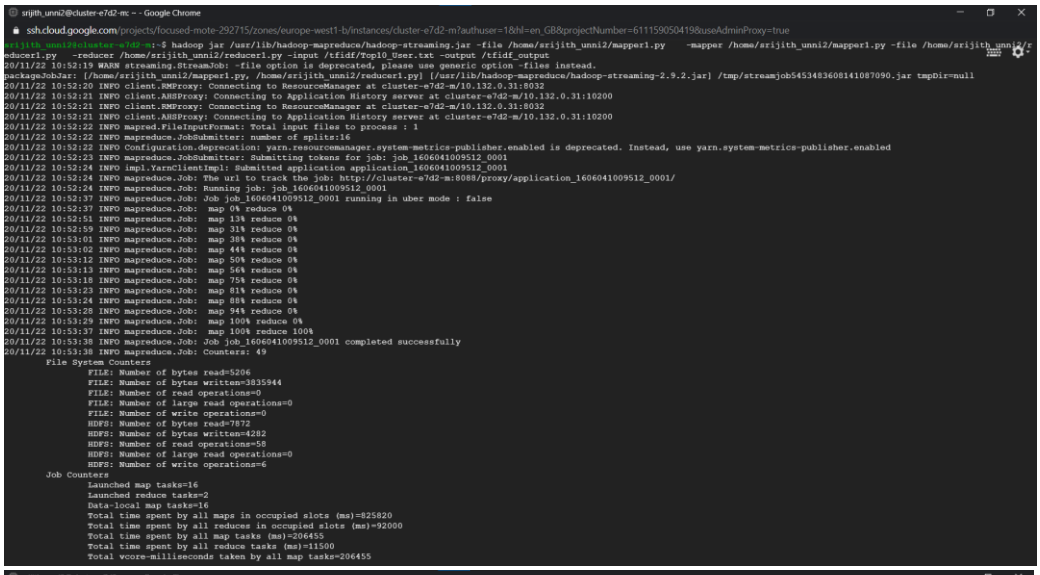

srijith_unni2@cluster-e7d2-m: ~ - Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true
hive> CREATE TABLE stackexdb.top10_post ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' AS SELECT DISTINCT id, score, viewcount, answercount, commentcount, favouritecount, title FROM stackexdb.post
s ORDER BY score DESC LIMIT 10;
Query ID = srijith_unni2_20201121084719_0e5c1451-0629-4490-a21b-cfde34ce200f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605944478645_0003)
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
Reducer 3 container SUCCEEDED 1 1 0 0 0 0
VERTICES: 03/03 [==>>] 100% ELAPSED TIME: 11.39 s
Moving data to directory hdfs://cluster-e7d2-m/user/hive/warehouse/stackexdb.db/top10_post
OK
id score viewcount answercount commentcount favouritecount title
Time taken: 13.27 seconds
hive> SELECT * FROM stackexdb.top10_post;
OK
top10_post.id top10_post.score top10_post.viewcount top10_post.answercount top10_post.commentcount top10_post.favouritecount top10_post.title
11227809 24969 1541165 26 3 11149 ''Why is processing a sorted array faster than processing an unsorted array?''
927358 21777 9000893 86 14 6970 ''How do I undo the most recent local commits in Git?''
2003505 17395 8435222 40 7 5477 ''How do I delete a Git branch locally and remotely?''
292357 12200 2868645 36 9 2382 ''What is the difference between 'git pull' and 'git fetch'?''
231767 10627 2333118 42 0 5902 '''What does the ""yield"" keyword do?'''
477816 10467 2946855 36 0 1463 ''What is the correct JSON content type?''
348170 9309 3316501 37 11 1590 ''How do I undo 'git add' before commit?''
1642028 9174 820505 24 26 2108 '''What is the ""-->"" operator in C++?'''
6591213 8919 3122723 34 0 1293 ''How do I rename a local Git branch?''
8767325 8762 7275538 97 1 1349 ''How can I remove a specific item from an array?''
Time taken: 0.289 seconds, Fetched: 10 row(s)
hive>

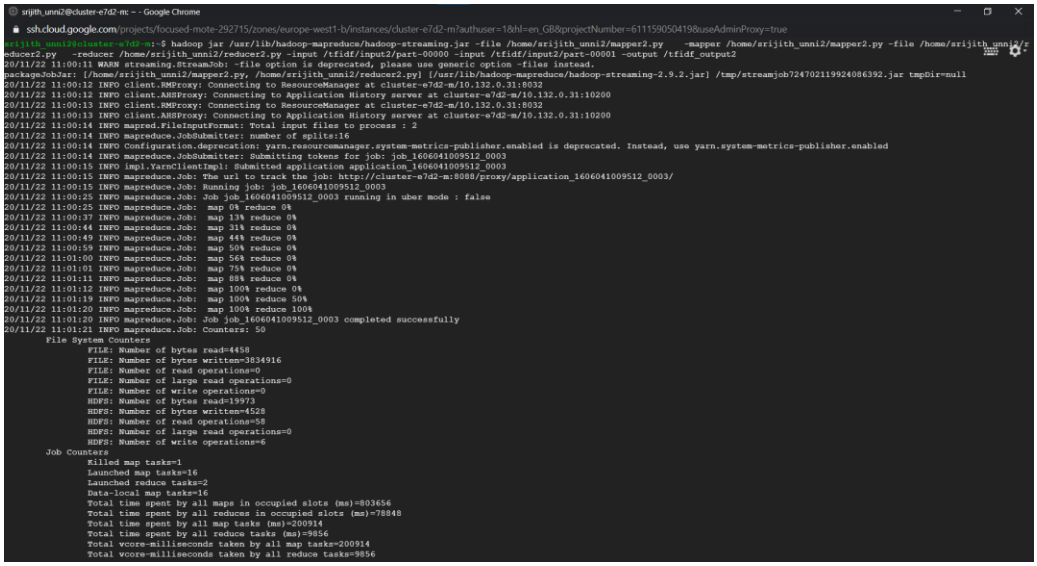

srijith_unni2@cluster-e7d2-m: ~ - Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true
hive> CREATE TABLE stackexdb.top10_user ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' AS SELECT DISTINCT id, score, viewcount, owneruserid, title FROM stackexdb.posts ORDER BY score DESC LIMIT 10
;
Query ID = srijith_unni2_20201121084834_7e765ae0-afe9-4294-b089-ace04b632f05
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605944478645_0003)
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
Reducer 3 container SUCCEEDED 1 1 0 0 0 0
VERTICES: 03/03 [==>>] 100% ELAPSED TIME: 11.67 s
Moving data to directory hdfs://cluster-e7d2-m/user/hive/warehouse/stackexdb.db/top10_user
OK
id score viewcount owneruserid title
Time taken: 13.469 seconds
hive> SELECT * FROM stackexdb.top10_user;
OK
top10_user.id top10_user.score top10_user.viewcount top10_user.owneruserid top10_user.title
11227809 24969 1541165 87234 ''Why is processing a sorted array faster than processing an unsorted array?''
927358 21777 9000893 89904 ''How do I undo the most recent local commits in Git?''
2003505 17395 8435222 95592 ''How do I delete a Git branch locally and remotely?''
292357 12200 2868645 6068 ''What is the difference between 'git pull' and 'git fetch'?''
231767 10627 2333118 18300 '''What does the ""yield"" keyword do?'''
477816 10467 2946855 12870 ''What is the correct JSON content type?''
348170 9309 3316501 14069 ''How do I undo 'git add' before commit?''
1642028 9174 820505 87234 '''What is the ""-->"" operator in C++?'''
6591213 8919 3122723 338204 ''How do I rename a local Git branch?''
8767325 8762 7275538 364969 ''How can I remove a specific item from an array?''
Time taken: 0.301 seconds, Fetched: 10 row(s)
hive>

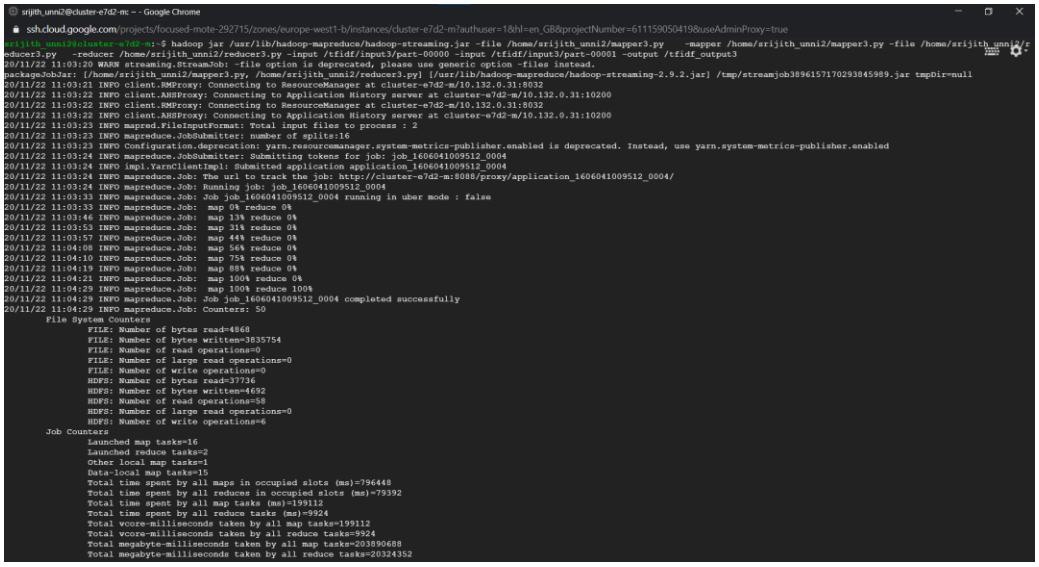

srijith_unni2@cluster-e7d2-m: ~ - Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true
hive> CREATE TABLE stackexdb.users_hadoop ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' AS SELECT DISTINCT id, owneruserid, title, tags from stackexdb.posts where tags rlike '.*(Hadoop|hadoop).*'
;
Query ID = srijith_unni2_20201121084156_001eb096-2b82-4a80-a879-2f64173f5ed8
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1605944478645_0003)
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: 15.58 s
Moving data to directory hdfs://cluster-e7d2-m/user/hive/warehouse/stackexdb.db/users_hadoop
OK
id owneruserid title tags
Time taken: 30.578 seconds
hive> SELECT * FROM stackexdb.users_hadoop;
OK
users_hadoop.id users_hadoop.owneruserid users_hadoop.title users_hadoop.tags
24179 2588 How does Hive compare to HBase? <hadoop><hbase><hive>
339344 30861 Is there a .NET equivalent to Apache Hadoop? <c#><.net><hadoop><mapreduce>
1152732 114196 How does the MapReduce sort algorithm work? <algorithm><sorting><parallel-processing><hadoop><mapreduce>
1482282 123067 Java vs Python on Hadoop <java><python><hadoop>
1533330 41717 Writing data to Hadoop <hadoop><hdfs>
2354525 77308 What should be hadoop.tmp.dir ? <hadoop><hdfs><config>
2358402 246677 Where HDFS stores files locally by default? <hadoop><hdfs>
2499585 114196 Chaining multiple MapReduce jobs in Hadoop <hadoop><mapreduce>
2669800 305105 Change block size of dfs file <hadoop>
2674421 152253 Free Large datasets to experiment with Hadoop <resources><hadoop><opendata>
2831507 154586 How does Hadoop perform input splits? <hadoop><mapreduce><hdfs>
3207238 218900 Where does hadoop mapreduce framework send my System.out.print() statements ? (stdout) <hadoop><mapreduce>
3356259 68920 ''Difference between Pig and Hive? Why have both?'' <hadoop><hive><apache-pig>
3515401 68920 Pig Latin: Load multiple files from a date range (part of the directory structure) <hadoop><apache-pig>
3548259 428475 Merging multiple files into one within Hadoop <hadoop><apache-pig>
4065999 211528 ''Does Hive have a String split function?'' <hadoop><hive>
4716961 215971 Hadoop copy a directory? <hadoop><hdfs>
5058400 2819 ''Where does Hive store files in HDFS?'' <hadoop><hive><hdfs>
5293446 71834 "HDFS error: could only be replicated to 0 nodes, instead of 1" <amazon-ec2><hadoop>
5377118 647952 How to convert .txt file to Hadoop's sequence file format <java><file><hadoop><type-conversion><hive>
5385163 319013 Create temporary table in Hive? <hadoop><hive>
5571156 463286 "Hadoop, how to compress mapper output but not the reducer output" <compression><hadoop><hdfs>
5700068 658560 merge output files after reduce phase <hadoop><mapreduce>
5864589 572138 "How to fix ""Task attempt_201104251139_0295_r_000006_0 failed to report status for 600 seconds."" <hadoop><mapreduce>
6153560 145360 Hbase client ConnectionLoss for /hbase error <java><ruby><hadoop><hbase><thrift>
6297533 324968 Search/Find a file and file content in Hadoop <file><filesystems><hadoop><distributed><distributed-computing>
6445339 578961 "COLLECT_SET() in Hive, keep duplicates?" <java><hadoop><user-defined-functions><hive>

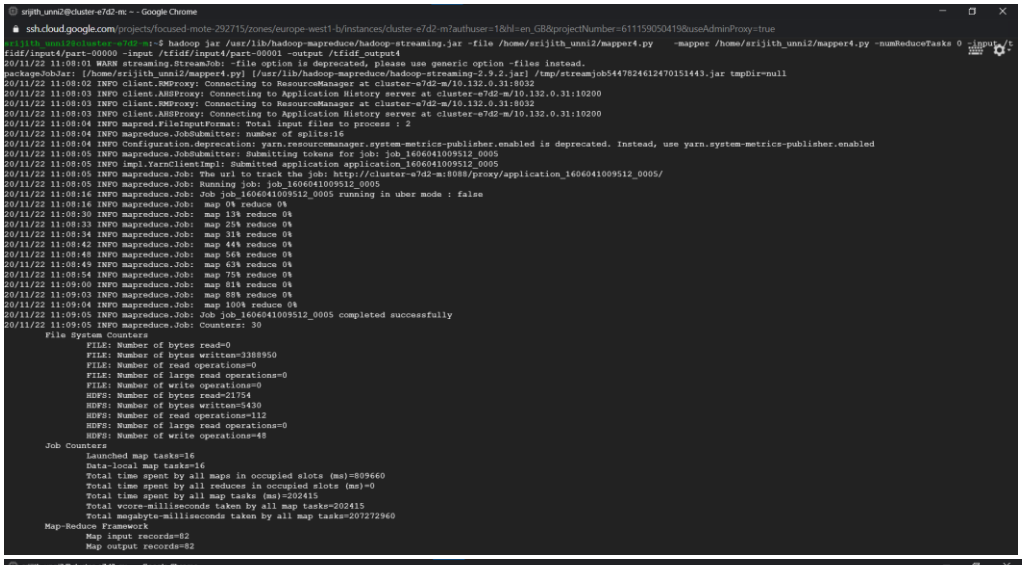

srijith_unni2@cluster-e7d2-m: ~ - Google Chrome
ssh.cloud.google.com/projects/focused-mote-292715/zones/europe-west1-b/instances/cluster-e7d2-m?authuser=1&hl=en_GB&projectNumber=611159050419&useAdminProxy=true
srijith_unni2@cluster-e7d2-m:~$ hadoop fs -ls /user/hive/warehouse/stackexdb.db;
Found 3 items
drwxrwxrwt - srijith_unni2 hadoop 0 2020-11-21 08:47 /user/hive/warehouse/stackexdb.db/top10_post
drwxrwxrwt - srijith_unni2 hadoop 0 2020-11-21 08:48 /user/hive/warehouse/stackexdb.db/top10_user
drwxrwxrwt - srijith_unni2 hadoop 0 2020-11-21 08:42 /user/hive/warehouse/stackexdb.db/users_hadoop
srijith_unni2@cluster-e7d2-m:~$

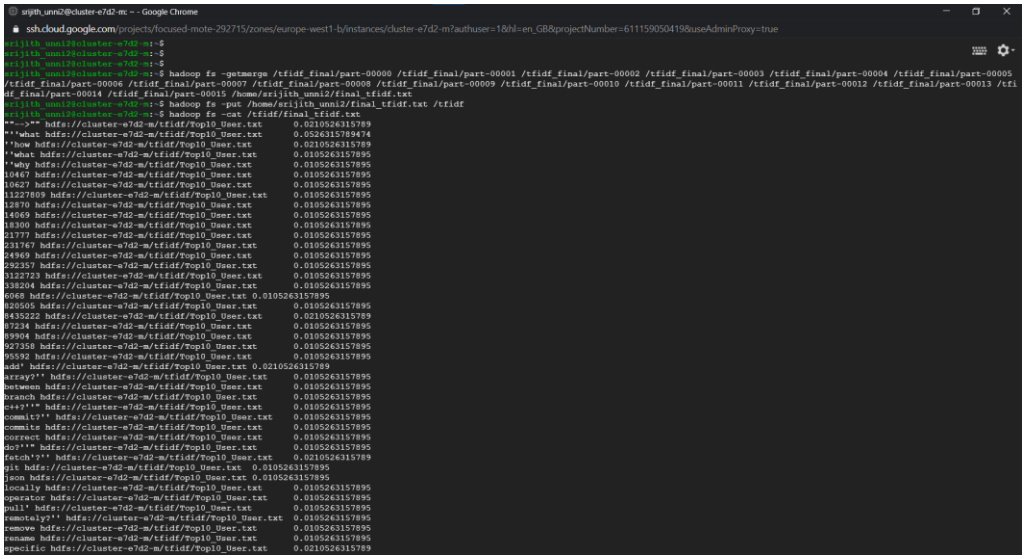
6. Using the Top10_User table, we shall perform the calculation of TF-IDF using MapReduce programs (*Source: https://github.com/devangpatel01/TF-IDF-implementation-using-map-reduce-Hadoop-python-*)
   We have changed the Hadoop commands from source according to our file path and jar version. Mapper and Reducer programs remain unchanged due to logic being consistent.











PS: All input files were uploaded using the “Upload File” option in SSH VM instance and all output files were downloaded from HDFS NameNode. Once data manipulation and extraction actions were completed, clusters were deleted from GCP.