**ETL Project Report**

By George Oddoye & Vidhya.J

**Data Source**

Kaggle.com is the Data source for this project. We chose a dataset named “Trending YouTube video statistics”( <https://www.kaggle.com/datasnaek/youtube-new>). This dataset includes several months of data on daily trending YouTube videos. Data is included for the US, GB, DE and CA (USA, Great Britain, Germany and Canada respectively), with up to 200 listed trending videos per day.

Each region’s data is in a separate .csv file . The data also includes a category\_id field, which varies between regions. To retrieve the categories for a specific video, an associated JSON has been provided. One such file is included for each of the five regions in the dataset.

**Extraction**

Data was extracted from CSV files and JSON format files for 4 countries.

**Transformation**

1. Category id and category title for each record was parsed from the JSON files and stored in lists that were used to create a new data frame (df1).
2. CSV files were read into data frames and transformed (df2).
3. Time of video publication was originally in the ISO format. Time ( hour, min, sec) of video publication was parsed from original format.
4. A new published time column was be added to df2.
5. Dropping irrelevant columns , renaming and rearrangement of columns were done along with changing of data types to maintain uniformity
6. Data frames df1 and df2 was merged to present meaningful data.

**Loading**

Both data frames for 4 countries were loaded into the relational database – Postgres.

**Confirmation of data being added**

Confirmation of date being added was performed by querying one country( Germany) table .