Commands used to copy .csv file from bucket to cluster:

* hadoop fs -mkdir /pigfile
* hadoop fs -cp 'gs://ah-cloudassignment1-6nov2022/all\_emails.csv' /pigfile

Commands to import .jar file to support CSVExcelStorage :

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

Then using pig:

* register /home/ajay\_hegde2/piggybank.jar

**PIG commands used to clean dataset**:

*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);

*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 ;

*Filtering 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) )

*Filtering data rows to eliminate rows with at least one blank field*

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

*Filtering 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'));

*Storing the filtered data into HDFS /FinalHive*

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

**HADOOP Commands:**

*Deleting the \_SUCCESS named log file*

* hadoop fs -rm /FinalHiveData/\_SUCCESS

*Merging files in FinalHiveData file and storing as ‘*hive\_allEmails\_input.csv’

* hadoop fs -getmerge /FinalHiveData /home/ajay\_hegde2/hive\_allEmails\_input.csv

*Placing the merged file into bucket ‘ah-cloudassignment1-6nov2022\_updated’*

* hadoop fs -put hive\_allEmails\_input.csv 'gs://ah-cloudassignment1-6nov2022\_updated'.