**ETL Project Report:**

**Trending YouTube Video Statistics**

**Prepared by:** Grace Ming, Belinda Liang, Lucy Wo

* **Summary/Objective:**

Youtube, as the most popular video-sharing platform in North America and providing commercial opportunities to uploaders, is attracting analysts to dive into its data for analysis. The ***Trending YouTube Video Statistics*** including daily trending videos, likes, dislikes, etc., in 10 countries, is taken from kaggle.com. Another dataset resource is ***YouTube API Video Category ID List*,** which could be an assistance to better understand what the YouTube categories are. The datasets have been provided as CSV and JSON files respectively which need to be transformed to the same format for further analysis. Python is the main tool to manipulate the data, such as retrieving data from different formats, cleaning data, merging datasets, uploading cleaned data to a relational database(PostgreSQL).

* **Data Sources:**
  + ***Trending YouTbe Video Statistics***

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

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US and CA regions, with up to 200 listed trending videos per day. This dataset contains all the details of the trending YouTube videos along with its likes, dislikes, comments, tags and views for each video for a particular year.

* + ***YouTube API Video Category ID List***

<https://gist.github.com/dgp/1b24bf2961521bd75d6c>

* **Process:**

1. **Determining the object**

* YouTube views in Canada and in the United States from mid of 2013 to the end of 2017

1. **Finding dataset, retrieving data and uploading into Jupyter notebook**

* Download dataset from Kaggle and from Github
* Formats: CSV files (Kaggle.com) and JSON document (Github)

1. **Extract - Displaying data through Pandas dataframe**

* Retrieving data from data files,
* Reading the details of the documents.

1. **Transform - Generating new data with subset columns and cleaning data frame**

YouTube CA and US dataset:

* Removing unwanted columns
* Formatting the date to year\_month format
* Resetting index by *year\_month*, and *category\_id* , then displaying by the sum of the month.
* Merging the Canada monthly youtube views with Americas’ on year\_month
* Renaming the column names to a clearer format

Github dataset:

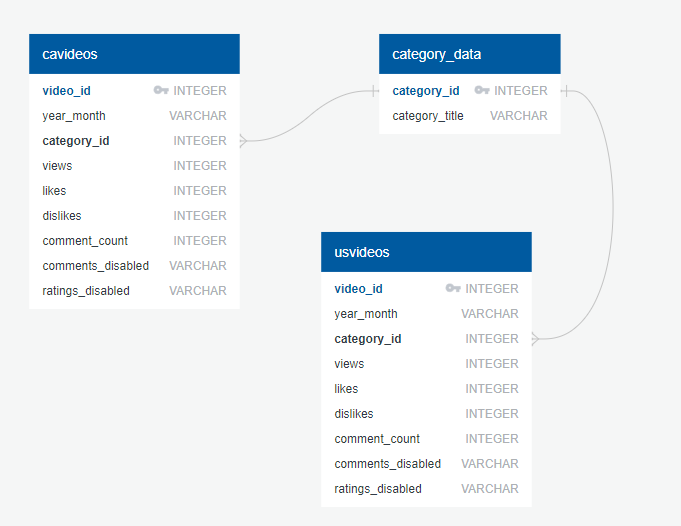
* Reading JSON file and retrieving data of the category numbers
* Creating a new object which is displaying the category information

Merging two datasets on *category\_id*

1. **Analysis**:

* Using matplotlib to plot lines and to visualize the differences between Canada and the United state from 2013 to 2018
* Comparing: views, likes, dislikes, comments and views by category

1. **Load - Uploading data to Postgres database:**

****

* Creating a database in PostgreSQL for storing data
* Connecting to the local database
* Uploading cleaned data to the database
* **Conclusion**:

YouTube views in the United States had slightly higher numbers of views, likes, dislikes and comments overall than that in Canadia, except for that during the end of 2017.

Problems: Difficulty in finding datasets, which does not share much common with what the YouTube resources had. Therefore, it is hard to compare the popularity among different video-sharing platforms (eg: Netflix, Amazon Prime Video) by looking into the views, likes, dislikes, etc.