Group 4 ETL Project

* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

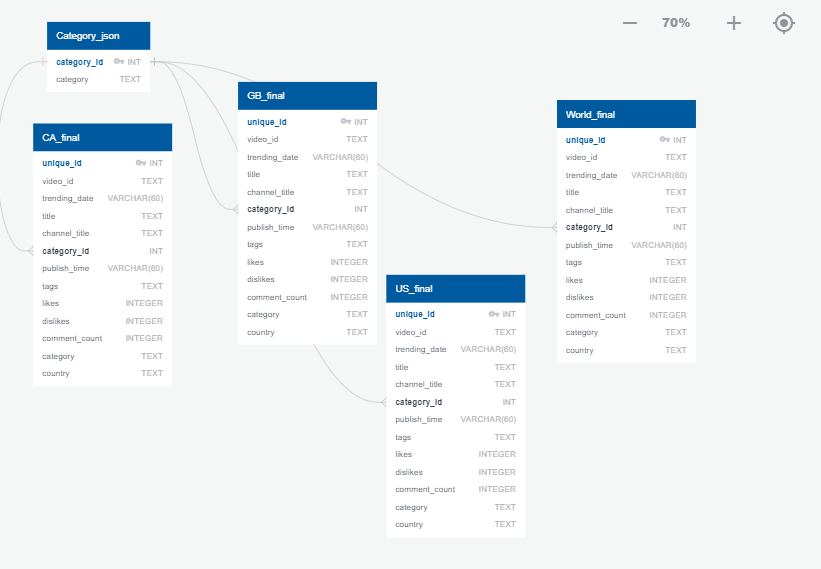
We started with the premise of what types of questions would we want to answer to obtain the data sets we would use for this project. With the pandemic prevalent during our class time and there being less in-person socializing, we thought it would be interesting to see how YouTube video categories and views had been impacted.

We did a search on Kaggle.com of youtube in their datasets and chose “Hottest” dataset by Mitchell J having the most votes at the time of 3260. The data presented was representative of 10 different countries, a .csv file and .json file of information for each. We narrowed our scope to the 3 countries that predominantly used the English language for ease of translation: United States (US), Canada (CA), and Great Britain (GB).

<https://www.kaggle.com/datasnaek/youtube-new>

* **T**ransform: what data cleaning or transformation was required.

ERD with Primary and Foreign key assignments



Jupyter notebook was used to transform the datasets through creating Panda dataframes, cleaning up columns, dropping unwanted data, and merging information from the .csv and .json files

* 1. Read in each of the datasets to Panda dataframe
  2. Drop unwanted columns
  3. Create a “for” loop to read the category title associated with each category id
  4. Merge the category title information from the .json data with the .csv data
     + Required removal of the previously defined index
     + Required datatypes to match so datatype conversion performed
  5. At this point column headers can be changed to accommodate desired format
  6. Create schemas in PgAdmin to accommodate the merged information
  7. Load the cleaned/merged data through PgAdmin to a final .csv file
* **L**oad: the final database, tables/collections, and why this was chosen.

We utilized PgAdmin to load our final database (because we were all more familiar with this software than MongoDB 😊). Postgres being relational made sense to use for this project since the data could be merged on a shared column “Category ID” across each country’s information. The data can be used to analyze youtube video popularity by categories and/or country as well as calculating % of likes and dislikes, etc.