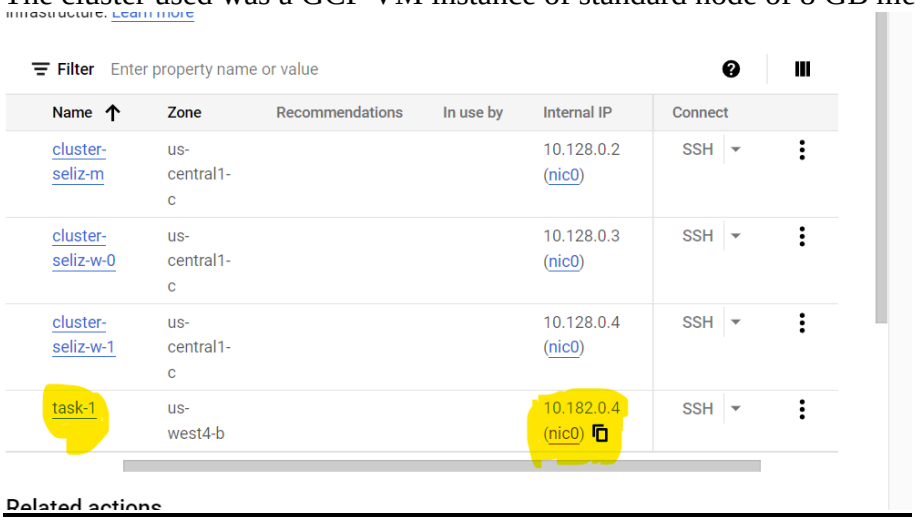
CLOUD TECHNOLOGIES - 1

1. The aim of this assignment is to design and develop a simple and rudimentary spam detection system using the following technologies:
   - Cloud Infrastructure using AWS, GCP or other industry-standard public cloud
   - Hadoop
   - MapReduce
   - Hive
   - Pig
2. **<u>TASK 1</u>**

   2.1. A project was created under the ancestory mail.dcu.ie named 'CA675 -Assignment 1'.

   2.2. This task was completed using the cluster 'task-1' with internal IP address 10.182.0.4. The cluster used was a GCP VM instance of standard node of 8 GB memory.



Filter Enter property name or value

Name	Zone	Recommendations	In use by	Internal IP	Connect
cluster-seliz-m	us-central1-c			10.128.0.2 (nic0)	SSH
cluster-seliz-w-0	us-central1-c			10.128.0.3 (nic0)	SSH
cluster-seliz-w-1	us-central1-c			10.128.0.4 (nic0)	SSH
task-1	us-west4-b			10.182.0.4 (nic0)	SSH

   2.3. Related actions

   2.4. Used the below command to install default jdk **sudo apt install default-jdk** and the latest java version 11 was installed.

   2.5. Created a user named **hadoop.**

   2.6. Now we also need to install ssh key's that is secured shell.
   sudo apt-get install openssh-server

   2.7. Login to the user **hadoop** and set its password as **password**
   su – hadoop

   2.8. Uploaded the Hadoop version from the site and extract the file using tar.
   wget https://archive.apache.org/dist/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz
   tar -xzvf hadoop-3.3.4.tar.gz

   2.9. Created a directory to save the Hadoop in the directory **/opt/hadoop.**
   sudo mkdir /opt/Hadoop
   Moved the extracted file into the Hadoop directory
   sudo mv hadoop-3.3.4/* /opt/hadoop

   2.10. Set the Hadoop path using the below commands

export JAVA_HOME=/usr
export HADOOP_HOME=/opt/hadoop
export
PATH=$PATH:$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME

export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_INSTALL=$HADOOP_HOME

2.11. To set the env variables we use the below command.
nano ~/.bashrc

2.12. Next we run the command **sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh** to run the file and set the user variables using **export JAVA_HOME=/usr**

2.13. Before we configure the Hadoop file we need to create responding name and data directorys.
mkdir /home/hadoop/hdfs/name
mkdir /home/hadoop/hdfs/data

2.14. Now we will configure the *hdfs-site.xml* for that open that file using below command.
sudo nano /opt/hadoop/etc/hadoop/hdfs-site.xml

Once the file opens copy the below text inside the configuration tag.

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:/home/hadoop/hdfs/name</value>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>file:/home/hadoop/hdfs/data</value>

</property>

2.15. Now we will configure the *yarn-site.xml* for that open that file using below command.

sudo nano /opt/hadoop/etc/hadoop/yarn-site.xml

Once the file opens copy the below text inside the configuration tag

<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>

2.16. Now we will configure the *mapred-site.xml*for that open that file using below command.

sudo nano /opt/hadoop/etc/hadoop/ mapred-site.xml

Once the file opens copy the below text inside the configuration tag

<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>

2.17. To run Hadoop first we need to format the namenode then you need to run the below command for first time when you starting the cluster if you use it again then all your metadata will get erase.

hdfs namenode –format

2.18. Now we need to start the DFS i.e. Distributed File System.

start-dfs.sh
start-yarn.sh

and run the following command **jps,** Now Hadoop is installed in the system.

2.19. Uploaded the Hive version from the site and extract the file using tar.'

wget https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz
tar xzf apache-hive-3.1.2-bin.tar.gz

2.20. Created a directory to save the hive in the directory **/opt/hive.**

sudo mkdir /opt/hive

Moved the extracted file into the Hive directory

sudo mv apache-hive-3.1.2-bin/* /opt/hive

2.21. To set the env variables we use the below command.

nano ~/.bashrc

2.22. Set the Hive path using the below commands

export HIVE_HOME=/opt/hive
export PATH=$PATH:$HIVE_HOME/bin

2.23. To set the env variables we use the below command.

nano ~/.bashrc

2.24. Next we run the command **sudo nano $HIVE_HOME/bin/hive-config.sh** to run the file and set the user variables using **export JAVA_HOME= /opt/hive**

2.25. The next command is run
$HIVE_HOME/bin/schematool -dbType derby –initSchema

2.26. On running the command **hive** we see that hive is installed,

2.27. Uploaded the Pig version from the site and extract the file using tar.
wget https://downloads.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz
tar -xzf pig-0.17.0.tar.gz

2.28. Created a directory to save the pig in the directory **/opt/pig.**
sudo mkdir /opt/pig
Moved the extracted file into the pig directory
sudo mv ./pig-0.17.0/* /opt/pig

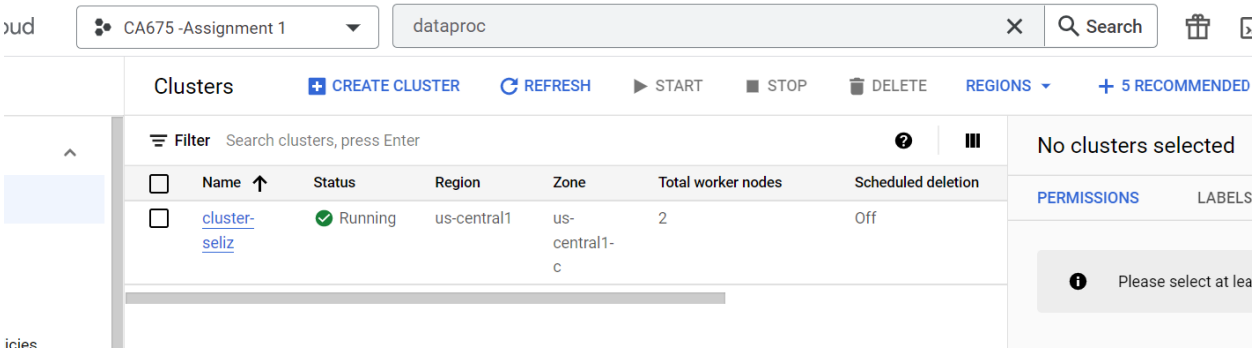
2.29. To set the env variables we use the below command.
nano ~/.bashrc

2.30. On running the command pig we see that it is installed.

2.31.

3. **<u>TASK 2</u>**
I have used an amazon reviews dataset for this assignment. The dataset is acquired from Kaggle (https://www.kaggle.com/datasets/naveedhn/amazon-product-review-spam-and-non-spam) and the used dataset is Cell_Phones_and_Accessories.json. Since this dataset is in json format I converted it into a csv file saved as amazon_review.csv, choosing only required columns i.e, reviewerID, reviewText, category and class using pandas. The file in which this work is done is in Assignment_1.ipynb in the git repo. After converting into csv the file size is of 1.1GB. I have converted the reviewtext column into lowercase string before converting it into csv file.

4. After that, the dataset is uploaded manually to Google Cloud Platform in the created project CA675 -Assignment 1.



Using DataProc in GCP, a Hadoop cluster called 'cluster-seliz' was created. Following that, the CSV file was uploaded to the cluster's name node machine (cluster-seliz-m).

Then, a new folder was created using the command **'hadoop fs -mkdir -p /user/cluster-seliz-m/newcsv/**' and the csv file was moved into this directory using the command '**hadoop fs -cp gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg/google-cloud-dataproc-metainfo/842f89cf-30e0-405a-9452-1ea0f7eedea3/cluster-seliz-m/Assignment1/amazon_reviews.csv cluster-seliz-m/newcsv/amazon_reviews.csv**'.

```
Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Mon Oct 31 14:16:16 2022 from 35.235.241.18
seliz_koshy2@cluster-seliz-m:~$ hdfs dfs -ls -R
drwxr-xr-x   - seliz_koshy2 hadoop          0 2022-10-31 15:07 cluster-seliz-m
drwxr-xr-x   - seliz_koshy2 hadoop          0 2022-10-31 15:33 cluster-seliz-m/newcsv
-rw-r--r--   2 seliz_koshy2 hadoop 1089095034 2022-10-31 15:33 cluster-seliz-m/newcsv/amazon_reviews.csv
drwxr-xr-x   - seliz_koshy2 hadoop          0 2022-10-31 14:38 new_csv
seliz_koshy2@cluster-seliz-m:~$
```

5. Next I am going to be using hive to create a table and perform cleaning operations as hive. The reason I chose hive over pig was that it has a SQL-like Query Language and can be used to analyze large and complex datasets easily.

   Before creating the hive table we have to create a database to store the table. So we initiate hive by using hive command. Then a database named selizdb is created and inside that a table was created named amazon1 using the below query :

   hive>CREATE TABLE IF NOT EXISTS amazon1 (reviewerID String, reviewText String,category String,class float)
   row format delimited fields terminated by ',';

```
hive> use reviewsdb;
OK
Time taken: 0.656 seconds
hive> CREATE TABLE IF NOT EXISTS amazon1 ( reviewerID String, reviewText String,category String,class float)
    > row format delimited fields terminated by ',';
OK
Time taken: 0.567 seconds
hive> desc amazon1;
OK
reviewerid              string
reviewtext              string
category                string
class                   float
Time taken: 0.187 seconds, Fetched: 4 row(s)
hive> SELECT * FROM amazon1 LIMIT 5;
Query ID = seliz_koshy2_20221106134439_e76c9afd-d713-467f-b16c-e96c364979bf
Total jobs = 1
Launching Job 1 out of 1
--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1            container     SUCCEEDED      0          0        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 00/01  [>>-------------------------] 0%    ELAPSED TIME: 0.78 s
--------------------------------------------------------------------------------------------
OK
Time taken: 10.235 seconds
hive> load data inpath 'cluster-seliz-m/newcsv/amazon_reviews.csv' overwrite into table amazon1;
Loading data to table reviewsdb.amazon1
OK
Time taken: 0.482 seconds
```

   To load the table with data I used the below query

   load data inpath 'cluster-seliz-m/newcsv/amazon_reviews.csv' overwrite into table amazon1;

6. **<u>TASK 3</u>**

   To further clean the dataset I have removed the punctuations present in the reviewtext column by saving the entire table into new table processed_data;

   Query:

   hive>create table processed_data as select reviewerID, REGEXP_REPLACE(reviewtext, '[^0-9A-Za-z]+', '') as reviewtext,category,class from amazon1;

Since the dataset is based on texts and has already been converted to lowercase there is no further processing to be done.

```
hive> create table processed_data as select reviewerID, REGEXP_REPLACE(reviewtext, '[^0-9A-Za-z ]+', '') as reviewtext,category,class from amazonl;
Query ID = seliz_koshy2_20221106140503_965ce30b-4af6-484f-8043-c1d80380a811
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_1667737918709_0030)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11          11        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 22.02 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-seliz-m/user/hive/warehouse/reviewsdb.db/processed_data
OK
Time taken: 30.889 seconds
hive> desc processed_data;
OK
reviewerid              string
reviewtext              string
category                string
class                   string
Time taken: 0.062 seconds, Fetched: 4 row(s)
```

7. **TASK 4**

Next the processed_data table is split into two separate datasets: one ham dataset and one spam dataset.

To obtain the spam dataset I used the below query to create the table spam_data.

QUERY :

hive>create table spam_data as select * from processed_data where class=1.0;

```
hive> create table spam_data as select * from processed_data where class=1.0;
Query ID = seliz_koshy2_20221106141215_dd7adff8-5478-4ac9-9d92-a903cfe94295
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667737918709_0030)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED     11          11        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 16.92 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-seliz-m/user/hive/warehouse/reviewsdb.db/spam_data
OK
Time taken: 18.144 seconds
hive> select * from spam_data limit 5;
Query ID = seliz_koshy2_20221106141340_789d3ab3-0f79-4833-b98d-7c33100e1219
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667737918709_0030)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8           8        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 12.00 s
--------------------------------------------------------------------------------
OK
A9G5UGDP9D0IC   these clips are way better  then the one sprint supplies  suck these  work  out much nice  and they work really great   Cell_Phones_and_Acce
ssories 1.0
AV5H8F4LQU50K   good item great to have very durable and dependable i love this product thank you and have a good day   Cell_Phones_and_Accessories     1.0
A3KJLJV2HHFKOG  beautiful colors and sturdy case easy on and easy off order processed quickly and product was shipped securely with no problems Cell_Phones_
and_Accessories 1.0
```

To find the top 10 spam accounts based on the number of spam reviews present in the dataset. I used the below query:

hive>SELECT reviewerID,reviewtext FROM spam_data ORDER BY reviewerID DESC LIMIT 10;

```
hive> SELECT DISTINCT reviewerID,reviewtext FROM spam_data DESC LIMIT 10;
Query ID = seliz_koshy2_20221107002132_136155e7-de58-41d4-ae84-d6d923380b4e
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667774375782_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 14.47 s
--------------------------------------------------------------------------------
OK
A0000488123JA1KQJTEM8   keeps my phone safe and sound when i drop it im a clumsy person when it comes to phones hehe
A00005303588WHRQZ6N4L   stunning look and i need a case that showcase the complete look of my iphone beautifulthis case suites me the best in looks
A00025741CVPCXCF3NHMR   much easier than dealing with verizon directly in a store they also wanted more money upfront  rebate to mess with this was much bet
ter
A000285218JCFNDXRN02X   wonderful little case fit perfect on my phone just stretched the case a little bit and it fits snug onto my phone  i love the neon g
reen color thanks so much
A00028781NF0U7YEN9U19   just what my daughter needed to use for her ipod
A00037441I8XOQJSUWCAG   i bought this for my samsung replenish syracuse my old battery wasnt working any more  this battery works great and lasts almost a w
hole day
A00044782UB564I4SGA0X   it is really good but it was hard to push out all the air bubbles but that is not a big problem
A00061202IT7XNIEW32MA   i love it it works amazing and it is very durable it does exactly what it says i like that it takes the same charger as my iphone 5c
A00062163COOP0O1EIMZL   it is agood car holder recomend u to buy it but moves alot while driving but generaly good stuff and the shipping was great i got it
 after 2 weeks not a month thank u
A00062283LKXEZFY9NQ8B   i had to be very carefull installing it but every thing went well i am very happy with the product and give this one thumbs up
Time taken: 15.644 seconds, Fetched: 10 row(s)
```

To obtain the ham dataset I used the below query to create the table ham_data.

QUERY :

hive>create table ham_data as select * from processed_data where class=0.0;

```
hive> SELECT DISTINCT reviewerID,reviewtext FROM ham_data DESC LIMIT 10;
Query ID = seliz_koshy2_20221107002305_73096de5-e274-4059-a8f4-60bf1a3dd0d2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667774375782_0005)

--------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.71 s
--------------------------------------------------------------------------------
OK
A00008281Z7EL3AO1UQD7   cute
A000123613N4QBI2JI0JD   the case feels a bit cheap but overall its a nice case a but hard to use when a call comes in but if you usually use headsets it won
t be a problem
A0002444350T5FBABPR4J   i was searching through amazon looking for a new case because i was tired of my old otterbox that was getting washed up i stumbled u
pon this one and thought it would be a great fit it came in the mail in about 5 days and i was excited to start using itat first i was very pleased with the
 case it looked sleek and it was protective i got the navy blue matte finish one for the iphone 4 the adapter hole to charge the phone was a little to small
 the adapter would fit but it would fall out if you were using the phone which became frustrating really fast but the major problem was that the corners of
the matte finish on the back of the case started to peel off this started about two weeks after i started to use it
A00026861LL8FKLSE1O0S   i liked it but it didt protect my screen from cracking so it was very disappointing  i wouldnt recommend it
A00026861LL8FKLSE1O0S   very poor quality it comes in two parts and it doesnt stay together  this case looks nice online but when it came to me it was very
flimsy
A0002730WOKVUCGRLYJU    the adapter snapped as soon as i put my phone to charge the headphone cable is extremely longu get 0000 stars dont buy this shit peo
ple
A000443821AD43TOGKNGZ   i chose this rating because it was hard to get my phone into the case and i broke a nail because of it
A00057621NM70611JIO97   the cords are cheap and it shows why the charger goes in the phone rough and it comes out worse the cords also break very easy
A0006066105Y63JM18F33   this product was very well made and will last forever but it just doesnt work as a belt clipdamages dress belts with minimum usenot
secure on wide beltsscreen faces out so it is exposed to possible damage and nosy peoplei am very impressed with all the other products from this company bu
t they need to put more through into a belt clip design
A00062283LKXEZFY9NQ8B   it is good for me to recharge the battery in my cell phone maybe it works in other cell phone but not mine
Time taken: 12.988 seconds, Fetched: 10 row(s)
```

To find the top 10 spam accounts based on the number of spam reviews present in the dataset. I used the below query:

hive>SELECT reviewerID,reviewtext FROM ham_data ORDER BY reviewerID DESC LIMIT 10;


```
hive> SELECT DISTINCT reviewerID,reviewtext FROM ham_data DESC LIMIT 10;
Query ID = seliz_koshy2_20221107002305_73096de5-e274-4059-a8f4-60bf1a3dd0d2
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667774375782_0005)

--------------------------------------------------------------------------------
        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
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 11.71 s
--------------------------------------------------------------------------------
OK
A00008281Z7EL3AO1UQD7   cute
A000123613N4QBI2JI0JD   the case feels a bit cheap but overall its a nice case a but hard to use when a call comes in but if you usually use headsets it won
t be a problem
A0002444350T5FBABPR4J   i was searching through amazon looking for a new case because i was tired of my old otterbox that was getting washed up i stumbled u
pon this one and thought it would be a great fit it came in the mail in about 5 days and i was excited to start using itat first i was very pleased with the
 case it looked sleek and it was protective i got the navy blue matte finish one for the iphone 4 the adapter hole to charge the phone was a little to small
 the adapter would fit but it would fall out if you were using the phone which became frustrating really fast but the major problem was that the corners of
the matte finish on the back of the case started to peel off this started about two weeks after i started to use it
A00026861LL8FKLSE1O0S   i liked it but it didt protect my screen from cracking so it was very disappointing  i wouldnt recommend it
A00026861LL8FKLSE1O0S   very poor quality it comes in two parts and it doesnt stay together  this case looks nice online but when it came to me it was very
flimsy
A0002730WOKVUCGRLYJU    the adapter snapped as soon as i put my phone to charge the headphone cable is extremely longu get 0000 stars dont buy this shit peo
ple
A000443821AD43TOGKNGZ   i chose this rating because it was hard to get my phone into the case and i broke a nail because of it
A00057621NM70611JIO97   the cords are cheap and it shows why the charger goes in the phone rough and it comes out worse the cords also break very easy
A0006066105Y63JM18F33   this product was very well made and will last forever but it just doesnt work as a belt clipdamages dress belts with minimum usenot
secure on wide beltsscreen faces out so it is exposed to possible damage and nosy peoplei am very impressed with all the other products from this company bu
t they need to put more through into a belt clip design
A00062283LKXEZFY9NQ8B   it is good for me to recharge the battery in my cell phone maybe it works in other cell phone but not mine
Time taken: 12.988 seconds, Fetched: 10 row(s)
```

7. Calculation of TF-IDF top 10 spam/ham keywords for each top 10 spam/ham accounts.

```
hive> create table top10spamdata as SELECT DISTINCT reviewerID,reviewtext FROM spam_data DESC LIMIT 10;
Query ID = seliz_koshy2_20221107002651_a00bab78-f4a3-4227-8ca1-783a8e18f908
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667774375782_0005)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      8          8        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reduce 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.05 s
--------------------------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-seliz-m/user/hive/warehouse/reviewsdb.db/top10spamdata
OK
Time taken: 17.174 seconds
hive> create table top10hamdata as SELECT DISTINCT reviewerID,reviewtext FROM ham_data DESC LIMIT 10;
Query ID = seliz_koshy2_20221107002712_f3b22387-33ad-49b2-a3aa-9686b908fb86
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667774375782_0005)

--------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      3          3        0        0       0       0
Reduce 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reduce 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 1.84 s
--------------------------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-seliz-m/user/hive/warehouse/reviewsdb.db/top10hamdata
OK
Time taken: 3.29 seconds
```

i. To calculate the TF-IDF of the top 10 spam keywords for each top 10 spam accounts a table was created containing the top 10 spam accounts and its reviews as shown in the picture above.

   QUERY: **hive>create table top10spamdata as SELECT DISTINCT reviewerID,reviewtext FROM spam_data DESC LIMIT 10;**

ii. Similarly to calculate top 10 ham keywords for each top 10 ham accounts another table was created containing the top 10 spam accounts and its reviews as shown in the picture above.

   QUERY: **hive>create table top10hamdata as SELECT DISTINCT reviewerID,reviewtext FROM ham_data DESC LIMIT 10;**

iii. Next, the top10spamdata and top10hamdata is stored into HDFS in the location 'hdfs://cluster-seliz-m/newcsv/'.

   QUERY: hive>insert overwrite directory 'hdfs://cluster-seliz-m/newcsv/top10spamdata' row format delimited fields terminated by ',' select * from top10spamdata;

   QUERY: hive>insert overwrite directory 'hdfs://cluster-seliz-m/newcsv/top10hamdata' row format delimited fields terminated by ',' select * from top10hamdata;

```
hive> insert overwrite directory 'hdfs://cluster-seliz-m/newcsv/top10spamdata' row format delimited fields terminated by ',' select * from top10spamdata;
Query ID = seliz_koshy2_20221107002823_d27605ac-b9b8-437c-9fd5-01c410d0f7b7
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667774375782_0005)

--------------------------------------------------------------------------------
        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: 4.86 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-seliz-m/newcsv/top10spamdata
OK
Time taken: 5.828 seconds
hive> insert overwrite directory 'hdfs://cluster-seliz-m/newcsv/top10hamdata' row format delimited fields terminated by ',' select * from top10hamdata;
Query ID = seliz_koshy2_20221107002837_1d7d1d91-679d-4a92-818e-fa639f485f45
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667774375782_0005)

--------------------------------------------------------------------------------
        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: 3.93 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://cluster-seliz-m/newcsv/top10hamdata
OK
Time taken: 4.828 seconds
```

iv. Currently our data is stored in the above mentioned HDFS location and we have to obtain this data so we can perform mapreduce on them.For that I copied the data into the gcp bucket by specifying its gsutil URL and renamed the files to spam_text_files and ham_text_files respectively.

QUERY for spam_data :
hadoop fs -cp hdfs://cluster-seliz-m/newcsv/top10spamdata/000000_0 gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg

QUERY for ham_data :
hadoop fs -cp hdfs://cluster-seliz-m/newcsv/top10hamdata/000000_0 gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg

```
seliz_koshy2@cluster-seliz-m:~$ hadoop fs -cp hdfs://cluster-seliz-m/newcsv/top10spamdata/000000_0 gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg
```

```
seliz_koshy2@cluster-seliz-m:~$ hadoop fs -cp hdfs://cluster-seliz-m/newcsv/top10hamdata/000000_0 gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg
```

NOTE: The data was stored in the path hdfs://cluster-seliz-m/newcsv/<dataset>/000000_0 as shown below hence the CLI commands above appears as so.

```
seliz_koshy2@cluster-seliz-m:~$ hdfs dfs -ls hdfs://cluster-seliz-m/newcsv/top10hamdata
Found 1 items
-rw-r--r--   2 seliz_koshy2 hadoop       2151 2022-11-07 00:28 hdfs://cluster-seliz-m/newcsv/top10hamdata/000000_0
[5]+  Killed                  hive
seliz_koshy2@cluster-seliz-m:~$ hdfs dfs -ls hdfs://cluster-seliz-m/newcsv/top10spamdata
Found 1 items
-rw-r--r--   2 seliz_koshy2 hadoop       1445 2022-11-07 00:28 hdfs://cluster-seliz-m/newcsv/top10spamdata/000000_0
```

v. To obtain the individual records from the dataset, we need to split the results into 10 text files containing each single record of reviewerID and reviewText. Using the jyupter file ‘tfdif.ipynb’ I was able to split the 10 user files into individual text files. The TF-IDF program then used these 10 files as input.

TASK 5

Now we begin to use the mapreduce programs by creating a directory **assignment** using the command below to store the map and reduce python programs.

QUERY: **hadoop fs -mkdir -p /assignment**

Loaded the input files into the directory as well as the map and reduce programs using the below commands

QUERY for loading map and reduce programs from the gcp bucket:
sudo gsutil cp -r gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg/google-cloud-dataproc-metainfo/842f89cf-30e0-405a-9452-1ea0f7eedea3/cluster-seliz-m/Assignment1/Mapper_programs/* ./assignment

QUERY for loading spam input text files:
sudo gsutil cp -r gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg/google-cloud-dataproc-metainfo/842f89cf-30e0-405a-9452-1ea0f7eedea3/cluster-seliz-m/Assignment1/spam_text_files/* ./assignment

QUERY for loading ham input text files:
sudo gsutil cp -r gs://dataproc-staging-us-central1-1031731324489-ozzu0ijg/google-cloud-dataproc-metainfo/842f89cf-30e0-405a-9452-1ea0f7eedea3/cluster-seliz-m/Assignment1/ham_text_files/* ./assignment

```
seliz_koshy2@cluster-seliz-m:~$ ls ./assignment/
A0000488123JA1KQJTEM8.txt  A00025741CVPCXCF3NHMR.txt  A00037441I8XOQJSUWCAG.txt  A00061202IT7XNIEW32MA.txt  mapper_3.py
A00005303588WHRQZ6N4L.txt  A00026861LL8FKLSEl0OS.txt  A000443821AD43TOGKNGZ.txt  A00062163COOP0O1EIMZL.txt  mapper_4.py
A000008281Z7EL3AO1UQD7.txt  A0002730WOKVUCGRLYJU.txt   A00044782UB564I4SGA0X.txt  A00062283LKXEZFY9NQ8B.txt  reducer_1.py
A000123613N4QBI2JI0JD.txt  A000285218JCFNDXRN02X.txt  A00057621NM70611JIO97.txt  mapper_1.py                reducer_2.py
A000244435OT5FBABPR4J.txt  A00028781NF0U7YEN9U19.txt  A0006066l05Y63JM18F33.txt  mapper_2.py                reducer_3.py
```

To run the map and reduce python programs for the individual text files I use the below sample command in the format.
hadoop jar /usr/lib/hadoop/hadoop-streaming-3.2.3.jar -file ./assignment/mapper_1.py -mapper "python mapper_1.py" -file ./assignment/reducer_1.py -reducer "python reducer_1.py" -input /user/seliz_koshy2/cluster-seliz-m/MapReduce/A0000488123JA1KQJTEM8.txt -output ./user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1

```
seliz_koshy2@cluster-seliz-m:~$ hadoop  jar /usr/lib/hadoop/hadoop-streaming-3.2.3.jar -file ./assignment/mapper_1.py -mapper "python mapper_1.py" -file ./a
ssignment/reducer_1.py -reducer "python reducer_1.py" -input /user/seliz_koshy2/cluster-seliz-m/MapReduce/A0000488123JA1KQJTEM8.txt -output ./user/seliz_kos
hy2/cluster-seliz-m/MapReduce/output_1
2022-11-07 16:55:27,278 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
packageJobJar: [./assignment/mapper_1.py, ./assignment/reducer_1.py] [/usr/lib/hadoop/hadoop-streaming-3.2.3.jar] /tmp/streamjob1312786492357119190.jar tmpD
ir=null
2022-11-07 16:55:28,751 INFO client.RMProxy: Connecting to ResourceManager at cluster-seliz-m/10.128.0.2:8032
2022-11-07 16:55:29,016 INFO client.AHSProxy: Connecting to Application History server at cluster-seliz-m/10.128.0.2:10200
2022-11-07 16:55:29,667 INFO client.RMProxy: Connecting to ResourceManager at cluster-seliz-m/10.128.0.2:8032
2022-11-07 16:55:29,668 INFO client.AHSProxy: Connecting to Application History server at cluster-seliz-m/10.128.0.2:10200
2022-11-07 16:55:29,894 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/seliz_koshy2/.staging/job_1667838324
597_0002
2022-11-07 16:55:30,309 WARN concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,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:750)
2022-11-07 16:55:30,310 INFO mapred.FileInputFormat: Total input files to process : 1
2022-11-07 16:55:30,405 INFO mapreduce.JobSubmitter: number of splits:24
2022-11-07 16:55:30,679 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1667838324597_0002
2022-11-07 16:55:30,682 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-11-07 16:55:30,934 INFO conf.Configuration: resource-types.xml not found
2022-11-07 16:55:30,934 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-11-07 16:55:31,380 INFO impl.YarnClientImpl: Submitted application application_1667838324597_0002
2022-11-07 16:55:31,493 INFO mapreduce.Job: The url to track the job: http://cluster-seliz-m:8088/proxy/application_1667838324597_0002/
2022-11-07 16:55:31,501 INFO mapreduce.Job: Running job: job_1667838324597_0002
2022-11-07 16:55:42,783 INFO mapreduce.Job: Job job_1667838324597_0002 running in uber mode : false
2022-11-07 16:55:42,784 INFO mapreduce.Job:  map 0% reduce 0%
2022-11-07 16:55:50,878 INFO mapreduce.Job:  map 8% reduce 0%
2022-11-07 16:55:51,883 INFO mapreduce.Job:  map 13% reduce 0%
2022-11-07 16:55:56,915 INFO mapreduce.Job:  map 29% reduce 0%
2022-11-07 16:55:57,920 INFO mapreduce.Job:  map 42% reduce 0%
2022-11-07 16:56:04,963 INFO mapreduce.Job:  map 54% reduce 0%
```

The output file is stored in the output folder and viewed using the command
hdfs dfs -ls ./user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1

```
seliz_koshy2@cluster-seliz-m:~$ hdfs dfs -ls ./user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1
Found 8 items
-rw-r--r--   2 seliz_koshy2 hadoop          0 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/_SUCCESS
-rw-r--r--   2 seliz_koshy2 hadoop        100 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00000
-rw-r--r--   2 seliz_koshy2 hadoop        323 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00001
-rw-r--r--   2 seliz_koshy2 hadoop          0 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00002
-rw-r--r--   2 seliz_koshy2 hadoop        101 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00003
-rw-r--r--   2 seliz_koshy2 hadoop        302 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00004
-rw-r--r--   2 seliz_koshy2 hadoop        304 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00005
-rw-r--r--   2 seliz_koshy2 hadoop          0 2022-11-07 16:56 user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00006
seliz_koshy2@cluster-seliz-m:~$ hdfs dfs -cat user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1
```

The words in the text file are counted and displayed in the format
Word | input_text_file | count respectively.

```
seliz_koshy2@cluster-seliz-m:~$ hdfs dfs -cat ./user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00000
hehe hdfs://cluster-seliz-m/user/seliz_koshy2/cluster-seliz-m/MapReduce/A0000488123JA1KQJTEM8.txt       1
seliz_koshy2@cluster-seliz-m:~$ hdfs dfs -cat ./user/seliz_koshy2/cluster-seliz-m/MapReduce/output_1/part-00001
a0000488123ja1kqjtem8,keeps hdfs://cluster-seliz-m/user/seliz_koshy2/cluster-seliz-m/MapReduce/A0000488123JA1KQJTEM8.txt
im hdfs://cluster-seliz-m/user/seliz_koshy2/cluster-seliz-m/MapReduce/A0000488123JA1KQJTEM8.txt 1
phones hdfs://cluster-seliz-m/user/seliz_koshy2/cluster-seliz-m/MapReduce/A0000488123JA1KQJTEM8.txt     1
seliz_koshy2@cluster-seliz-m:~$
```

The map and reduce program was invoked and tfdif was calculated.
Finally, the output is stored as a text file and the resultant screenshots of the above task can be found