#Task 3 – Data Cleaning

#The process of data cleaning is carried out in Pig. First of all the dataset which is to be cleaned should be copied from the bucket to the local terminal disk space/Hadoop system. This will be done by the following command

1. `hadoop fs -cp 'gs://dataproc-staging-europe-west1-773978562921-grrhd7uh/assignment1DataBucket/ElectronicsCSV1.csv' /DataCSV`

#One the data is copied, start the pig. Once the Pig is starts, include the library of piggybank.jar. I included this file through my GitHub with wget command in hadoop directory as follows:

2. `wget https://github.com/hamzaqadeer1/CA-675---Cloud-Technologies-Assignment-1/blob/main/piggybank.jar`

#then I was able to register this in my Pig cleaning session.

3. `register /home/hamza_qadeer2/pig/piggybank.jar`

Then starts the reading and loading of the data from Hadoop file system in Pig by including hdfs path and columns names along with their datatypes. Command for Loading data with the following command. The command used is as follows:

4.
```
newElectronicsCSV1_pigLoad = Load 'hdfs://cluster-assignment1-m/user/hamza_qadeer2/LabTasks/DataCSV/ElectronicsCSV1.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULTILINE') AS (sno:chararray,id:chararray,reviewerID:chararray,asin:chararray,reviewerName:chararray,helpful:chararray,reviewText:chararray,overall:chararray,summary:chararray,unixReviewTime:chararray,reviewTime:chararray,category:chararray,class:chararray);
```

#After this command the data loaded was checked then check for NULL values present in the data through the following query. Though I checked the in initial exploration as well but it's tried here again as a requirement of this task.

5.
```
checkNULL = FILTER q1newFileData by NOT ((sno IS NULL) OR (reviewerID IS NULL) OR (asin IS NULL) OR (reviewerName IS NULL) OR (helpful IS NULL) OR (reviewText IS NULL) OR (overall IS NULL) OR (summary IS NULL));
```

#After checking the NULL values, I will now check the blank values in the dataset by the following query

6.
```
checkBlank1 = FILTER checkNULL by NOT  ((reviewerID =='') OR (asin =='') OR (reviewerName =='') OR (helpful =='') OR (reviewText =='') OR (overall =='') OR (summary ==''));
```

Once the blank values are confirmed, then we will check the 'N/A' values if present

7.
```
checkNA = FILTER checkBlank1 by NOT ((sno =='N/A') OR (reviewerID =='NA') OR (asin =='N/A') OR (reviewerName =='N/A') OR (helpful =='N/A') OR (reviewText =='N/A') OR (overall =='N/A') OR (summary =='N/A'));
```

#At this step we will store the values in newly created file.

```
8. STORE checkNA INTO '/Electronics_Clean_Data' USING
   org.apache.pig.piggybank.storage.CSVExcelStorage(',','YES_MULT
   ILINE');
```

#After success message of output creation, we would need to quit pig and get back on Hadoop. Where we will first locate those files which were created and then, merge those chunks of output files created by the pig cleaning process. Moreover, the 2nd command then stores the newly cleaned dataset file back to dataproc bucket.

```
9. hadoop fs -getmerge /Electronics_Clean_Data/ /home/
   hamza_qadeer2/CleanData/MergeCleanElectronics.csv

10.     hadoop fs -put MergeCleanElectronics.csv 'gs://dataproc-
   staging-europe-west1-773978562921-grrhd7uh/
   assignment1DataBucket/'
```