ETL Project  
compiling a data lake of youtube and billboard trending stats

November 2018

**Team Members:** Erica Rosa Thach Tran

## Project Background and Description

We prepared a data lake with trending statistics from YouTube and Billboard, which could be used to uncover relationships between most popular songs and related YouTube content over time.

## Extract

Data was loaded into a Jupyter Notebook for analysis and transformation.

**Data Source:** US Trending YouTube Video Stats | 11/14/17 – 6/13/18 (<https://www.kaggle.com/datasnaek/youtube-new>)

**Format:** CSV

**Data Source:** YouTube Video Categories

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

**Format:** JSON

**Data Source:** Billboard Hot 100 for 11/14/17

(<https://www.billboard.com/charts/hot-100>)

**Format:** HTML

## Transform

A dataframe was created from the YouTube Trending Stats CSV. Python was used to navigate the YouTube Categories JSON file; and pandas was used to create a dataframe of the Category ID and Category Title. A pandas merge was performed on the two YouTube dataframes to create a comprehensive table of YouTube video trending stats and their respective categories. This final dataframe was filtered to include only relevant columns. Date objects were set as datetime types.

Splinter and the Chrome driver was used to access the Billboard Top 100 site. BeautifulSoup was used to parse through the html and create lists of the artist and song title. These lists were transformed into a pandas dataframe.

## Load

The final database was created in MySQL to efficiently and reliably store the data in tables. SQL Alchemy was used to create the engine to connect the Jupyter Notebook to MySQL. Three tables were created:

* trending\_table : the final YouTube Dataframe
* trending\_table\_111417 : the YouTube Dataframe for the date 11/14/17
* billboard\_table: the Billboard Top 100 dataframe (for the date 11/14/17)

## Limitations:

The “title” field of the YouTube data was excluded, due to strange characters that were not able to be processed by the to\_SQL function. If this field were able to be imported into the MySQL database, it could be used to enhance analysis of correlations between Billboard top songs and YouTube video trends.

## Conclusion

The final loaded MySQL database can be used for further analysis, to observe trends in YouTube categories with the most uploads, views, comments, likes, and dislikes. If the title field were available, the YouTube data could be compared to the Billboard data to observe trends in YouTube videos related to popular songs for a certain date range.