**ETL REPORT**

Completed By: Angela Spirou, Jing Wang, William Escobar & Richard Chounoune

**E**XTRACT:

The original datasets were three CSV files with data of several months’ worth of daily trending YouTube videos from the United States, United Kingdom and Canada. All datasets were downloaded from Kaggle at the following link: <https://www.kaggle.com/datasnaek/youtube-new>.

**T**RANSFORM:

All csv files were first read into a Jupyter Notebook in their raw forms and new DataFrames were created for each csv using python. The original csv files can be found saved within a folder named resources. In order to start cleaning the data, relevant columns were selected, renamed and added into new filtered DataFrames, with the video\_id set as the index for all three data sets.

**L**OAD:

A SQL Database was chosen due to the similarities within schematic structure of the csvs which is conducive of a relational database. Furthermore, with each dataset having unique IDs, PostgrSQL was used due to the allowance of using primary keys and foreign keys in order to assure congruency of data. Within PostgrSQL, new tables were created based off of the columns of the finalized DataFrames from the transformation step, with the video\_id data set as the primary key. Once the tables were created, an engine connection was initiated within Jupyter Notebook and all DataFrames were loaded into the database using the .to\_sql function.