Assignment 1 - Data Analysis

CA675 - Cloud Technologies

Student details

a. Name : Ajay Anni Hegde

b. Student ID : 22266405

c. Email : ajay.hegde2@mail.dcu.ie

Link for the Git repository: https://github.com/ajayhegde007/CA675-CloudTechnologies.git

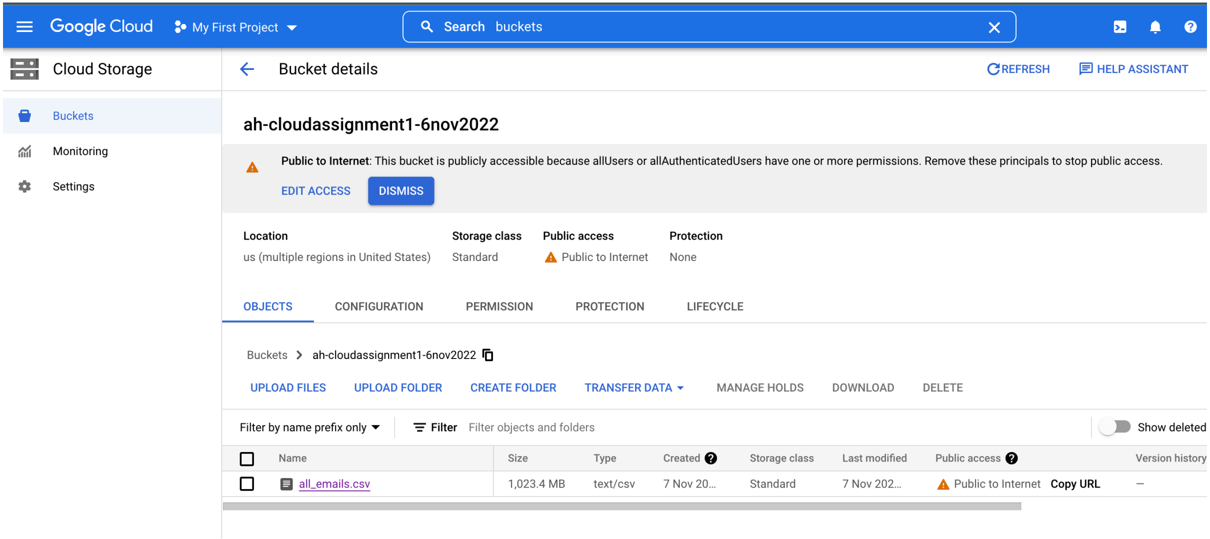
Link for the project on the cloud system:
https://console.cloud.google.com/home/dashboard?project=profound-coda-362616

Dataset source:

The dataset used is *all_emails.csv* which contains email data from 'ENRON' mail dataset and SPAMASSASSIN mail dataset. The dataset is in .csv format.
Below is the link for more details:
https://www.kaggle.com/datasets/fabioscopeta/email-datasets-for-inference-attacks?select=all_emails.csv

Description:

1. Initially a cluster named 'cloudassignment' is created in GCP under project "My First Project". Project link: https://console.cloud.google.com/home/dashboard?project=profound-coda-362616
2. A Hadoop cluster namely "*cloudassignment*" on project id: profound-coda-362616 has been used as a working environment to implement data cleaning and querying with Pig and Hive.
3. The dataset "all_emails.csv" is downloaded and uploaded to the bucket name *ah-cloudassignment1-6nov2022* in project namely "*My First Project*", Project-ID:profound-coda-362616



ah-cloudassignment1-6nov2022

⚠ **Public to Internet**: This bucket is publicly accessible because allUsers or allAuthenticatedUsers have one or more permissions. Remove these principals to stop public access.

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	⚠ Public to Internet	None

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Version history
□	all_emails.csv	1,023.4 MB	text/csv	7 Nov 20...	Standard	7 Nov 202...	⚠ Public to Internet Copy URL	–

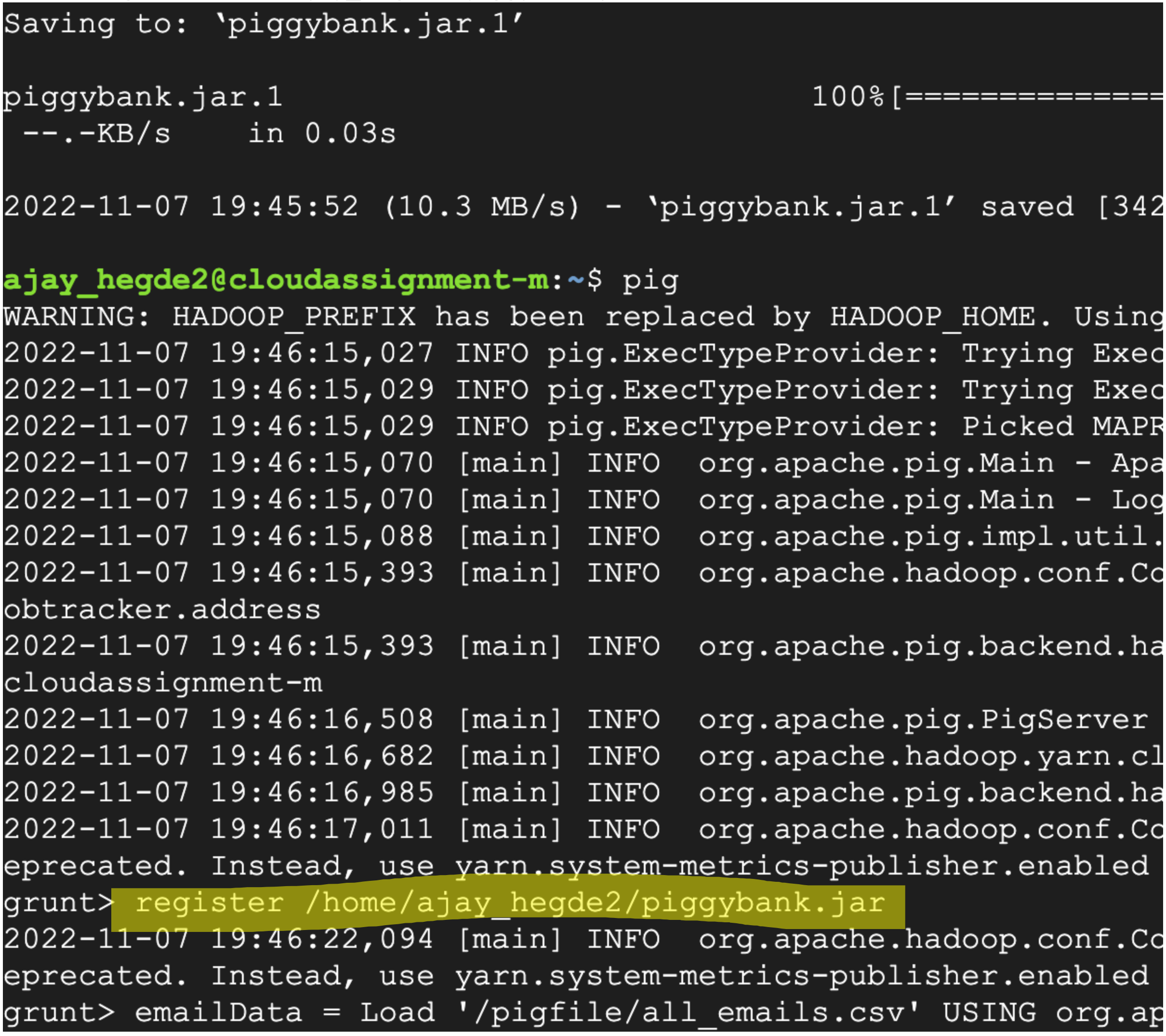
4. ---HADOOP--

The CSV file was copied to the cluster from the bucket:

- hadoop fs -mkdir /pigfile
- hadoop fs -cp 'gs://ah-cloudassignment1-6nov2022/all_emails.csv' /pigfile

5. -------PIG-----------
   The entry data contains a field with commas,line-break characters. Therefore, the functions such as CSVLoader and PigStorage did not work properly to handle them. Instead, CSVExcelStorage has been selected due to its support for loading multi line data. This function is available in the piggybank library (link:https://cwiki.apache.org/confluence/display/PIG/PiggyBank) and registered into Hadoop as follows:
   - wget https://github.com/prasad1825/CA675-Assignment2/raw/main/Data%20Cleaning/piggybank.jar

   Then using pig:
   - register /home/ajay_hegde2/piggybank.jar



```
Saving to: 'piggybank.jar.1'

piggybank.jar.1                          100%[=============
 --.-KB/s     in 0.03s

2022-11-07 19:45:52 (10.3 MB/s) - 'piggybank.jar.1' saved [342

ajay_hegde2@cloudassignment-m:~$ pig
WARNING: HADOOP_PREFIX has been replaced by HADOOP_HOME. Using
2022-11-07 19:46:15,027 INFO pig.ExecTypeProvider: Trying Exec
2022-11-07 19:46:15,029 INFO pig.ExecTypeProvider: Trying Exec
2022-11-07 19:46:15,029 INFO pig.ExecTypeProvider: Picked MAPR
2022-11-07 19:46:15,070 [main] INFO  org.apache.pig.Main - Apa
2022-11-07 19:46:15,070 [main] INFO  org.apache.pig.Main - Log
2022-11-07 19:46:15,088 [main] INFO  org.apache.pig.impl.util.
2022-11-07 19:46:15,393 [main] INFO  org.apache.hadoop.conf.Co
obtracker.address
2022-11-07 19:46:15,393 [main] INFO  org.apache.pig.backend.ha
cloudassignment-m
2022-11-07 19:46:16,508 [main] INFO  org.apache.pig.PigServer
2022-11-07 19:46:16,682 [main] INFO  org.apache.hadoop.yarn.cl
2022-11-07 19:46:16,985 [main] INFO  org.apache.pig.backend.ha
2022-11-07 19:46:17,011 [main] INFO  org.apache.hadoop.conf.Co
eprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> register /home/ajay_hegde2/piggybank.jar
2022-11-07 19:46:22,094 [main] INFO  org.apache.hadoop.conf.Co
eprecated. Instead, use yarn.system-metrics-publisher.enabled
grunt> emailData = Load '/pigfile/all_emails.csv' USING org.ap
```

6. Load data from the five CSV files into Pig

emailData = Load 'hdfs://cloudassignment-m/pigfile/all_emails.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE') AS (date:chararray, to:chararray, from:chararray, body:chararray, label:chararray);

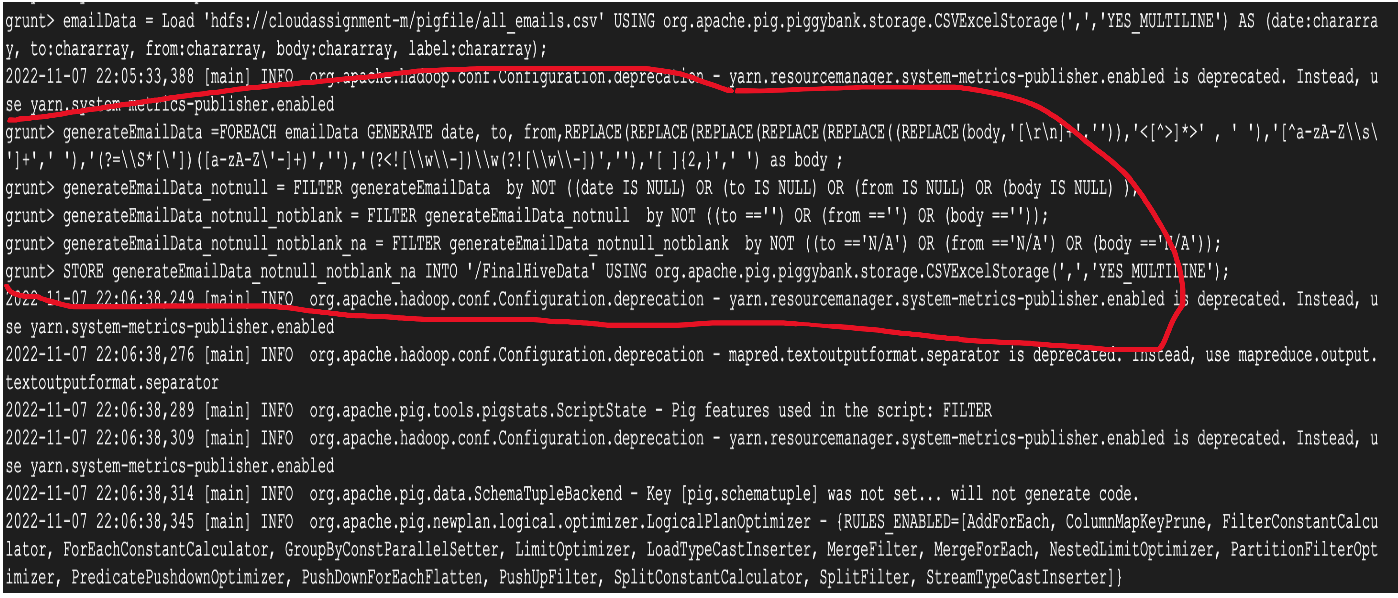
7. Extracting only the required columns and cleaning the email body
    - generateEmailData =FOREACH emailData GENERATE date, to, from,REPLACE(REPLACE(REPLACE(REPLACE(REPLACE((REPLACE(body,'[\r\n]+', '')),'<[^>]*>' , ' '),'[^a-zA-Z\\s\']+',' '),'(?=\\S*[\'])([a-zA-Z\'-]+)',''),'(?<![\\w\\-])\\w(?![\\w\\-])',''),'[ ]{2,}',' ') as body ;

8. Filtered data rows to eliminate rows with at least one null field
    - generateEmailData_notnull = FILTER generateEmailData by NOT ((date IS NULL) OR (to IS NULL) OR (from IS NULL) OR (body IS NULL));

9. Filtered data rows to eliminate rows with at least one blank field
    - generateEmailData_notnull_notblank = FILTER generateEmailData_notnull by NOT ((to =='') OR (from =='') OR (body ==''));

10. Filtered data rows to eliminate rows with at least one 'N/A' field
    - generateEmailData_notnull_notblank_na = FILTER generateEmailData_notnull_notblank by NOT ((to =='N/A') OR (from =='N/A') OR (body =='N/A'));

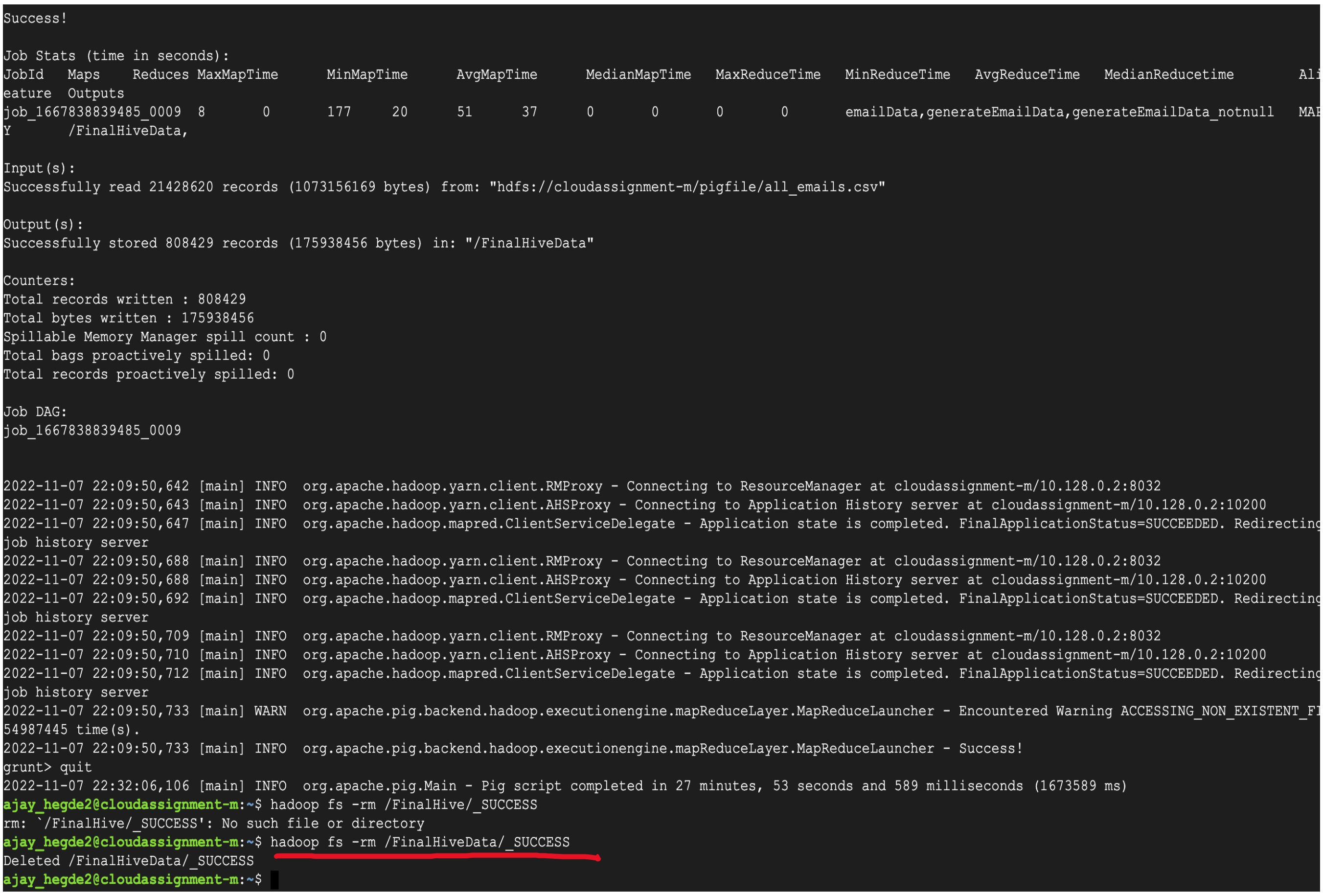
11. Stored The filtered data into HDFS /FinalHive
    - STORE generateEmailData_notnull_notblank_na INTO '/FinalHiveData' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE');



```
grunt> emailData = Load 'hdfs://cloudassignment-m/pigfile/all_emails.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE') AS (date:chararra
y, to:chararray, from:chararray, body:chararray, label:chararray);
2022-11-07 22:05:33,388 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, u
se yarn.system-metrics-publisher.enabled
grunt> generateEmailData =FOREACH emailData 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 ;
grunt> generateEmailData_notnull = FILTER generateEmailData  by NOT ((date IS NULL) OR (to IS NULL) OR (from IS NULL) OR (body IS NULL) );
grunt> generateEmailData_notnull_notblank = FILTER generateEmailData_notnull  by NOT ((to =='') OR (from =='') OR (body ==''));
grunt> generateEmailData_notnull_notblank_na = FILTER generateEmailData_notnull_notblank  by NOT ((to =='N/A') OR (from =='N/A') OR (body =='N/A'));
grunt> STORE generateEmailData_notnull_notblank_na INTO '/FinalHiveData' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE');
2022-11-07 22:06:38,249 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, u
se yarn.system-metrics-publisher.enabled
2022-11-07 22:06:38,276 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.
textoutputformat.separator
2022-11-07 22:06:38,289 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2022-11-07 22:06:38,309 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, u
se yarn.system-metrics-publisher.enabled
2022-11-07 22:06:38,314 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-11-07 22:06:38,345 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, FilterConstantCalcu
lator, ForEachConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NestedLimitOptimizer, PartitionFilterOpt
imizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitConstantCalculator, SplitFilter, StreamTypeCastInserter]}
```

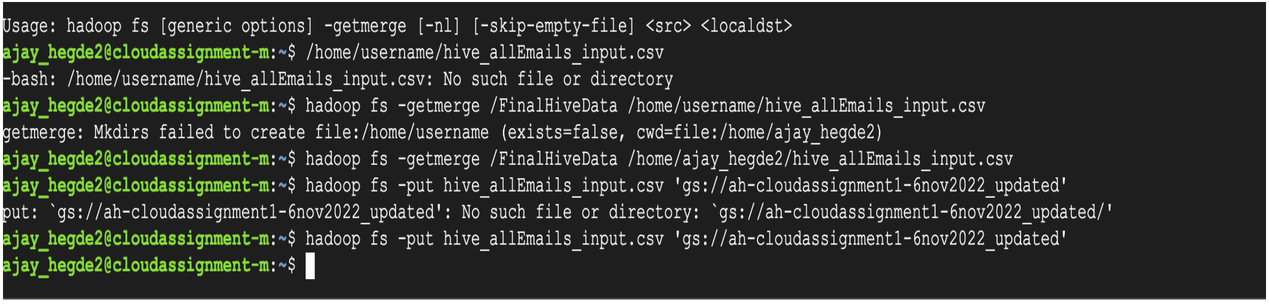
12. After the storage, Pig has divided the result in _SUCCESS file and part-m- files in /FinalHive in HDFS. The log file namely SUCCESS will block the load function of Hive so this file needs to be deleted with the following command:

    - hadoop fs -rm /FinalHiveData/_SUCCESS



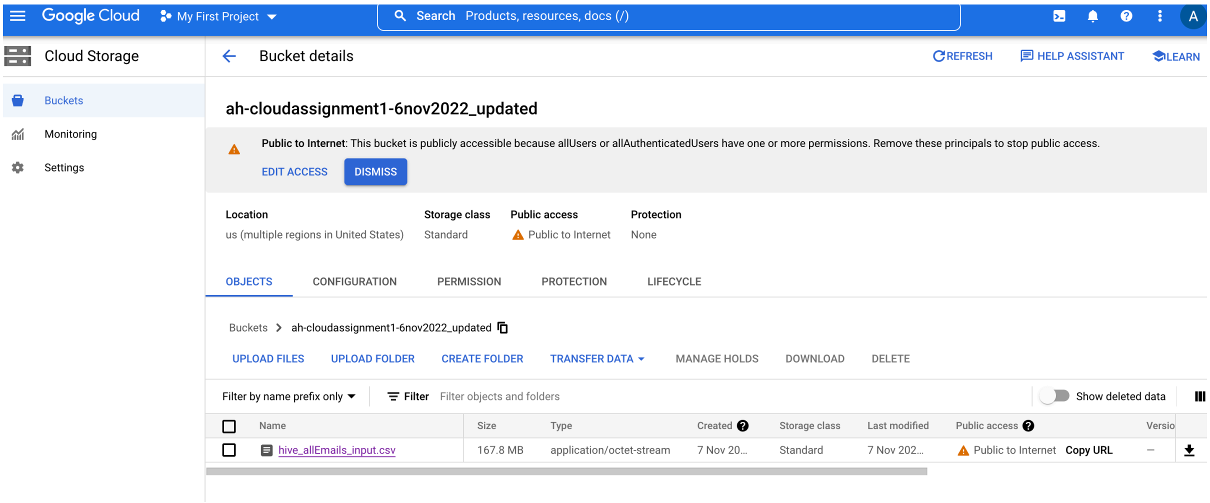
```
Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTime      AvgMapTime      MedianMapTime   MaxReduceTime   MinReduceTime   AvgReduceTime   MedianReducetime        Ali
eature  Outputs
job_1667838839485_0009  8       0       177     20      51      37      0       0       0       0       emailData,generateEmailData,generateEmailData_notnull   MAI
Y       /FinalHiveData,

Input(s):
Successfully read 21428620 records (1073156169 bytes) from: "hdfs://cloudassignment-m/pigfile/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_1667838839485_0009


2022-11-07 22:09:50,642 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cloudassignment-m/10.128.0.2:8032
2022-11-07 22:09:50,643 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cloudassignment-m/10.128.0.2:10200
2022-11-07 22:09:50,647 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
job history server
2022-11-07 22:09:50,688 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cloudassignment-m/10.128.0.2:8032
2022-11-07 22:09:50,688 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cloudassignment-m/10.128.0.2:10200
2022-11-07 22:09:50,692 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
job history server
2022-11-07 22:09:50,709 [main] INFO  org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at cloudassignment-m/10.128.0.2:8032
2022-11-07 22:09:50,710 [main] INFO  org.apache.hadoop.yarn.client.AHSProxy - Connecting to Application History server at cloudassignment-m/10.128.0.2:10200
2022-11-07 22:09:50,712 [main] INFO  org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting
job history server
2022-11-07 22:09:50,733 [main] WARN  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Encountered Warning ACCESSING_NON_EXISTENT_FI
54987445 time(s).
2022-11-07 22:09:50,733 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> quit
2022-11-07 22:32:06,106 [main] INFO  org.apache.pig.Main - Pig script completed in 27 minutes, 53 seconds and 589 milliseconds (1673589 ms)
ajay_hegde2@cloudassignment-m:~$ hadoop fs -rm /FinalHive/_SUCCESS
rm: `/FinalHive/_SUCCESS': No such file or directory
ajay_hegde2@cloudassignment-m:~$ hadoop fs -rm /FinalHiveData/_SUCCESS
Deleted /FinalHiveData/_SUCCESS
ajay_hegde2@cloudassignment-m:~$
```

13. part-m- files in /FinalHiveData were merged into only file :
    - hadoop fs -getmerge /FinalHiveData /home/ajay_hegde2/hive_allEmails_input.csv
    - hadoop fs -put hive_allEmails_input.csv 'gs://ah-cloudassignment1-6nov2022_updated'.



```
Usage: hadoop fs [generic options] -getmerge [-nl] [-skip-empty-file] <src> <localdst>
ajay_hegde2@cloudassignment-m:~$ /home/username/hive_allEmails_input.csv
-bash: /home/username/hive_allEmails_input.csv: No such file or directory
ajay_hegde2@cloudassignment-m:~$ hadoop fs -getmerge /FinalHiveData /home/username/hive_allEmails_input.csv
getmerge: Mkdirs failed to create file:/home/username (exists=false, cwd=file:/home/ajay_hegde2)
ajay_hegde2@cloudassignment-m:~$ hadoop fs -getmerge /FinalHiveData /home/ajay_hegde2/hive_allEmails_input.csv
ajay_hegde2@cloudassignment-m:~$ hadoop fs -put hive_allEmails_input.csv 'gs://ah-cloudassignment1-6nov2022_updated'
put: `gs://ah-cloudassignment1-6nov2022_updated': No such file or directory: `gs://ah-cloudassignment1-6nov2022_updated/'
ajay_hegde2@cloudassignment-m:~$ hadoop fs -put hive_allEmails_input.csv 'gs://ah-cloudassignment1-6nov2022_updated'
ajay_hegde2@cloudassignment-m:~$
```

14. Link to bucket *'ah-cloudassignment1-6nov2022_updated'* where cleaned data has been saved: https://storage.googleapis.com/ah-cloudassignment1-6nov2022_updated/hive_allEmails_input.csv



ah-cloudassignment1-6nov2022_updated

Public to Internet: This bucket is publicly accessible because allUsers or allAuthenticatedUsers have one or more permissions. Remove these principals to stop public access.

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Public to Internet	None

Buckets > ah-cloudassignment1-6nov2022_updated

□	Name	Size	Type	Created	Storage class	Last modified	Public access	Versio
□	hive_allEmails_input.csv	167.8 MB	application/octet-stream	7 Nov 20…	Standard	7 Nov 202…	Public to Internet Copy URL	–

THE END