**Screenshots along with 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

Graphical user interface, text, application, email

Description automatically generated

1. -----------------------------------------------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

1. -------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

Text

Description automatically generated



1. 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);

1. 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 ;

1. 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) );

1. 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 ==''));

1. Filtered data rows to eliminate rows with at least one 'N/A' field

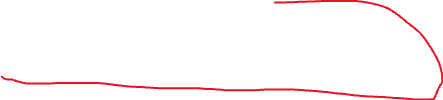
* generateEmailData\_notnull\_notblank\_na = FILTER generateEmailData\_notnull\_notblank by NOT ((to =='N/A') OR (from =='N/A') OR (body =='N/A'));

1. Stored The filtered data into HDFS /FinalHive

* STORE generateEmailData\_notnull\_notblank\_na INTO '/FinalHiveData' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES\_MULTILINE');

![Graphical user interface, text, website

Description automatically generated



1. On storage completion, Pig segregates the result in \_SUCCESS file and part-m- files in /FinalHive in HDFS. The SUCCESS named log file will obstruct the loading work of Hive so this file needs to be removed using the below code:

* hadoop fs -rm /FinalHiveData/\_SUCCESS

Text

Description automatically generated



1. part-m- files in /FinalHiveData were combined 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'.

Text

Description automatically generated

1. Link to bucket *‘ah-cloudassignment1-6nov2022\_updated’* where cleaned dataset has been stored: <https://storage.googleapis.com/ah-cloudassignment1-6nov2022_updated/hive_allEmails_input.csv>

Graphical user interface, application, email

Description automatically generated

**END**