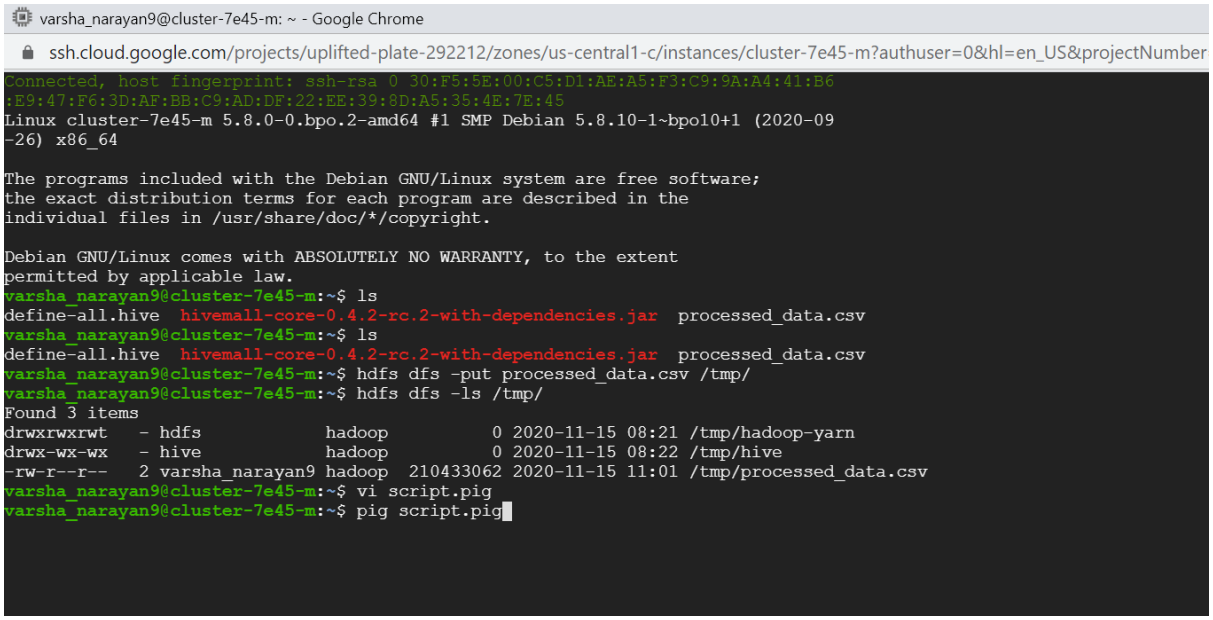
CA675- Assignment 1 - SCREENSHOTS

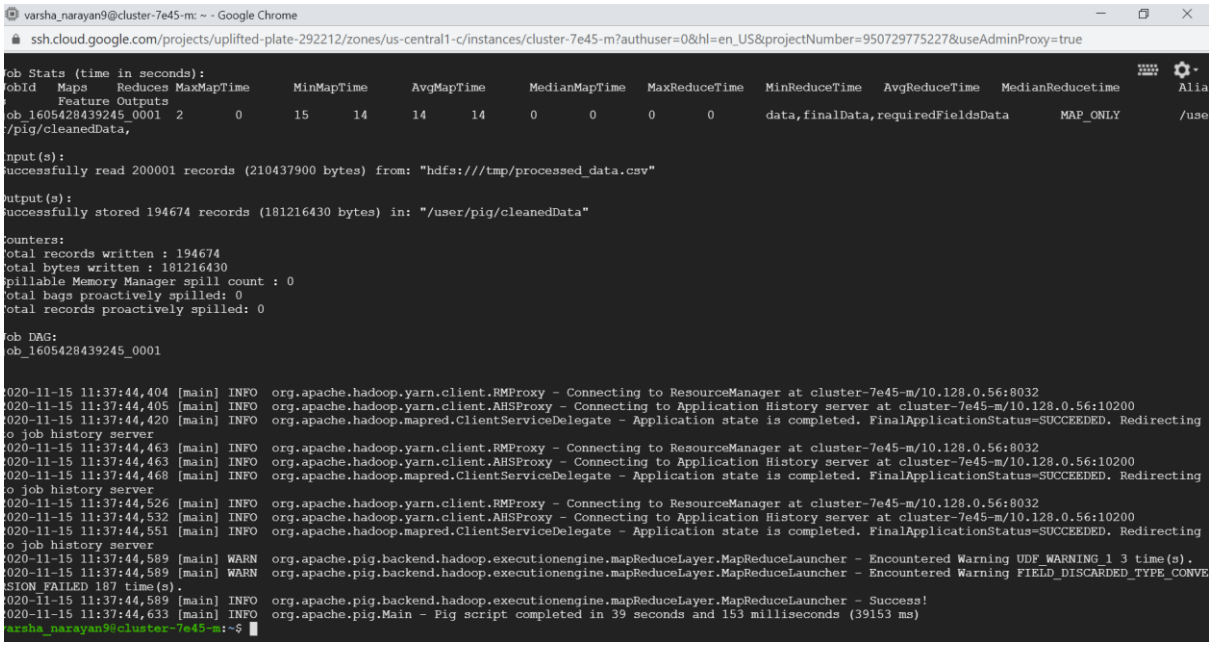
TASK 2 - Data Cleaning/ processing

1) script.pig

```
csvFile = LOAD '/tmp/processed_new.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (Id:int, Post
TypeId:int,  AcceptedAnswerId:int, ParentId:int, CreationDate:chararray, DeletionDate:chararray, Score:int, ViewCount:int, Body:chararray, OwnerUserId:int, OwnerDis
playName:chararray, LastEditorUserId:int, LastEditorDisplayName:chararray, LastEditDate:chararray, LastActivityDate:chararray, Title:chararray, Tags:chararray, Answ
erCount:int, CommentCount:int, FavoriteCount:int, ClosedDate:chararray, CommunityOwnedDate:chararray);

requiredFieldsData = FOREACH csvFile GENERATE $0 AS Id:int, $6 AS Score:int, REPLACE($8,'\\n|\\r|<br>|\\t|<.+?>',' ') AS Body:chararray, $9 AS OwnerUserId:int, $15
AS Title:chararray, $16 AS Tags:chararray;

finalData = FILTER requiredFieldsData BY (OwnerUserId IS NOT NULL);

STORE finalData INTO '/user/pig/cleanedData' USING PigStorage(',');
```

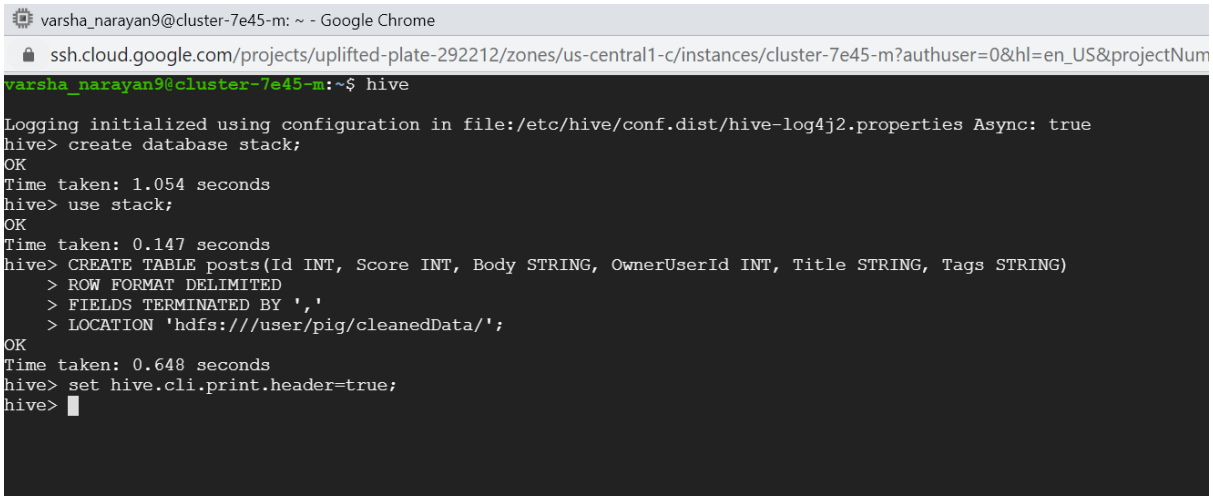
2) Running pig script in MapReduce mode

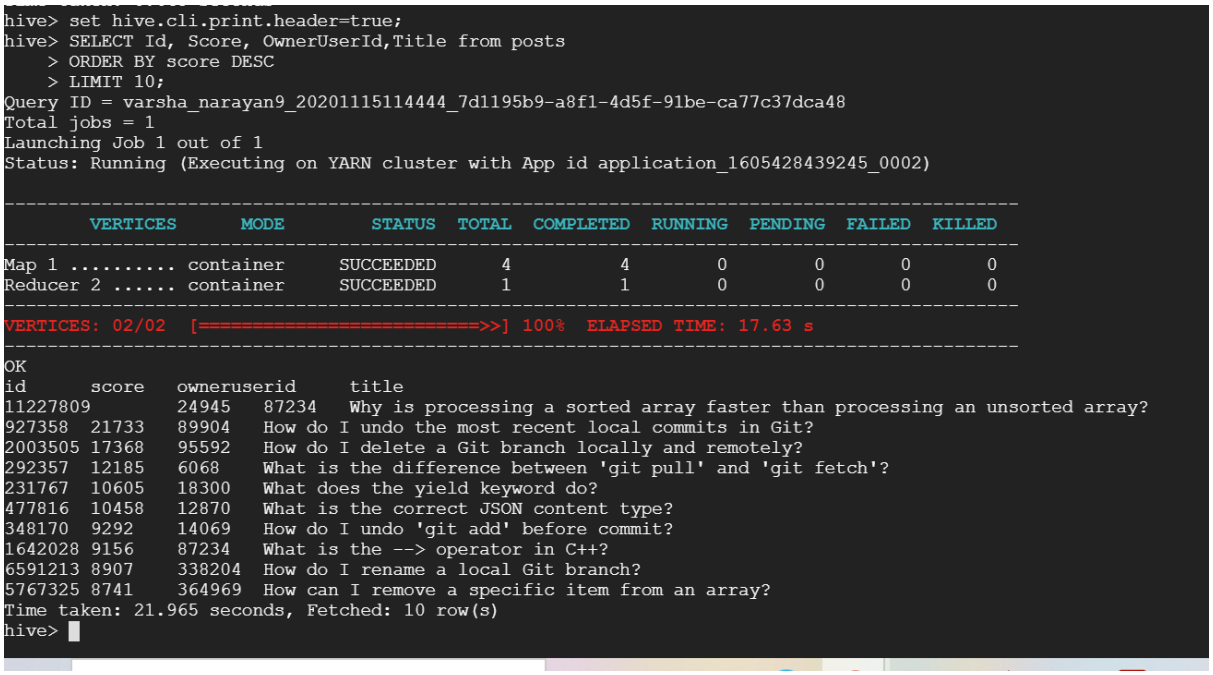




TASK 3 – HIVE QUERIES

1) The top 10 posts by score





2) The top 10 users by post score

```
hive> SELECT OwnerUserId, SUM(Score) AS Total_Score from posts
    > GROUP BY OwnerUserId
    > ORDER BY Total_Score DESC
    > LIMIT 10;
Query ID = varsha_narayan9_20201115114605_662e8e71-0408-403a-b37b-c0fd0d3ee14d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605428439245_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        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: 13.57 s
--------------------------------------------------------------------------------
OK
owneruserid     total_score
87234   36166
4883    26892
9951    25365
6068    24534
89904   22381
51816   21223
49153   18796
95592   18263
63051   18108
179736  17322
Time taken: 14.959 seconds, Fetched: 10 row(s)
hive>
```

3) The number of distinct users, who used the word “Hadoop” in one of their posts

```
hive> SELECT COUNT(DISTINCT OwnerUserId) AS unique_user_Count from posts
    > WHERE (LOWER(BODY) like '%hadoop%' OR LOWER(Title) like '%hadoop%' or LOWER(Tags) like '%hadoop%');
Query ID = varsha_narayan9_20201120194617_f007e01f-f075-4305-9021-fb1da4abbb46
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605891365788_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        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: 28.64 s
--------------------------------------------------------------------------------
OK
unique_user_count
179
Time taken: 29.786 seconds, Fetched: 1 row(s)
```

TASK 4 – TF-IDF Calculation

1) Adding jar and source files to HDFS hive location

```
varsha_narayan9@cluster-7e45-m: ~ - Google Chrome
ssh.cloud.google.com/projects/uplifted-plate-292212/zones/us-central1-c/instances/cluster-7e45-m?authuser=0&hl=en_US&projectNumber=9507
varsha_narayan9@cluster-7e45-m:~$ ls
define-all.hive  hivemall-core-0.4.2-rc.2-with-dependencies.jar  processed_data.csv  script.pig
varsha_narayan9@cluster-7e45-m:~$ hdfs dfs -put hivemall-core-0.4.2-rc.2-with-dependencies.jar /tmp/hive/
varsha_narayan9@cluster-7e45-m:~$ hdfs dfs -put define-all.hive /tmp/hive/
varsha_narayan9@cluster-7e45-m:~$
```

2) Data Preparation

```
hive> CREATE TABLE TopUsersScore AS SELECT OwnerUserId, SUM(Score) AS Total_Score from posts
    > GROUP BY OwnerUserId
    > ORDER BY Total_Score DESC
    > LIMIT 10;
Query ID = varsha_narayan9_20201115115305_66d5d7b8-898a-4481-9ea6-b5c1574cbd4c
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605428439245_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        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: 18.39 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-7e45-m/user/hive/warehouse/stack.db/topusersscore
OK
Time taken: 24.475 seconds
hive> CREATE TABLE TopUsersPosts AS
    > SELECT OwnerUserId,Body,Title,Tags from posts
    > WHERE OwnerUserId IN (SELECT OwnerUserId from TopUsersScore)
    > GROUP BY OwnerUserId, Body, Title, Tags;
Query ID = varsha_narayan9_20201115115344_8dd21a87-e944-41b6-9055-c684239618ef
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605428439245_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      4          4        0        0       0       0
Map 3 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 4 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 17.28 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-7e45-m/user/hive/warehouse/stack.db/topusersposts
```

```
Time taken: 19.091 seconds
hive> INSERT OVERWRITE DIRECTORY '/user/hive/hiveTableContent'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT * FROM TopUsersPosts;
Query ID = varsha_narayan9_20201115115529_c7b2d03f-b5b9-4407-9cda-637cd4bcda0d
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605428439245_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 4.67 s
--------------------------------------------------------------------------------
Moving data to directory /user/hive/hiveTableContent
OK
Time taken: 5.303 seconds
hive>
```

3) Pig Processing of Body, Title, Tags columns for TF-IDF calculation (hiveScript.pig)

```
hiveData = LOAD '/user/hive/hiveTableContent' USING PigStorage(',');
CombineFields = FOREACH hiveData GENERATE $0, CONCAT ($1, $2, $3) AS POSTS;
CleanPosts = FOREACH CombineFields GENERATE $0, REPLACE($1,'\\n|\\r|<br>|\\t|<.+?>|([^a-zA-Z\\s]+)',' ') AS POSTS;
STORE CleanPosts INTO '/user/pig/hiveClean3' USING PigStorage(',');
```

```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Ali
s       Feature Outputs
job_1605844482112_0016  1       0       5       5       5       5       0       0       0       0       CleanPosts,CombineFields,hiveData       MAP_ONLY        /us
r/pig/hiveClean3,

Input(s):
Successfully read 321 records (131535 bytes) from: "/user/hive/hiveTableContent"

Output(s):
Successfully stored 321 records (127670 bytes) in: "/user/pig/hiveClean3"

Counters:
Total records written : 321
Total bytes written : 127670
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1605844482112_0016


2020-11-20 07:46:29,140 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-83b7-m/10.128.0.62:8032
2020-11-20 07:46:29,141 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-83b7-m/10.128.0.62:10200
2020-11-20 07:46:29,153 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
to job history server
2020-11-20 07:46:29,179 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-83b7-m/10.128.0.62:8032
2020-11-20 07:46:29,179 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-83b7-m/10.128.0.62:10200
2020-11-20 07:46:29,189 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
to job history server
2020-11-20 07:46:29,240 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-83b7-m/10.128.0.62:8032
2020-11-20 07:46:29,241 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-83b7-m/10.128.0.62:10200
2020-11-20 07:46:29,247 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
to job history server
2020-11-20 07:46:29,302 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2020-11-20 07:46:29,342 [main] INFO  org.apache.pig.Main - Pig script completed in 28 seconds and 160 milliseconds (28160 ms)
```

4) Hive Mall commands for TF-IDF calculation

```
hive> use stack;
OK
Time taken: 0.564 seconds
hive> add jar hivemall-core-0.4.2-rc.2-with-dependencies.jar;
Added [hivemall-core-0.4.2-rc.2-with-dependencies.jar] to class path
Added resources: [hivemall-core-0.4.2-rc.2-with-dependencies.jar]
hive> source define-all.hive;
```

```
hive> create temporary macro max2(x INT, y INT) if(x>y,x,y);
OK
Time taken: 0.053 seconds
hive> create temporary macro tfidf(tf FLOAT, df_t INT, n_docs INT) tf * (log(10, CAST(n_docs as FLOAT)/max2(1,df_t)) + 1.0);
OK
Time taken: 0.076 seconds
```

```
hive> create external table TF_IDFCalc1 (
    >   userid int,
    >   posts string
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.116 seconds
hive> LOAD DATA LOCAL INPATH '/home/varsha_narayan9/hiveClean3/part-m-00000' INTO TABLE TF_IDFCalc1;
Loading data to table stack.tf_idfcalc1
OK
Time taken: 0.727 seconds
hive> SELECT * FROM TF_IDFCalc1;
OK
```

```
OK
Time taken: 0.402 seconds
hive> create external table TF_IDFCalc1 (
    >   userid int,
    >   posts string
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.135 seconds
hive> LOAD DATA LOCAL INPATH '/home/varsha_narayan9/hiveClean3/part-m-00000' INTO TABLE TF_IDFCalc1;
Loading data to table stack.tf_idfcalc1
OK
Time taken: 0.575 seconds
hive> create or replace view posts_exploded
    > as
    > select
    >   userid,
    >   word
    > from
    >   TF_IDFCalc1 LATERAL VIEW explode(tokenize(posts,true)) t as word
    > where
    >   not is_stopword(word);
OK
Time taken: 0.315 seconds
hive> create or replace view term_frequency
    > as
    > select
    >   userid,
    >   word,
    >   freq
    > from (
    > select
    >   userid,
    >   tf(word) as word2freq
    > from
    >   posts_exploded
    > group by
    >   userid
    > ) t
    > LATERAL VIEW explode(word2freq) t2 as word, freq;
OK
```

```
hive> create or replace view document_frequency
    > as
    > select
    >   word,
    >   count(distinct userid) docs
    > from
    >   posts_exploded
    > group by
    >   word;
OK
Time taken: 0.221 seconds
hive> select count(distinct userid) from TF_IDFCalc1;
Query ID = varsha_narayan9_20201120211736_463b11e3-4032-430d-8989-37c5e4ad4d2a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605891365788_0007)

--------------------------------------------------------------------------------
        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: 9.09 s
--------------------------------------------------------------------------------
OK
10
Time taken: 14.496 seconds, Fetched: 1 row(s)
hive> set hivevar:n_docs=10;
hive> create or replace view tfidf
    > as
    > select
    >   tf.userid,
    >   tf.word,
    >   -- tf.freq * (log(10, CAST(${n_docs} as FLOAT)/max2(1,df.docs)) + 1.0) as tfidf
    >   tfidf(tf.freq, df.docs, ${n_docs}) as tfidf
    > from
    >   term_frequency tf
    >   JOIN document_frequency df ON (tf.word = df.word)
    > Order BY tf.userid;
OK
```

5) TF-IDF results (10 terms for top 10 users) – 100 records fetched

```
hive> SELECT S.userid, S.word, S.tfidf
    > FROM
    >  (SELECT userid, word, tfidf,  row_number() over (partition by userid) as r FROM tfidf) S
    > WHERE S.r < 11;
Query ID = varsha_narayan9_20201120211919_38e3c3c9-da7b-4f4f-b7a2-34bda75e4432
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605891365788_0007)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 4 .......... 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
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 18.11 s
--------------------------------------------------------------------------------
OK
4883    example 0.003313927528660059
4883    waiting 0.003021148033440113
4883    without 0.003021148033440113
4883    behavior        0.002566419943736752
4883    quot    0.004600842126471439
4883    another 0.001656963764330029 4
4883    needed  0.00211169185403339
4883    control 0.0018456929735323581
4883    deployed        0.003021148033440113
4883    whether 0.0036913859470647163
6068    integer 0.0014556040987372398
6068    january 0.0012365138509715676
6068    evidence        0.005822416394948959
6068    point   0.0010174236032058951
6068    professional    0.0014556040987372398
6068    new     0.0030444177790905787
6068    odd     0.0014556040987372398
6068    thread  0.0012365138509715676
6068    open    0.00177852804797854
6068    mock    0.0029112081974744797
```

```
9951    command 0.0019345855023047657
9951    javascript      0.010325635482444879
9951    learning        0.0013429265424919476
9951    hours   0.0013429265424919476
9951    concrete        0.0017636683769524097
9951    attempt 0.0014982098350190673
9951    difference      0.0010774680005586053
9951    successfully    0.0014982098350190673
9951    urls    0.0029964196700381345
9951    question        0.0022945854608190947
49153   savebtn 7.443245267495513E-4
49153   netbeans        0.0014886490534991026
49153   notifications   7.443245267495513E-4
49153   robots  7.443245267495513E-4
49153   audio   7.443245267495513E-4
49153   gap     7.443245267495513E-4
49153   either  0.001040521035379078
49153   purpose 0.0014886490534991026
49153   keep    4.841942679047818E-4
49153   apples  0.002232973463833332
51816   framework       0.0012012012302875519
51816   lists   0.0033584145165700846
51816   zip     0.0012012012302875519
51816   bind    0.004081609718860146
51816   individual      0.0012012012302875519
51816   execution       0.002040804859430073
51816   cache   9.146419112047525E-4
51816   also    0.0041618089655891385
51816   luck    0.0012012012302875519
51816   subtract        0.0010204024297150365
63051   ascii   0.008965541287370659
63051   gz      0.0035180298145860434
63051   output  0.008682665366417786
63051   sandbox 0.0017590149072930217
63051   lengths 0.0017590149072930217
63051   lstrh   0.0017590149072930217
63051   columns 0.002678766414947806
63051   connections     0.0035180298145860434
63051   allow   0.0017590149072930217
63051   suggesting      0.0017590149072930217
87234   sorting 0.008580656579564291
87234   li      0.0174985150667349
```

```
87234   sorting 0.008580656579564291
87234   li      0.0174985150667349
87234   seconds 0.00706030306752578
87234   public  0.0058328381763139234
87234   snippet 0.0038456533935910257
87234   silly   0.005050505045801401
87234   times   0.004290328289782146
87234   coding  0.00353015153376289
87234   favorite        0.005050505045801401
87234   formatting      0.00353015153376289
89904   setautoresizemode       0.006230529397726059
89904   recommend       0.005292741278935168
89904   called  0.0038063822770297677
89904   first   0.0038063822770297677
89904   modify  0.010585482557870335
89904   resize  0.018691588193178177
89904   property        0.004744170395820659
89904   size    0.009488340791641318
89904   get     0.010251495124060154
89904   appending       0.005292741278935168
95592   http    0.003658612333356212
95592   jesse   0.002111932495608926
95592   homebrew        0.002111932495608926
95592   full    0.0035881099612220754
95592   co      0.002111932495608926
95592   global  0.0014761774656131498
95592   named   0.0016081085545148466
95592   whole   0.0017940549806110377
95592   peak    0.002111932495608926
95592   path    0.00774138595261792
179736  jar     0.001212965935082691
179736  let     0.006123990546670973
179736  bash    6.064829675413455E-4
179736  loop    5.181322118231699E-4
179736  css     0.002226905608950817
179736  strong  0.0014597953510070809
179736  manually        6.064829675413455E-4
179736  theurl  0.005575467832386494
179736  request 0.002425931870165382
179736  splitting       7.964954129420221E-4
Time taken: 19.55 seconds, Fetched: 100 row(s)
hive>
```

6) Store Results

```
hive> INSERT OVERWRITE DIRECTORY '/user/hive/TF_IDF_FinalResult'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > SELECT S.userid, S.word, S.tfidf
    > FROM
    >  (SELECT userid, word, tfidf,  row_number() over (partition by userid) as r FROM tfidf) S
    > WHERE S.r < 11;
Query ID = varsha_narayan9_20201120212223_04aa9502-ea3b-416f-90d7-dd4c520722eb
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1605891365788_0007)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Map 4 .......... 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
Reducer 5 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 05/05  [==========================>>] 100%  ELAPSED TIME: 17.98 s
----------------------------------------------------------------------------------------------
Moving data to directory /user/hive/TF_IDF_FinalResult
OK
Time taken: 19.084 seconds
hive>
```