 # commands for setting the permissions on a private key file to ensure it's secure, and then using that private key to establish an SSH connection

chmod 600 labsuser.pem

ssh -i labsuser.pem hadoop@ec2-3-86-185-26.compute-1.amazonaws.com

#Creating a Directory

[hadoop@ip-172-31-80-141 ~]$ mkdir csv

[hadoop@ip-172-31-80-141 ~]$ cd csv

#Copying the file from the S3 location to the local file System assignment.csv

[hadoop@ip-172-31-80-141 csv]$ aws s3 cp "s3://youtubespam01/Youtube.csv" assignment.csv

download: s3://youtubespam01/Youtube.csv to ./assignment.csv

[hadoop@ip-172-31-80-141 csv]$ ls

assignment.csv

 # Copying the Local file to HDFS directory assignment

[hadoop@ip-172-31-80-141 csv]$ hdfs dfs -mkdir /cloudassignment

[hadoop@ip-172-31-80-141 csv]$ hdfs dfs -copyFromLocal './assignment.csv' /cloudassignment

#Accessing Pig

[hadoop@ip-172-31-80-141 csv]$ pig -x MapReduce

#Spam and Ham Dataset creation

grunt> maintable = LOAD '/cloudassignment/assignment.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'YES_MULTILINE', 'UNIX', 'SKIP_INPUT_HEADER') AS (Name:chararray, Comment:chararray);

grunt>dump maintable;

#Cleaning using TRIM and FILTER functions

grunt> trimedtable = FOREACH maintable GENERATE TRIM(Name) as Name, TRIM(Comment) as Comment;

=====

trimedtable= FILTER maintable BY NOT (Name matches '.*\\?.*' or Comment matches '.*\\?.*');

#New Dataset that contains Clean Data

spamandham= FOREACH trimedtable GENERATE Name, Comment, (CASE WHEN Comment MATCHES
'.*(pokemon|stanford|legal|youtube|god|loan|ff|whatsapp|investment|subscribe|out).*' THEN 'spam' ELSE 'ham' END) AS predicted_label;

#Creating Spam Dataset:

spamdata= FILTER spamandham BY predicted_label == 'spam';

#Creating Ham Dataset

hamdata= FILTER spamandham BY predicted_label == 'ham';

#Codes to check the spamdata and hamdata tables

dump spamdata;

dump hamdata;

#Top 10 Spam Accounts

grunt> spamdata2= GROUP spamdata BY Name;

grunt> spamcount = FOREACH spamdata2 GENERATE group AS NAme, COUNT(spamdata) AS spamcount;

grunt> mainspamdata= ORDER spamcount BY spamcount DESC;

grunt> top10spam = LIMIT mainspamdata 10;

grunt> dump top10spam;

Results:

(Sunday karisai,6)

(Dineo O'Brien,4)

(Fernando Joseph,4)

(Wisdom Great,3)

(Scott Brooklyn,3)

(MAHMOUD AHMED,3)

(chentradingi.. on telegram,3)

(MARK GOLDSMITH,3)

(Angela Bella Gill,3)

(Kroll Brian,3)

==========

#Top 10 Ham Accounts

grunt> hamdata2= GROUP hamdata BY Name;

grunt> hamcount = FOREACH hamdata2 GENERATE group AS NAme, COUNT(hamdata) AS hamcount;

grunt> mainhamdata= ORDER hamcount BY hamcount DESC;

grunt> top10ham = LIMIT mainhamdata 10;

grunt> dump top10ham;

Results:

(Charlie Steve,5)

(Robert Spear,4)

(INVEST WITH LEGITHACKS7 ON TELEGRAM,4)

(CONTACT CRYPTOCURRENCYKEY ON TELEGRAM,4)

(Jack Nathaniel,3)

(BLACK DIAMOND,3)

(Zaynab Hussain,3)

(Andy Steele,3)

(Amanda De Wet,3)

(Lopez Finley,3)

==========