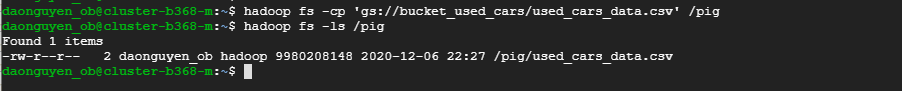
**PRE-PROCESSING DATASET on GOOGLE CLOUD PLATFORM (GCP)**

The original dataset used in our project is downloaded at <https://www.kaggle.com/ananaymital/us-used-cars-dataset>. Then this dataset is uploaded to the bucket name bucket\_used\_cars in our project namely CA675-A2-2020-new on GCP.

A Hadoop cluster namely cluster-b368-m on project CA675-A2-2020-new has been used as a working environment to implement data cleaning and querying with Pig and Hive, with following steps:

1. In Hadoop

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

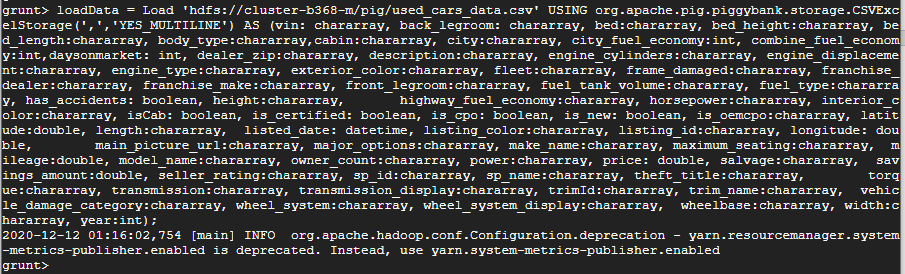


1. In PIG

It is the fact that the entry data contains a field with commas and line-break characters. Consequently, the functions such as CSVLoader and PigStorage did not work properly to handle that issue. Instead, CSVExcelStorage has been selected due to its support for loading multi line data. This function is available in the piggybank library which can be downloaded at <https://cwiki.apache.org/confluence/display/PIG/PiggyBank> and uploaded into Hadoop and registered in PIG as follows:

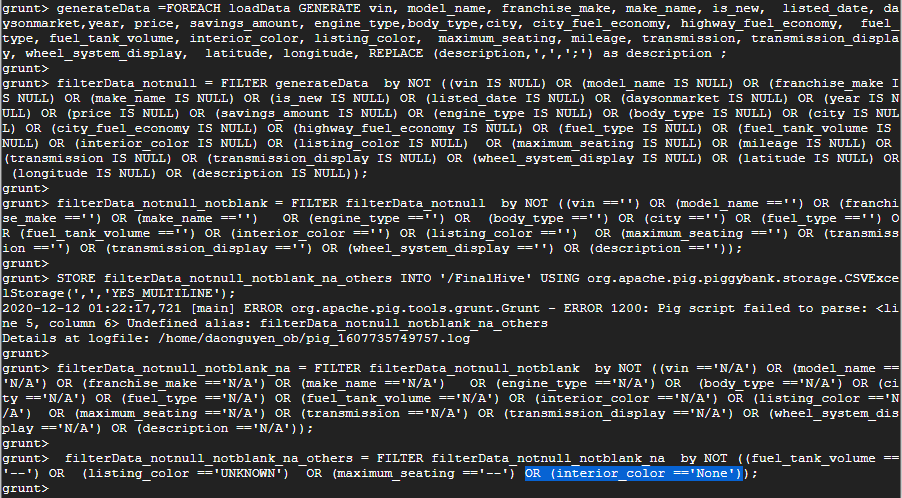
**register** piggybank.jar

The following command is used to load data from the five CSV files into Pig.



In the next step, the loaded data can be generated into a table with only necessary fields. Specially, the field Description has been cleaned using REPLACE functions. In particular, commas have been eliminated and replaced by semi-colon characters. It is an essential process so that the data can be properly loaded into Hive tables in the later task.

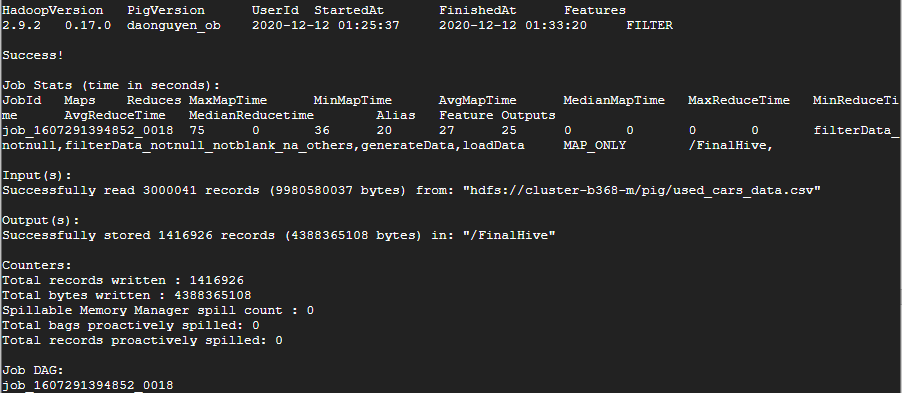
In addition, the generated data can be filtered to eliminate rows with at least one null/blank/NA/None/UNKNOWN/-- field



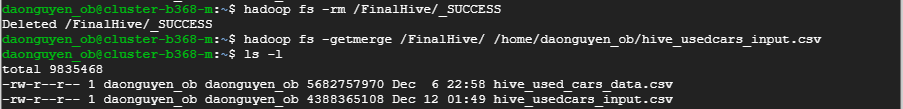
The filtered data can be stored into HDFS /FinalHive



The result can be seen at the below Figure.



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 need to be deleted, while part-m- files in /FinalHive should be merged into only file.



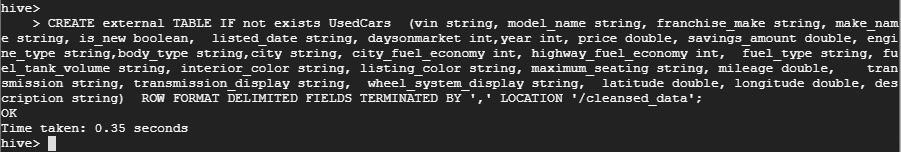
Next, the csv file can be put to a bucket and shared to public, which will support for connecting to BigQuery or visualisation tool.

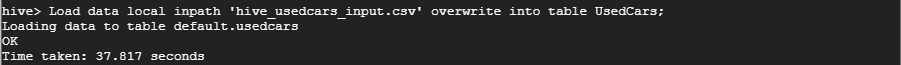
**hadoop** **fs -put** hive\_usedcars\_input.csv 'gs://bucket\_usedcars\_updated'

Now, the csv file can be accessed at the bucket namely bucket\_usedcars\_updated as well as the public link <https://storage.googleapis.com/bucket_usedcars_updated/hive_usedcars_input.csv>

1. In HIVE

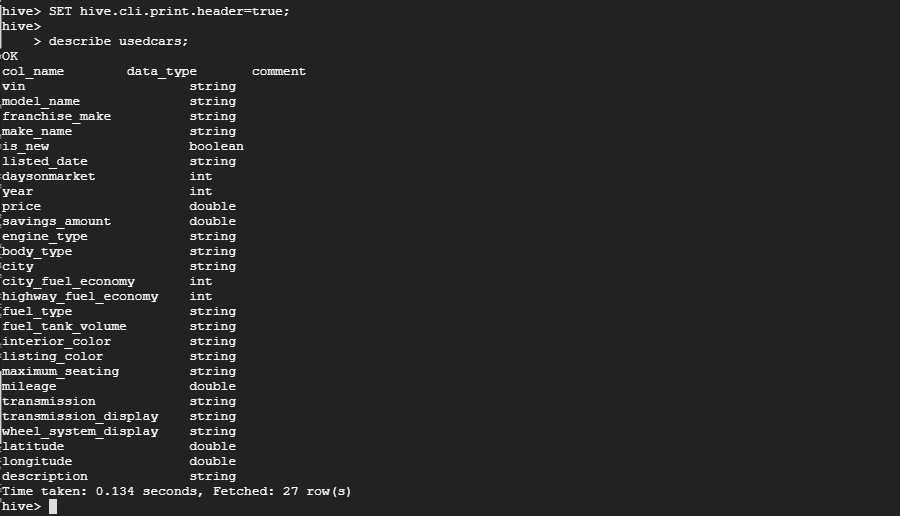
First, the filtered data in CSV files extracted from PIG can be loaded into Hive as follows.





Second, the column headers of printed tables can be presented by following Setup





Now, cleansed used car data has loaded in Hive table ‘usedcars’ already and be available for querying.