CA675 Cloud Technologies

Assignment 1: Individual Assignment

Spam Detection System

Submitted by: **Radhika Vyas**

Student Number: **22260148**

Student Email: **radhika.vyas2@mail.dcu.ie**

M.Sc. Computing majors in Data Analytics

Submitted to: **Manoj Kesavulu**

LINK FOR GIT REPOSITORY

"https://github.com/radhikavyas12/Assignment-1-Cloud"

LINK FOR CLOUD Project

https://console.cloud.google.com/compute/instances?authuser=1&project=sample-project-366621

Difference between SPAM and HAM

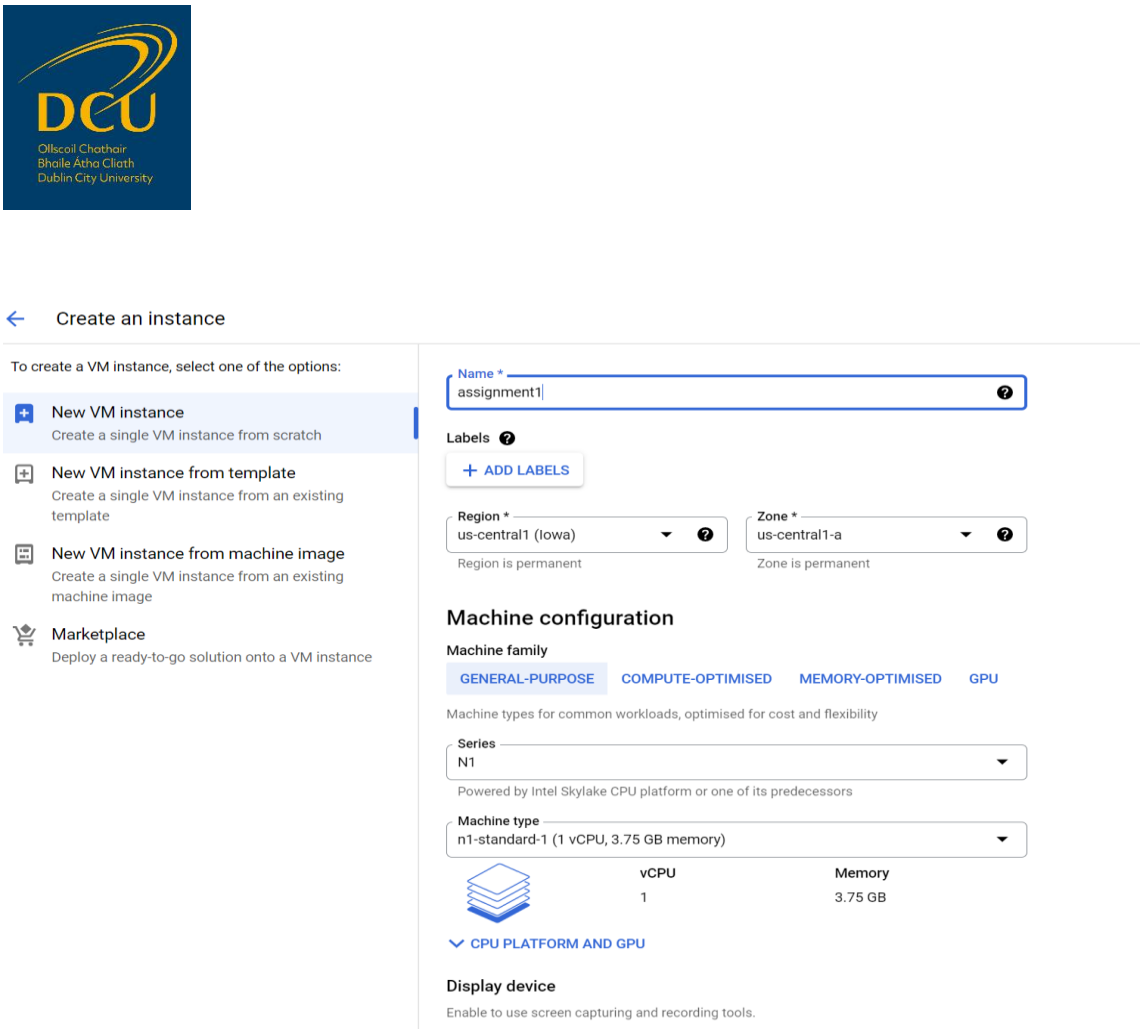
A spam is basically undesirable bulk messages, it can be in the form of an emails, text messages or tweets etc which one might haven't asked for but still gets in contact with many a times. If we talk about emails common spam emails generally have an email address of some random word which are not legit for example @nits.cem , @heckererth.com etc.

A Ham on the other side are messages which one gets which are generally desired and one might have opted for it in one or the other way. For example, while logging into any ecommerce website there might be some suggestions of receiving latest updates of clothing or related stuff via mails to keep you updated because of which we receive emails on daily basis which we don't even realise that we opted for.

SPAM DETECTION SYSTEM

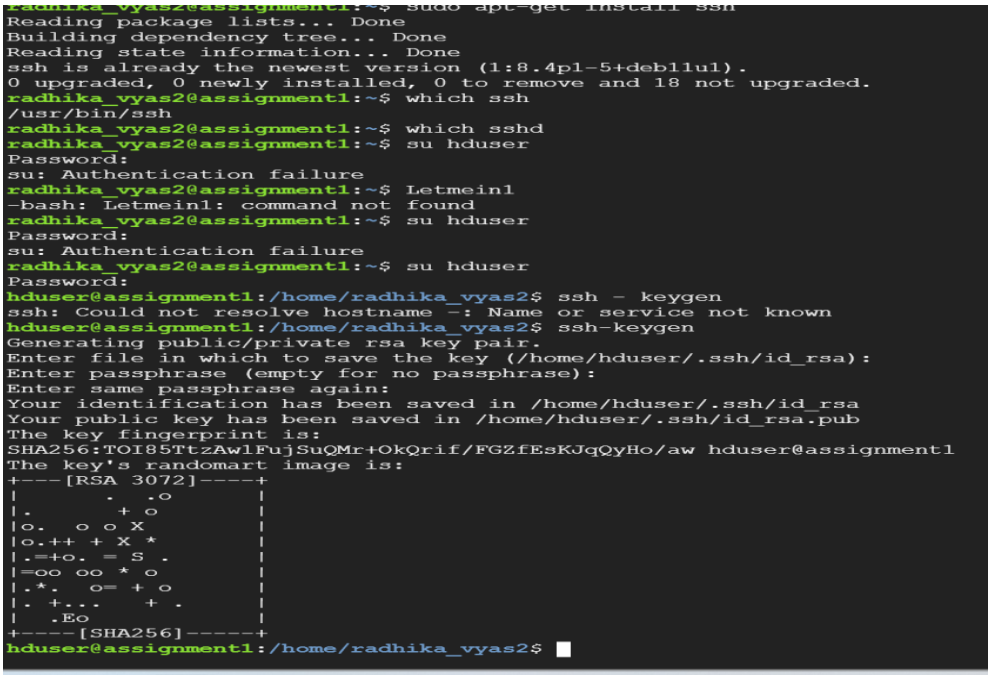
Task 1:

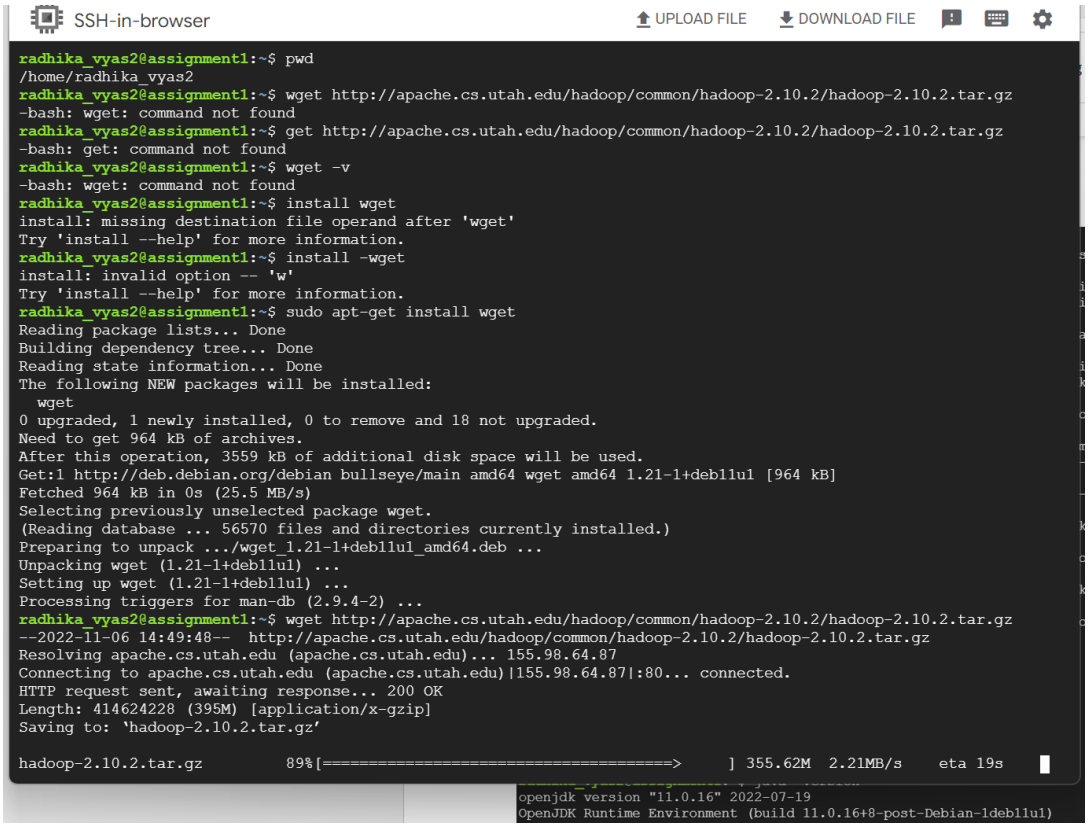
To perform the task Google cloud Platform was used for creation of a vm instance and single node cluster



Java Installation

Creation of HDuser for Hadoop




SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
radhika_vyas2@assignment1:~$ pwd
/home/radhika_vyas2
radhika_vyas2@assignment1:~$ wget http://apache.cs.utah.edu/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz
-bash: wget: command not found
radhika_vyas2@assignment1:~$ get http://apache.cs.utah.edu/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz
-bash: get: command not found
radhika_vyas2@assignment1:~$ wget -v
-bash: wget: command not found
radhika_vyas2@assignment1:~$ install wget
install: missing destination file operand after 'wget'
Try 'install --help' for more information.
radhika_vyas2@assignment1:~$ install -wget
install: invalid option -- 'w'
Try 'install --help' for more information.
radhika_vyas2@assignment1:~$ sudo apt-get install wget
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
wget
0 upgraded, 1 newly installed, 0 to remove and 18 not upgraded.
Need to get 964 kB of archives.
After this operation, 3559 kB of additional disk space will be used.
Get:1 http://deb.debian.org/debian bullseye/main amd64 wget amd64 1.21-1+deb11u1 [964 kB]
Fetched 964 kB in 0s (25.5 MB/s)
Selecting previously unselected package wget.
(Reading database ... 56570 files and directories currently installed.)
Preparing to unpack .../wget_1.21-1+deb11u1_amd64.deb ...
Unpacking wget (1.21-1+deb11u1) ...
Setting up wget (1.21-1+deb11u1) ...
Processing triggers for man-db (2.9.4-2) ...
radhika_vyas2@assignment1:~$ wget http://apache.cs.utah.edu/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz
--2022-11-06 14:49:48-- http://apache.cs.utah.edu/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz
Resolving apache.cs.utah.edu (apache.cs.utah.edu)... 155.98.64.87
Connecting to apache.cs.utah.edu (apache.cs.utah.edu)|155.98.64.87|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 414624228 (395M) [application/x-gzip]
Saving to: 'hadoop-2.10.2.tar.gz'
hadoop-2.10.2.tar.gz 89%[=====>] 355.62M 2.21MB/s eta 19s
openjdk version "11.0.16" 2022-07-19
OpenJDK Runtime Environment (build 11.0.16+8-post-Debian-1deb11u1)

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
    test -r ~/.dircolors && eval "$(dircolors -b ~/.dircolors)" || eval "$(dircol
    alias ls='ls --color=auto'
    #alias dir='dir --color=auto'
    #alias vdir='vdir --color=auto'

    #alias grep='grep --color=auto'
    #alias fgrep='fgrep --color=auto'
    #alias egrep='egrep --color=auto'
fi

# colored GCC warnings and errors
#export GCC_COLORS='error=01;31:warning=01;35:note=01;36:caret=01;32:locus=01:quo

# some more ls aliases
#alias ll='ls -l'
#alias la='ls -A'
#alias l='ls -CF'

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_INSTALL=/home/radhika_vyas2/hadoop
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
export HADOOP_HOME_WARN_SUPPRESS=1
export HADOOP_ROOT_LOGGER="WARN,DRFA"
".bashrc" 125L, 4057B                                                  125,37
```

Hive installation

```
Hive 3.1.2
Git git://HW13934/Users/gates/tmp/hive-branch-3.1/hive -r 8190d2be7b7165effa62bd21b7d60ef81fb0e4af
Compiled by gates on Thu Aug 22 15:01:18 PDT 2019
From source with checksum 0492c08f784b188c349f6afb1d8d9847
```

PIG INSTALLATION

```
Apache Pig version 0.17.0 (r1797386)
compiled Jun 02 2017, 15:41:58
```

The dataset used is a dataset of spam and ham in SMS which was extracted from Kaggle https://github.com/radhikavyas12/Assignment-1-Cloud/blob/main/spamdataset.csv

Firstly the file was uploaded in the cloud system cluster.

New directory was created in HDFS

SSH-in-browser

```
Linux hadoopcluster-w-0 5.10.0-0.deb10.16-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Mon Oct 31 23:21:28 2022 from 35.235.241.18
radhika_vyas2@hadoopcluster-w-0:~$ hdfs dfs -put -f /home/radhika_vyas2/dataset.xls /home/dataproc/sd/dataset/
-put: java.net.UnknownHostException: hadoopcluster-m
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-copyFromLocal [-f] [-p] [-l] [-d] [-t <thread count>] [-q <thread pool queue size>] <localsrc> ... <dst>]
        [-copyToLocal [-f] [-p] [-crc] [-ignoreCrc] [-t <thread count>] [-q <thread pool queue size>] <src> ... <localdst>]
        [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] [-e] <path> ...]
        [-cp [-f] [-p | -p[topax]] [-d] [-t <thread count>] [-q <thread pool queue size>] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] [-v] [-x] <path> ...]
        [-expunge [-immediate]]
        [-find <path> ... <expression> ...]
        [-get [-f] [-p] [-crc] [-ignoreCrc] [-t <thread count>] [-q <thread pool queue size>] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getfattr [-R] {-n name | -d} [-e en] <path>]
        [-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
        [-head <file>]
        [-help [cmd ...]]
        [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [-e] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] [-l] [-d] [-t <thread count>] [-q <thread pool queue size>] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <path>]]
        [-setfattr {-n name [-v value] | -x name} <path>]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] [-s <sleep interval>] <file>]
        [-test -[defswrz] <path>]
```

for creating external table in HIVE

create external table if not exists sms_scam(Category string,SMS string)

> comment 'SMS SPAM DETECTION'

> row format delimited

> fields terminated by','

> stored as textfile

> location '/rv/dataset/';

```
put: `/home/radhika_vyas2/dataset.xlsx': No such file or directory
radhika_vyas2@hadoopcluster-m:~$ pwd
/home/radhika_vyas2
radhika_vyas2@hadoopcluster-m:~$ ls
Dataset.xlsx  dataset  sataset
radhika_vyas2@hadoopcluster-m:~$ hdfs dfs -put -f /home/radhika_vyas2/Da
et
radhika_vyas2@hadoopcluster-m:~$ hive
Hive Session ID = 9e0b2f2e-9f53-4d31-93d2-b547960077d2

Logging initialized using configuration in file:/etc/hive/conf.dist/hive
Async: true
Hive Session ID = 111f9c91-1e8b-4bec-9c1e-d3c2e963995d
hive> create external table if not exists sms_scam(Category string,SMS s
    > comment 'SMS SPAM DETECTION'
    > row format delimited
    > fields terminated by','
    > stored as textfile
    > location '/rv/dataset/';
OK
Time taken: 1.602 seconds
hive>
```

SSH-in-browser

```
    > stored as textfile
    > location '/rv/dataset/';
OK
Time taken: 0.249 seconds
hive> select * from sms_scam
    > select * from sms_scam;
FAILED: ParseException line 2:0 missing EOF at 'select' near 'sms_scam'
hive> select*from sms_scam;
Query ID = radhika_vyas2_20221107223425_488022f0-9a2d-4882-a3cd-79f2e86f5203
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1667854260581_0004)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
        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: 6.63 s
--------------------------------------------------------------------------------
OK
Category        SMS
ham     "Go until jurong point
ham     Ok lar... Joking wif u oni...
spam    Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham     U dun say so early hor... U c already then say...
ham     "Nah I don't think he goes to usf
spam    "FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send
ham     Even my brother is not like to speak with me. They treat me like aids patent.
ham     As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Calle
spam    WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours onl
spam    Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
ham     "I'm gonna be home soon and i don't want to talk about this stuff anymore tonight
spam    "SIX chances to win CASH! From 100 to 20
spam    "URGENT! You have won a 1 week FREE membership in our £100
ham     I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have be
ing at all times.
ham     I HAVE A DATE ON SUNDAY WITH WILL!!
spam    "XXXMobileMovieClub: To use your credit
ham     Oh k...i'm watching here:)
ham     Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham     Fine if that's the way u feel. That's the way its gota b
spam    "England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES
ham     Is that seriously how you spell his name?
ham     I'm going to try for 2 months ha ha only joking
ham     So yeah first lar... Then when is da stock comin...
```

```
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:821)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:759)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:683)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
FAILED: ParseException line 1:39 cannot recognize input near 'is' 'ham' '<EOF>' in expression specification
hive> select* from sms_scam where Category='ham';
Query ID = radhika_vyas2_20221107224103_04ada162-9622-4e24-9621-1b690339666e
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_1667854260581_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: 7.29 s
--------------------------------------------------------------------------------
OK
ham     "Go until jurong point
ham     Ok lar... Joking wif u oni...
ham     U dun say so early hor... U c already then say...
ham     "Nah I don't think he goes to usf
ham     Even my brother is not like to speak with me. They treat me like aids patent.
ham     As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to c
ham     "I'm gonna be home soon and i don't want to talk about this stuff anymore tonight
ham     I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil m
ing at all times.
ham     I HAVE A DATE ON SUNDAY WITH WILL!!
ham     Oh k...i'm watching here:)
ham     Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham     Fine if that's the way u feel. That's the way its gota b
ham     Is that seriously how you spell his name?
ham     I'm going to try for 2 months ha ha only joking
```