Cloud Technologies- CA675

Student details

Name: Archit Garg
Student ID: 22262959
Email: archit.garg2@mail.dcu.ie

Link for the Git repository:

https://github.com/ItsArchit/Cloud-Technologies-CA675-.git

Link for the project on the cloud system:

https://console.cloud.google.com/home/dashboard?project=kinetic-axle-367820

Dataset Link:

https://www.kaggle.com/datasets/fabioscopeta/email-datasets-for-inference-attacks?select=all_emails.csv

Reasons to Select this dataset-

As per the requirements,

1. The aim was to find a dataset with data containing both SPAM as well as HAM contents to be filtered

2. To find a fairly large and complex dataset.

About dataset-

The selected dataset is a merged CSV of raw email data from ENRON and SPAMASSASSIN.

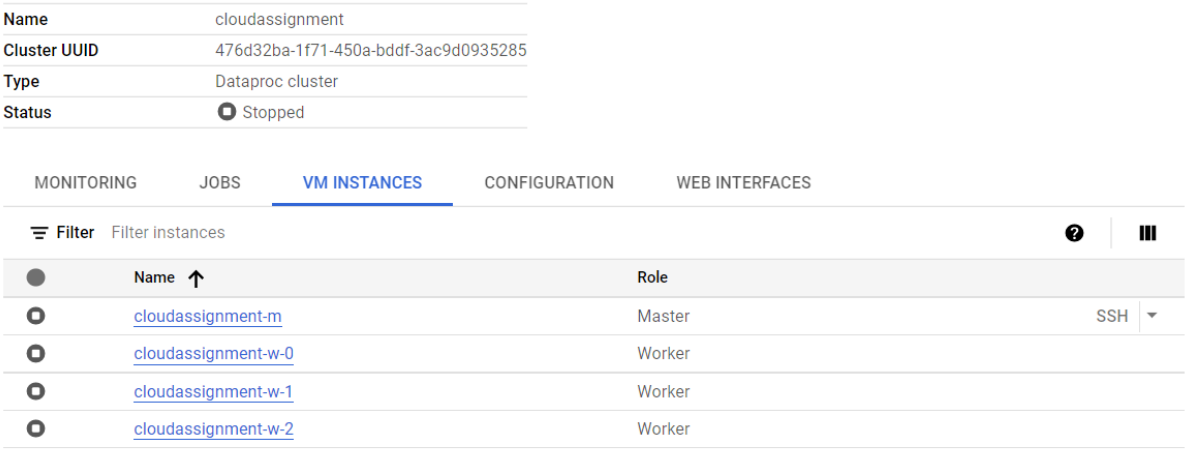
The dataset contains the following columns:

1. Date
2. To
3. From
4. Label

The dataset was fairly raw and hence ideal to be used to perform Data cleaning and data pre-processing.

Description

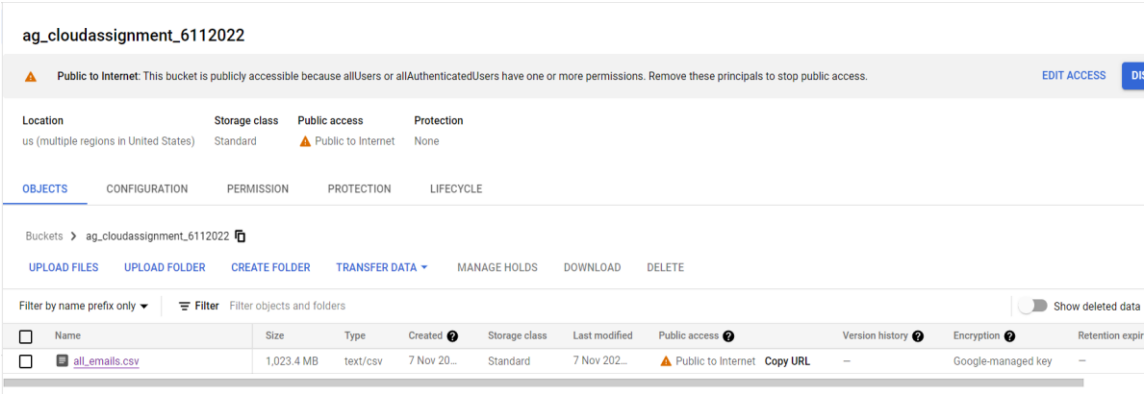
1. Started by creating a DataProc Cluster named “cloudassignment” with 1 Master node a 3 Worker node. Hadoop cluster namely "cloudassignment" on project id: profound-coda-362616 is used.



2. Created a Bucket name ag_cloudassignment_6112022.

3. Updated the permissions, to allow access into the cluster.

4. Uploaded the dataset named all_emails.csv into the bucket.

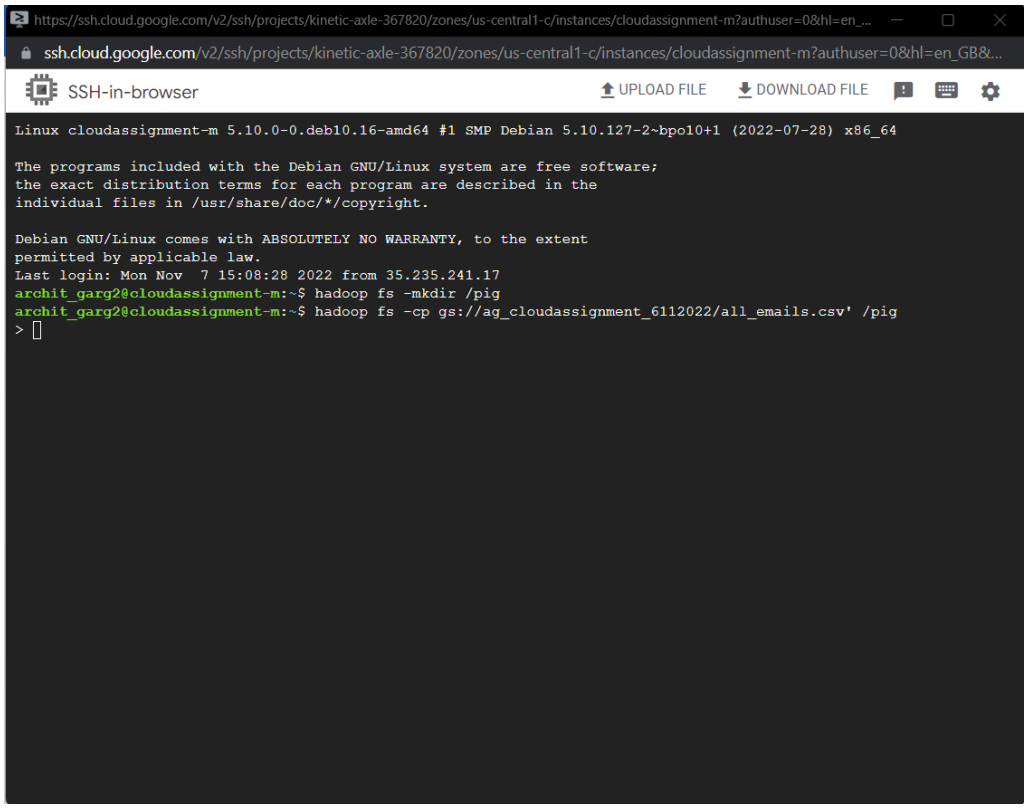


5. Launched **SSH** console through the master node of the Cluster created.

 # PRE-PROCESSING DATASET on GOOGLE CLOUD PLATFORM (GCP)#

----------HADOOP----------

#The CSV file was copied to the cluster from the bucket.
hadoop fs -mkdir /pig
hadoop fs -cp 'gs://ah-cloudassignment1-6nov2022/all_emails.csv' /pig



#FILE UPLOADED#

'hdfs://cloudassignment-m/piggy /all_emails.csv'

```
archit_garg2@cloudassignment-m:~$ hadoop fs -ls  /piggy
Found 1 items
-rw-r--r--   2 archit_garg2 hadoop 1073124505 2022-11-07 15:27 /piggy/all_emails.csv
```

#PIGGYBANK INSATLLED#

wget https://github.com/prasad1825/CA675-Assignment2/raw/main/Data%20Cleaning/piggybank.jar

```
archit_garg2@cloudassignment-m:~$ wget https://github.com/prasad1825/CA675-Assignment2/raw/main/Data%20Cleaning
/piggybank.jar
--2022-11-07 16:53:28--  https://github.com/prasad1825/CA675-Assignment2/raw/main/Data%20Cleaning/piggybank.jar
Resolving github.com (github.com)... 140.82.113.3
Connecting to github.com (github.com)|140.82.113.3|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/prasad1825/CA675-Assignment2/main/Data%20Cleaning/piggybank.jar [fo
llowing]
--2022-11-07 16:53:28--  https://raw.githubusercontent.com/prasad1825/CA675-Assignment2/main/Data%20Cleaning/pi
ggybank.jar
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.11
0.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 342415 (334K) [application/octet-stream]
Saving to: 'piggybank.jar'

piggybank.jar               100%[===========================================>] 334.39K  --.-KB/s    in 0.03s

2022-11-07 16:53:29 (9.61 MB/s) - 'piggybank.jar' saved [342415/342415]
```

#LAUNCHING PIG#

Pig

Functions such as CSVLoader and PigStorage would prove to be insufficient as the data was raw and contained certain fields with special characters and line breaks.
Therefore, CSVExcelStorage is used as it supports in loading multi line data.
Location: https://cwiki.apache.org/confluence/display/PIG/PiggyBank

#REGISTERING CSVExcelStorage#

register /home/Archit_garg2/piggybank.jar

```
grunt> register /home/architgarg/piggybank.jar
2022-11-07 16:54:06,462 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.sy
stem-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
2022-11-07 16:54:06,491 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 101: file '/home/architgarg/piggy
bank.jar' does not exist.
Details at logfile: /home/archit_garg2/pig_1667840015910.log
grunt> register /home/archit_garg2/piggybank.jar
2022-11-07 16:54:27,003 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.sy
stem-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt>
```

#Load data from the five CSV files into Pig

mailDataFile = Load '/piggy/all_emails.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE') AS (date:chararray, to:chararray, from:chararray, body:chararray, label:chararray);

```
grunt> mailDataFile = Load '/piggy/all_emails.csv' USING org.apache.pig.piggybank.storage.CSVE
xcelStorage(',','YES_MULTILINE') AS (date:chararray, to:chararray, from:chararray, body:charar
ray, label:chararray);
2022-11-07 22:18:20,349 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.r
esourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metric
s-publisher.enabled
```

#EXTRACTED REQUIRED COLUMNS AND USED REPLACE FOR CLEANING THE EMAIL BODY#

generateMailDataFile =FOREACH mailDataFile GENERATE date, to, from,REPLACE(REPLACE(REPLACE(REPLACE(REPLACE((REPLACE(body,'[\r\n]+','')),'<[^>]*>' , ' '),'[^a-zA-Z\\s\']+',' '),'(?=\\S*[\'])([a-zA-Z\'-]+)',''),'(?<![\\w\\-])\\w(?![\\w\\-])',''),'[ ]{2,}',' ') as body ;

 # ELIMINATED ROWS WITH AT LEAST ONE NULL FIELD#
generateMailDataFile_notnull = FILTER generateMailDataFile by NOT ((date IS NULL) OR (to IS NULL) OR (from IS NULL) OR (body IS NULL));

#ELIMINATED ROWS WITH AT LEAST ONE BLANK FIELD#

generateMailDataFile_notnull_notblank = FILTER generateMailDataFile_notnull by NOT ((to ==\'\') OR (from ==\'\') OR (body ==\'\'));

 # ELIMINATE ROWS WITH AT LEAST ONE 'N/A' FIELD #

generateMailDataFile_notnull_notblank_na = FILTER generateMailDataFile_notnull_notblank by NOT ((to ==\'N/A\') OR (from ==\'N/A\') OR (body ==\'N/A\'));

```
grunt> generateMailDataFile =FOREACH mailDataFile GENERATE date, to, from,REPLACE(REPLACE(REPL
ACE(REPLACE(REPLACE((REPLACE(body,'[\r\n]+','')),'<[^>]*>' , ' '),'[^a-zA-Z\\s\']+',' '),'(?=\
\S*[\'])([a-zA-Z\'-]+)',''),'(?<![\\w\\-])\\w(?![\\w\\-])',''),'[ ]{2,}',' ') as body ;
grunt> generateMailDataFile_notnull = FILTER generateMailDataFile  by NOT ((date IS NULL) OR (
to IS NULL) OR (from IS NULL) OR (body IS NULL) );
grunt> generateMailDataFile_notnull_notblank = FILTER generateMailDataFile_notnull  by NOT ((t
o ==\'\') OR (from ==\'\') OR (body ==\'\'));
grunt> generateMailDataFile_notnull_notblank_na = FILTER generateMailDataFile_notnull_notblank
  by NOT ((to ==\'N/A\') OR (from ==\'N/A\') OR (body ==\'N/A\'));
```

#STORING FILTERED DATA INTO -> HDFS/FinalHiveData#

STORE generateMailDataFile_notnull_notblank_na INTO '/FinalHiveData' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE');

```
grunt> STORE generateMailDataFile_notnull_notblank_na INTO '/FinalHiveData' USING org.apache.p
ig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE');
2022-11-07 22:20:36,986 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.r
esourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metri
s-publisher.enabled
2022-11-07 22:20:37,018 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred
.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.sepa
rator
2022-11-07 22:20:37,040 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
used in the script: FILTER
2022-11-07 22:20:37,060 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.r
esourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metri
s-publisher.enabled
2022-11-07 22:20:37,073 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schemat
uple] was not set... will not generate code.
2022-11-07 22:20:37,104 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptim
izer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, FilterConstantCalculator, ForEachConstan
tCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, Me
rgeForEach, NestedLimitOptimizer, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDo
wnForEachFlatten, PushUpFilter, SplitConstantCalculator, SplitFilter, StreamTypeCastInserter]}
2022-11-07 22:20:37,130 [main] INFO  org.apache.pig.newplan.logical.rules.ColumnPruneVisitor -
 Columns pruned for mailDataFile: $4
```

#STORING WAS A SUCCESS#

```
HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
3.2.3   0.18.0-SNAPSHOT archit_garg2    2022-11-07 22:20:37     2022-11-07 22:23:55     FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxRed
uceTime   MinReduceTime  AvgReduceTime   MedianReducetime        Alias   Feature Outputs
job_1667858620139_0001  8       0       177     22      52      39      0       0       0
  0       generateMailDataFile,generateMailDataFile_notnull,mailDataFile MAP_ONLY        /Fina
lHiveData,

Input(s):
Successfully read 21428620 records (1073156153 bytes) from: "/piggy/all_emails.csv"

Output(s):
Successfully stored 808429 records (175938456 bytes) in: "/FinalHiveData"

Counters:
Total records written : 808429
Total bytes written : 175938456
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1667858620139_0001
```

After the storage, Pig divided the result into _SUCCESS file and part-m- files in /FinalHiveData in HDFS.

#THE LOG FILE NAMELY SUCCESS DELETED#

hadoop fs -rm /FinalHiveData/_SUCCESS

```
archit_garg2@cloudassignment-m:~$ hadoop fs -rm /FinalHiveData/_SUCCESS
Deleted /FinalHiveData/_SUCCESS
```

 # part-m- files in /FinalHiveData were merged into only file

hadoop fs -getmerge /FinalHiveData /home/archit_garg2/hive_allmails_input.csv

---hadoop fs -put hive_allmails_input.csv 'gs://ag_cloudassignment_6112022_updated'

```
archit_garg2@cloudassignment-m:~$ hadoop fs -getmerge /FinalHiveData /home/archit_garg2/hive_a
llmails_input.csv
archit_garg2@cloudassignment-m:~$ hadoop fs -put hive_allmails_input.csv 'gs://ag_cloudassignm
ent_6112022_updated'
```

#CREATED A BUCKET TO STORE THE UPDATED DATASET#

Created a csv file to 'gs:// ag_cloudassignment_6112022_updated'

Location of the updated cleaned and processed dataset in the bucket:

https://storage.googleapis.com/ag_cloudassignment_6112022_updated/hive_allmails_input.csv

THANK YOU