ETL Project

YouTube Data Analysis with Data Sourced from Kaggle



Project Team: Murali Calambur & Santosh Singh

Submission of Murali Calambur

TAble of  
Contents

**Introduction………………………………………………………………….1**

**Data Source.……………………….………………………………..……….2**

**Data Description.………..……….………………………………..……….2**

**Data Conversion.………..……….………………………………..…….….2**

**Data Loading…….………..……….………………………………..……….3**

**Final Analysis & Conclusion.………..……….……… ………..……….3**

## INTRODUCTION

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year”. Top performers on the YouTube trending list are music videos (such as the famously virile “Gangam Style”), celebrity and/or reality TV performances, and the random viral videos that YouTube is well-known for. This dataset is sourced from Kaggle.

It is a daily record of the top trending YouTube videos. The Kaggle dataset includes several months of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

Possible uses for this dataset could include:

* Sentiment analysis in a variety of forms
* Categorizing YouTube videos based on their comments and statistics.
* Training ML algorithms like RNNs to generate their own YouTube comments.
* Analyzing what factors affect how popular a YouTube video will be.
* Statistical analysis over time

In this projects, we focused on loading the data for North America [US & Canada]. Focus was on the clean-up & loading. A small analysis was performed as demonstration that the data was loaded successfully.

## Data SOURCES

Kaggle:

**Source:** [**https://www.kaggle.com/datasnaek/youtube-new/data**](https://www.kaggle.com/datasnaek/youtube-new/data).

This dataset was collected using the YouTube API and downloaded from Kaggle for analysis.

## Data DESCRIPTION

Data is available is split into a CSV dataset and a JSON datasets. CSV contains the following data: Columns: [video\_id, trending\_date, title, channel\_title, category\_id, publish\_time, tags, views, likes, dislikes, comment\_count, thumbnail\_link, comments\_disabled, ratings\_disabled, video\_error\_or\_removed, description]

The JSON includes a category\_id field, which varies between regions. To retrieve the categories for a specific video, it has to be parsed from the associated JSON. One such file is included for each of the five regions in the dataset.

The category\_id field in JSON can be used to join the category\_id in the CSV to get a description of the category of the video.

**Project focus was on the North America region: USA & Canada**

CSV & JSON Files were loaded into 2 data frames

## Data TRANSFORMATION

A number of transformations were done on the data:

1. Country named was added to initial data frames to relate the data loaded from the country specific file.
2. The US & Canada data frames were merged
3. Records that showed videos that had errored out were removed
4. A new dataframe with columns needed for analysis was created
5. A series of manipulations were done on the **trending\_date** column:
   1. A fixed string of “20” was added to the “trending\_date” as a string field to convert date from YY.DD.MM format to YYYY.DD.MM
   2. The string field was modified and its format was changed in the data frame from YYYY.DD.MM to YYYY/DD/MM
   3. The YYYY/DD/MM was converted to date field as “YYYY/MM/DD”
6. The JSONs were parsed to extract the category\_id and category\_title using Python. The script identify the category\_id field and category\_tile from the “items” list

## Data LOADING

The data sets were loaded into a Postgres database. A relational database was selected to make it easier to join & query the data from the tables.

1. A database "youtube\_db" was created in Postgres
2. Two tables "youtube\_data" and "youtube\_category" were created in this database
3. The CSV data and category data from the JSON were loaded into the tables

## FINAL ANALYSIS & CONCLUSION

An analysis of the data was done to sample if the data was loaded successfully. A query to see total views by category in descending order in USA was executed through the Postgres. It showed that Music was the highest viewed category, followed by Entertainment and Film and Animation. It is clear that YouTube viewers look to the application for music than pets, despite the popular notion that cat videos on YouTube are popular!

A query to see the top most viewed video showed that from May 19th to June 2nd of 2018, 'Childish Gambino - This Is America’ was the most watched video.

This demonstrated that files were cleaned properly and loaded into the database for further queries.

## APPENDIX

1. Jupyter Notebook jointly done by Murali Calambur & Santosh Singh: youTube\_Statistics.ipynb file
2. Postgres “youtube\_db” database tables schema:

CREATE TABLE public.youtube\_data

(

video\_id character varying COLLATE pg\_catalog."default",

trending\_date date,

title character varying COLLATE pg\_catalog."default",

category\_id integer,

views integer,

likes integer,

dislikes integer,

comment\_count integer,

country character varying

)

CREATE TABLE public.youtube\_category

(

category\_id integer,

category\_title character varying COLLATE pg\_catalog."default"

)