CA675 Cloud Technologies : Assignment 1

Name: SNEHASIS NAYAK
Student ID: 21260962
Mail: snehasis.nayak2@mail.dcu.ie

Tasks

2.1. Task 1 - Acquire the top 200,000 posts by ViewCount (see Section 3 - Data
Acquisition for more details)

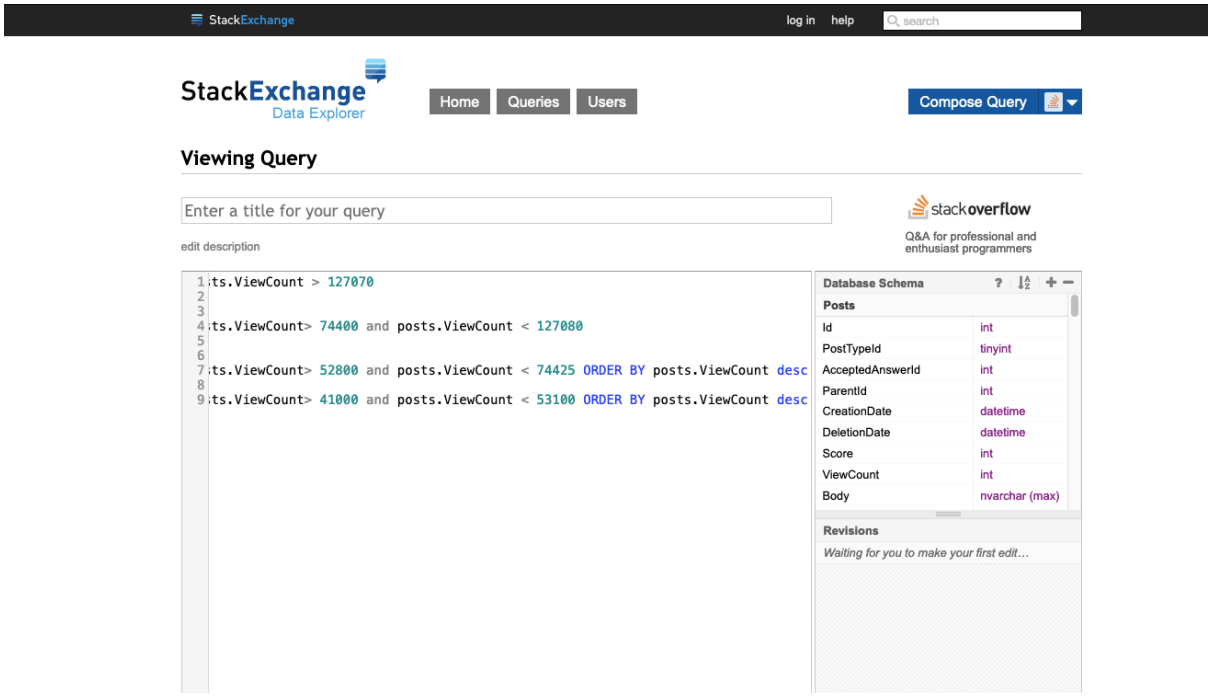
2.2. Task 2 & 3 - Use Pig/Hive/MapReduce - Extract, Transform and Load the data as
applicable to get:
2.2.1. The top 10 posts by score
2.2.2. The top 10 users by post score
2.2.3. The number of distinct users, who used the word “cloud” in one of their posts

2.3. 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

STEP 1 Extracting the data

Firstly, I pulled the data from StackExchange for top 200,000 posts using SQL query. The data was spread across 4 csv files with each file having 50,000 rows which I named 1.csv,2.csv,3.csv and 4.csv respectively.

Below is the screenshot of the queries I used.

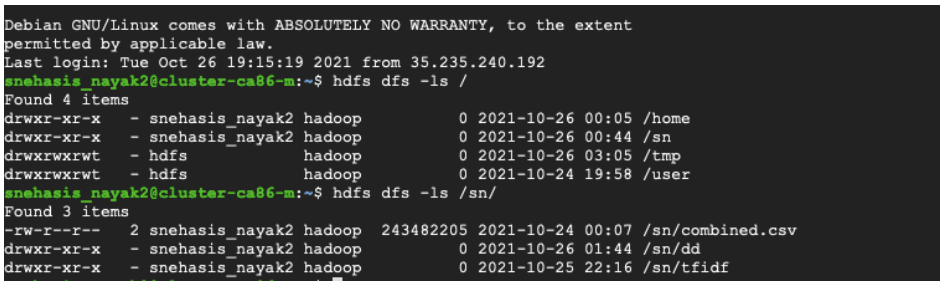


Then I combined all the 4 files into one combined csv file using below command.

cd /Users/snehasis/Desktop/Combine
cat *.csv >combined.csv

STEP 2 (Loading the Data)

I created a cluster in GCP using Data Proc with 1 namenode and 2 worker node. After that I SSH into the cloud terminal and successfully uploaded the data into the cluster. Then I created a directory in HDFS file system and uploaded the data into it using PUT command. Here the below screenshot the combined CSV file.



STEP 3 (Cleaning the data)

I used Pig technology for cleaning the data. I used PIG command to get into the grunt shell the loaded the combined CSV file into newdata variable and then cleaned the data by removing duplicate rows and kept necessary columns that were required for Data Analysis. The next step was to create a database and create a table in hive and store the newdata into the table.

Below is the screenshot of loading the data into newdata variable and cleaning the data.

```
snehasis_nayak2@cluster-ca86-m:~$ pig -useHCatalog
ls: cannot access '/usr/lib/hive/lib/slf4j-api-*.jar': No such file or directory
ls: cannot access '/usr/lib/hive/lib/hive-hbase-handler-*.jar': No such file or directory
ls: cannot access '/usr/lib/hive-hcatalog/lib/*hbase-storage-handler-*.jar': No such file or directory
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using value of HADOOP_PREFIX.
2021-10-25 20:20:49,599 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2021-10-25 20:20:49,601 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2021-10-25 20:20:49,601 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2021-10-25 20:20:49,706 [main] INFO  org.apache.pig.Main - Apache Pig version 0.18.0-SNAPSHOT (r: unknown) compiled Dec 21 1969, 06:05:27
2021-10-25 20:20:49,706 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/snehasis_nayak2/pig_1635193249702.log
2021-10-25 20:20:49,745 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/snehasis_nayak2/.pigbootup not found
2021-10-25 20:20:50,297 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2021-10-25 20:20:50,298 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://cluster-ca86-m
2021-10-25 20:20:51,730 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-b6f79058-39b6-49b2-a030-d23f2ae55a36
2021-10-25 20:20:51,923 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeline service address: cluster-ca86-m:8188
2021-10-25 20:20:52,340 [main] INFO  org.apache.pig.backend.hadoop.PigATSClient - Created ATS Hook
2021-10-25 20:20:52,417 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-me
trics-publisher.enabled
grunt> newdata = LOAD '/sn/combined.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (id:int, posttypeid:int, accepted
answerid:int, parentid:int, creationdate:chararray, deletiondate:chararray, score:int, viewcount:int, body:chararray, owneruserid:int, ownerdisplayname:chararray, lasteditoruserid:int,
 lasteditordisplayname:chararray, lasteditdate:chararray, lastactivitydate:chararray, title:chararray, tags:chararray, answercount:int, commentcount:int, favoritecount:int, closeddate:
chararray, communityowneddate:chararray, contentlicense:chararray );
2021-10-25 20:21:04,278 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-me
trics-publisher.enabled
grunt> newdata = FOREACH newdata GENERATE id, score, viewcount, body, owneruserid, title, tags;
grunt> data = DISTINCT newdata;
grunt> newdata = FOREACH newdata GENERATE id, score, viewcount , owneruserid, title, tags, (REPLACE(body,'[\r\n]+','')) AS body;
```

Below is the screenshot shows the DB creation and table creation in Hive.

```
snehasis_nayak2@cluster-ca86-m:~$ hive
Hive Session ID = 0c87e936-923c-4da0-abd8-dcc30566cab6

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
Hive Session ID = 79058b80-c6f1-4398-b0e2-1e32024c0a65
hive> show databases;
OK
db
default
Time taken: 1.19 seconds, Fetched: 2 row(s)
hive> CREATE DATABASE dc;
OK
Time taken: 0.104 seconds
hive> show tables;
OK
postsdata
Time taken: 0.075 seconds, Fetched: 1 row(s)
hive> use dc;
OK
Time taken: 0.04 seconds
hive> show tables;
OK
Time taken: 0.056 seconds
hive> CREATE TABLE stark (id INT, Score INT, ViewCount INT, Body STRING,  OwnerUserId INT, Title STRING, Tags STRING) ;
OK
Time taken: 0.374 seconds
```

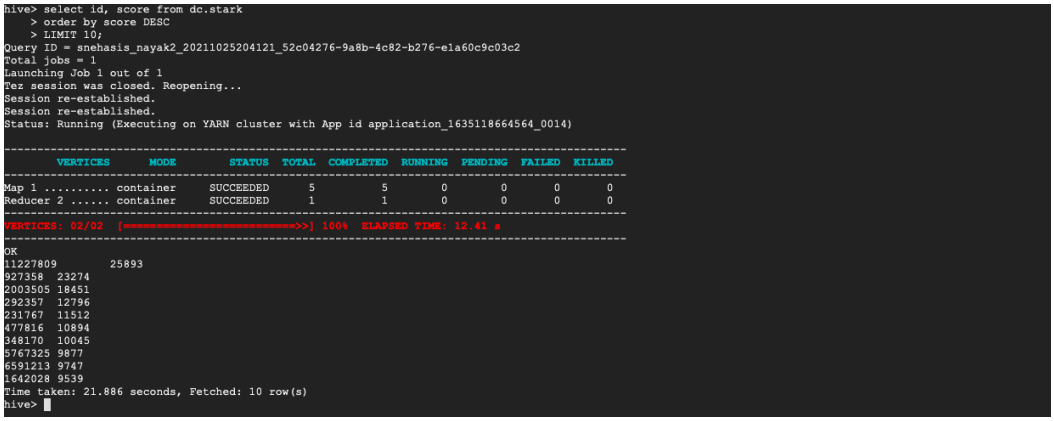
Below screenshot represents the newdata successfully uploaded in the Table.

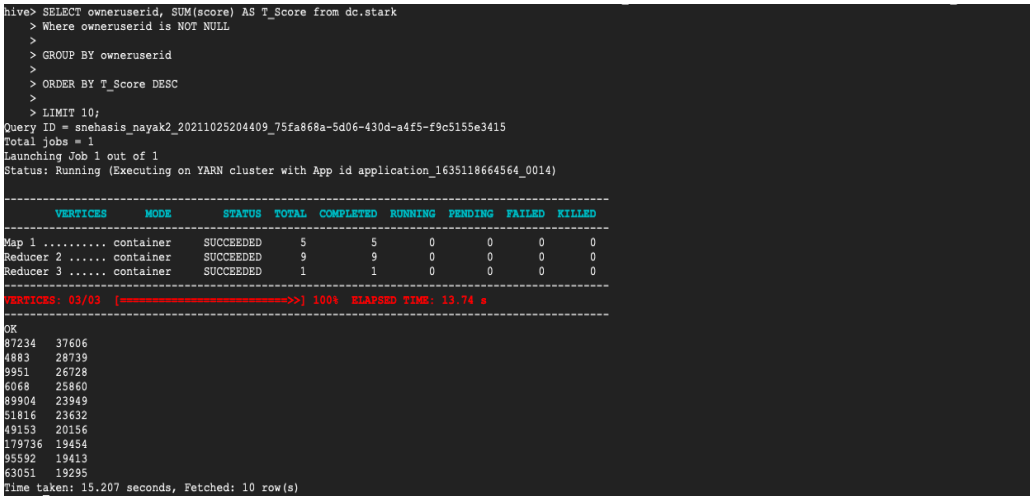
```
grunt> describe newdata;
2021-10-25 20:35:11,608 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-me
trics-publisher.enabled
newdata: {id: int,score: int,viewcount: int,owneruserid: int,title: chararray,tags: chararray,body: chararray}
grunt> 2021-10-25 20:36:23,542 [HiveClientCache-cleaner-0] INFO  org.apache.hadoop.hive.metastore.HiveMetaStoreClient - Closed a connection to metastore, current connections: 0

grunt> STORE newdata INTO 'dc.stark' USING org.apache.hive.hcatalog.pig.HCatStorer();
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.2   0.18.0-SNAPSHOT snehasis_nayak2 2021-10-25 20:37:16    2021-10-25 20:38:29     DISTINCT

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1635118664564_0013  2       1       41      36      38      38      12      12      12      12      newdata DISTINCT        dc.stark,

Input(s):
Successfully read 200001 records (243487029 bytes) from: "/sn/combined.csv"

Output(s):
Successfully stored 199811 records (215683684 bytes) in: "dc.stark"

Counters:
Total records written : 199811
Total bytes written : 215683684
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1635118664564_0013


2021-10-25 20:38:29,173 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-ca86-m/10.164.0.2:8032
2021-10-25 20:38:29,175 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-ca86-m/10.164.0.2:10200
2021-10-25 20:38:29,179 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-25 20:38:29,220 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-ca86-m/10.164.0.2:8032
2021-10-25 20:38:29,221 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-ca86-m/10.164.0.2:10200
2021-10-25 20:38:29,228 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-25 20:38:29,253 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-ca86-m/10.164.0.2:8032
2021-10-25 20:38:29,254 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-ca86-m/10.164.0.2:10200
2021-10-25 20:38:29,260 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-25 20:38:29,298 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 185 tim
e(s).
2021-10-25 20:38:29,298 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

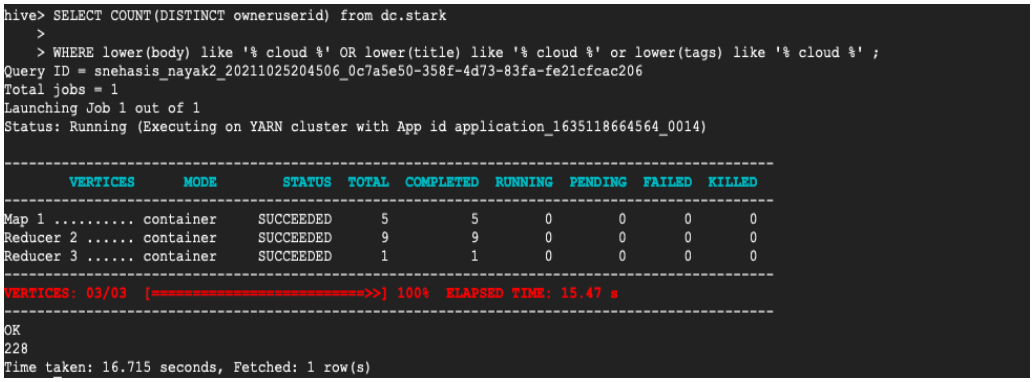
STEP 4 (Query Results)

I successfully ran the queries for top 10 posts by score, top 10 score by post score and number of distinct users, who used the word “cloud” in one of their posts.

Below are the three screenshots showing the query and the resultant output.







STEP 5 (CLEANING AND STORING DATA FOR CALULATING TFIDF)

The data was uploaded into a variable and performed data cleaning and extracted the posts of the top 10 users.

Below are the screenshots of the Actions performed.

```
grunt> newdata = LOAD '/sn/combined.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE','NOCHANGE','SKIP_INPUT_HEADER') AS (id:int, posttypeid:int, accepted
answerid:int, parentid:int, creationdate:chararray, deletiondate:chararray, score:int, viewcount:int, body:chararray, owneruserid:int, ownerdisplayname:chararray, lasteditoruserid:int,
 lasteditordisplayname:chararray, lasteditdate:chararray, lastactivitydate:chararray, title:chararray, tags:chararray, answercount:int, commentcount:int, favoritecount:int, closeddate:
chararray, communityowneddate:chararray, contentlicense:chararray );
2021-10-25 21:48:16,315 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-me
trics-publisher.enabled
grunt> newdata = FOREACH newdata GENERATE id, score, owneruserid, REPLACE(REPLACE(REPLACE(REPLACE((REPLACE(body,'[\r\n]+','')),'<[^>]>' , ' '),'[^a-zA-Z\\s\']+',' '),'(?=\\S[\'
])([a-zA-Z\'-]+)','$1'),'(?<![\\w\\-])\\w(?![\\w\\-])','')),'[ ]{2,}',' ') AS body;
grunt> newdata_distinctuserspost = GROUP newdata BY owneruserid;
grunt> newdata_usersbymaxscore = FOREACH newdata_distinctuserspost GENERATE group AS userid, MAX(newdata.score) AS maxscore;
grunt> newdata_usersbymaxscoredescorder = ORDER newdata_usersbymaxscore BY maxscore DESC;
grunt> newdata_limit_10 = LIMIT postData_usersbymaxscoredescorder 10;
2021-10-25 22:06:23,737 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse: <line 6, column 25> Undefined alias: postData_usersbymaxscoredescorder
Details at logfile: /home/snehasis_nayak2/pig_1635198466370.log
grunt> newdata_limit_10 = LIMIT newdata_usersbymaxscoredescorder 10;
grunt> newdata_top10userid = FOREACH newdata_limit_10 GENERATE userid;
grunt> newdata_postsby10users = JOIN newdata BY owneruserid, newdata_top10userid BY userid;
grunt> newdata_postsby10users = FOREACH newdata_postsby10users GENERATE newdata::owneruserid, LOWER(TRIM(REPLACE(newdata::body,'[ ]{2,}',' '))) AS newdata::body;
grunt> STORE newdata_postsby10users INTO '/sn/tfidf' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','NOCHANGE','SKIP_OUTPUT_HEADER');
```

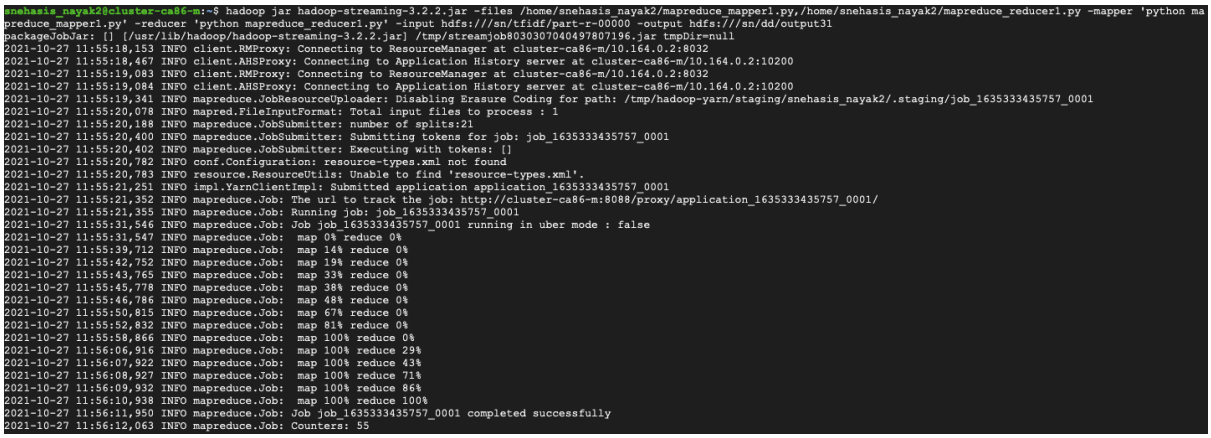
```
Input(s):
Successfully read 200001 records (243487029 bytes) from: "/sn/combined.csv"

Output(s):
Successfully stored 92 records (53456 bytes) in: "/sn/tfidf"

Counters:
Total records written : 92
Total bytes written : 53456
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

r
2021-10-25 22:16:51,407 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-ca86-m/10.164.0.2:8032
2021-10-25 22:16:51,408 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-ca86-m/10.164.0.2:10200
2021-10-25 22:16:51,412 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-25 22:16:51,448 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cluster-ca86-m/10.164.0.2:8032
2021-10-25 22:16:51,448 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cluster-ca86-m/10.164.0.2:10200
2021-10-25 22:16:51,452 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history serve
r
2021-10-25 22:16:51,483 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning FIELD_DISCARDED_TYPE_CONVERSION_FAILED 184 tim
e(s).
2021-10-25 22:16:51,483 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
```

The data was loaded into HDFS directory /sn/tfidf

```
snehasis_nayak2@cluster-ca86-m:~$ hadoop fs -ls /sn/
Found 2 items
-rw-r--r--   2 snehasis_nayak2 hadoop  243482205 2021-10-24 00:07 /sn/combined.csv
drwxr-xr-x   - snehasis_nayak2 hadoop          0 2021-10-25 22:16 /sn/tfidf
snehasis_nayak2@cluster-ca86-m:~$
```

The MapReduce function is executed in Python. There were 4 mapper functions and 3 reducer function to calculate the TF-IDF. Then I ran python code on mapper1 and reducer1 on the input data and the output was taken in as input in mapper2 and reducer 2 respectively.

Below is the screenshot of the action performed.



The final data was uploaded into the Hive using below command.

```
-load data inpath '/user/sn/dd/output4/part-0000*' into table
user_post_tfidf;
```

The final results of hive query results for the top 10 tfidf terms of top 10 users Is presented below.

```
OK
6068    0.034966        sys     1
6068    0.019426        aspx    2
6068    0.018131        block   3
6068    0.016836        border  4
6068    0.016836        puts    4
6068    0.014245        sqlconnection   6
6068    0.01295 books   7
6068    0.01295 separate        7
6068    0.01295 daily   7
6068    0.011655        nerddinner      10
7473    0.182383        reach   1
7473    0.10259 commands        2
7473    0.056995        jrschulz        3
7473    0.056995        rename  3
7473    0.045596        jscrollpane     5
7473    0.034197        radio   6
7473    0.023903        documentation   7
7473    0.022798        mac     8
7473    0.022798        getcolumn       8
7473    0.022798        checks  8
7473    0.022798        'ba     8
7473    0.022798        assemblystring  8
7473    0.022798        merging 8
7473    0.022798        functio 8
9951    0.034367        potentially     1
```