**ETL PROJECT**

***Data sources***

Our project is primarily based on a Data set found in Kaggle, titled: *Trending YouTube Video Statistics.* This dataset includes several months of data on daily trending YouTube videos. Data is included for the US, Mexico and Canada regions (among others) with up to 200 listed trending videos per day. The data also includes a *category\_id* field, which varies between regions. To retrieve the categories for a specific video, it can be found in the associated JSON. One such file is included for each of the regions in the dataset.

***Extracting***

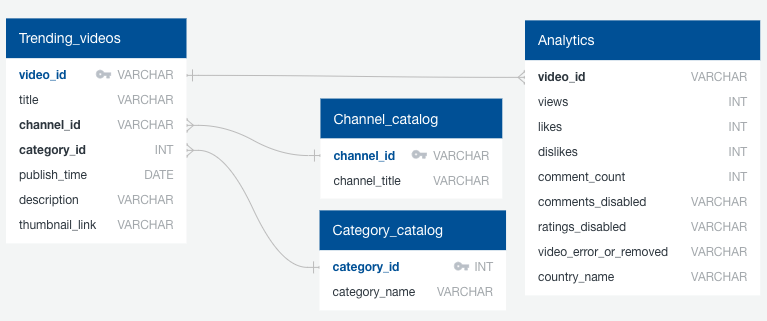
For starters, we decided to focus mainly on the region of North America in our analysis, including Canada, Mexico and US information for a better scope of trending videos in the area. For this, we needed to read a CSV file for each country and create a pandas data frame that shows the information gathered. Here, we found that the *category\_id* was numbered, and we needed to specify what each category number meant.

The information we needed was in the JSON files for each country, so we checked under the *ID*, *SNIPPET* and *TITLE* boxes to create a *Category Catalog* (Pandas Data Frame)for Mexico, USA and Canada. This is an example of the code used:



***Transforming Data frames***

In second place, we needed to transform the data to suit our needs. For this we needed to create different related DBs that could help us clearly understand the data we had. We accomplished this guided by the following schema:



So, we needed to start by creating a complete *Channel Catalog.* This was donecombining all data from the three different data frames, dropping channel title duplicates to generate our channel catalog, and finally creating the catalog and adding an index column.

Then we created our *Trending\_videos* Data Frame by merging our *videos* combined\_df from all the countries (done previously) with our *channel\_catalog* in order to add *channel\_id* and then dropping *channel\_title.* We also used the combined\_df to create an analytics table that only included categories important for statistical analysis.

Finally, we made sure that categories were the same across the three countries by creating a *Category Catalog* that merged all countries categories and dropped duplicates. This finally gave us the tools to start with the loading process.

**Loading Data frames into the database**