Task 4: Calculate TF-IDF with MapReduce/Pig/Hive

1. Creating a table with the columns required to calculate TF-IDF and putting it into the local.

   A. hive> create table user_db.grouped_users_posts as select owneruserid as a, body as b, SUM(score) as c from user_db.stackdata_analysis group by owneruserid, body;

   B. hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/hduser/tfidfdata'
   ROW FORMAT DELIMITED
   FIELDS TERMINATED BY ','
   select owneruserid, body from user_db.stackdata_analysis where owneruserid in (select id from user_db.grouped_users_posts);

2. The MapReduce program needs the input file to be separated by space rather than comma. Therefore, replacing the comma with spaces and then moving the file into a folder in the local from HDFS.

   A. > cd tfidfdata
   B. > sed 's/,/ /g' 000000_0 > inputfile

   C. > hadoop fs -mkdir /mappred
   D. > hadoop fs -put inputfile /mappred

3. Now, the Implementation of TF-IDF in Hadoop using Python will be in three phases using three mappers and three reducers. Using the following commands to run TF-IDF on the cluster.

   A. > hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -file /home/hduser/MAPPER_01.py /home/hduser/REDUCER_01.py -mapper "python MAPPER_01.py" -reducer "python REDUCER_01.py" -input hdfs://cluster-3299-m/mapinput/inputfile -output hdfs://cluster-3299-m/mappred1

   B. > hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -file /home/hduser/MAPPER_02.py /home/hduser/REDUCER_02.py -mapper "python MAPPER_02.py" -reducer "python REDUCER_02.py" -input hdfs://cluster-3299-m/mappred1/part-00000 hdfs://cluster-3299-m/mappred1/part-00001 hdfs://cluster-3299-m/mappred1/part-00002 hdfs://cluster-3299-m/mappred1/part-00003 hdfs://cluster-3299-m/mappred1/part-00004 -output hdfs://cluster-3299-m/mappred2

C. > hadoop jar /usr/lib/hadoop-mapreduce/hadoop-streaming.jar -file /home/hduser/MAPPER_03.py /home/hduser/REDUCER_03.py -mapper "python MAPPER_03.py" -reducer "python REDUCER_03.py" -input hdfs://cluster-3299-m/mappred2/part-00000 hdfs://cluster-3299-m/mappred2/part-00001 hdfs://cluster-3299-m/mappred2/part-00002 hdfs://cluster-3299-m/mappred2/part-00003 hdfs://cluster-3299-m/mappred2/part-00004 -output hdfs://cluster-3299-m/mappredf

4. Using the hadoop fs -getmerge command to merge the output files into a single CSV file (tfidfout.csv). Now, replacing the spaces and saving the file into another CSV file(tfidfout1.csv).

```
> sed -e 's/\s/,/g' tfidfout.csv > tfidfout1.csv
```

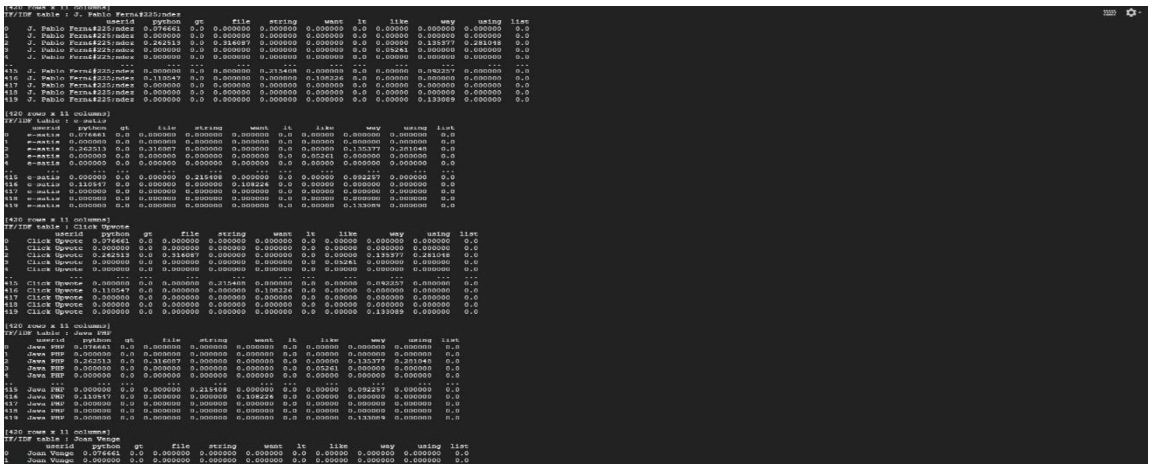
5. Thereafter, creating an external table in HIVE and loading the CSV file into the table.

A. hive> create external table if not exists TFIDF_data2 (Term String, Id int, tfidf float)
```
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
```

B. hive> load data local inpath 'tfidfout1.csv' overwrite into table TFIDF_data2;

6. Finally, running the following query to calculate the per-user TF-IDF of the top 10 terms for each of the top 10 users.

```
hive> SELECT *
FROM (
SELECT ROW_NUMBER()
OVER(PARTITION BY Id
ORDER BY tfidf DESC) AS TfidfRank, *
FROM TfIDF_data2) n
WHERE TfidfRank IN (1,2,3,4,5,6,7,8,9,10);
```



```
[420 rows x 11 columns]
TF/IDF table : J. Pablo Fern&#225;ndez
                   userid    python   gt      file    string      want   lt     like       way     using  list
0    J. Pablo Fern&#225;ndez  0.076661  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
1    J. Pablo Fern&#225;ndez  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
2    J. Pablo Fern&#225;ndez  0.262513  0.0  0.316087  0.000000  0.000000  0.0  0.00000  0.135377  0.281048   0.0
3    J. Pablo Fern&#225;ndez  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.05261  0.000000  0.000000   0.0
4    J. Pablo Fern&#225;ndez  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
..                       ...       ...  ...       ...       ...       ...  ...      ...       ...       ...   ...
415  J. Pablo Fern&#225;ndez  0.000000  0.0  0.000000  0.215408  0.000000  0.0  0.00000  0.092257  0.000000   0.0
416  J. Pablo Fern&#225;ndez  0.110547  0.0  0.000000  0.000000  0.108226  0.0  0.00000  0.000000  0.000000   0.0
417  J. Pablo Fern&#225;ndez  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
418  J. Pablo Fern&#225;ndez  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
419  J. Pablo Fern&#225;ndez  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.133089  0.000000   0.0

[420 rows x 11 columns]
TF/IDF table : e-satis
      userid    python   gt      file    string      want   lt     like       way     using  list
0    e-satis  0.076661  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
1    e-satis  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
2    e-satis  0.262513  0.0  0.316087  0.000000  0.000000  0.0  0.00000  0.135377  0.281048   0.0
3    e-satis  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.05261  0.000000  0.000000   0.0
4    e-satis  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
..       ...       ...  ...       ...       ...       ...  ...      ...       ...       ...   ...
415  e-satis  0.000000  0.0  0.000000  0.215408  0.000000  0.0  0.00000  0.092257  0.000000   0.0
416  e-satis  0.110547  0.0  0.000000  0.000000  0.108226  0.0  0.00000  0.000000  0.000000   0.0
417  e-satis  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
418  e-satis  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
419  e-satis  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.133089  0.000000   0.0

[420 rows x 11 columns]
TF/IDF table : Click Upvote
           userid    python   gt      file    string      want   lt     like       way     using  list
0    Click Upvote  0.076661  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
1    Click Upvote  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
2    Click Upvote  0.262513  0.0  0.316087  0.000000  0.000000  0.0  0.00000  0.135377  0.281048   0.0
3    Click Upvote  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.05261  0.000000  0.000000   0.0
4    Click Upvote  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
..            ...       ...  ...       ...       ...       ...  ...      ...       ...       ...   ...
415  Click Upvote  0.000000  0.0  0.000000  0.215408  0.000000  0.0  0.00000  0.092257  0.000000   0.0
416  Click Upvote  0.110547  0.0  0.000000  0.000000  0.108226  0.0  0.00000  0.000000  0.000000   0.0
417  Click Upvote  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
418  Click Upvote  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
419  Click Upvote  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.133089  0.000000   0.0

[420 rows x 11 columns]
TF/IDF table : Java PHP
       userid    python   gt      file    string      want   lt     like       way     using  list
0    Java PHP  0.076661  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
1    Java PHP  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
2    Java PHP  0.262513  0.0  0.316087  0.000000  0.000000  0.0  0.00000  0.135377  0.281048   0.0
3    Java PHP  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.05261  0.000000  0.000000   0.0
4    Java PHP  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
..        ...       ...  ...       ...       ...       ...  ...      ...       ...       ...   ...
415  Java PHP  0.000000  0.0  0.000000  0.215408  0.000000  0.0  0.00000  0.092257  0.000000   0.0
416  Java PHP  0.110547  0.0  0.000000  0.000000  0.108226  0.0  0.00000  0.000000  0.000000   0.0
417  Java PHP  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
418  Java PHP  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
419  Java PHP  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.133089  0.000000   0.0

[420 rows x 11 columns]
TF/IDF table : Joan Venge
         userid    python   gt      file    string      want   lt     like       way     using  list
0    Joan Venge  0.076661  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
1    Joan Venge  0.000000  0.0  0.000000  0.000000  0.000000  0.0  0.00000  0.000000  0.000000   0.0
```