**YouTube ETL Project**

Team 3 – Armin, Hayden, Lovjeet, Karan

**PROJECT GOAL:**

The aim of this YouTube (YT) ETL project is to use old and new data sources to:

* Extract most trending YouTube video data from data sources
* Transform and Format to prepare filtered datasets
* Load datasets to make final production database of YouTube trending videos with Categories, Channel title, Subscribers, Network, Language, and Channel ID

**EXTRACT:**

For this project, we have used three data sources.

* HTML (web scrapping) – Wikipedia, Techpostplus
* APIs – YouTube Data API
* CSV files

HTML(web scrapping)

* 1st source of HTML table is ‘List of Most Subscribed YT channels’ from Wikipedia
* 2nd source of HTML table is ‘List of Most viewed YT channels’ from Wikipedia
* 3rd source of HTML table is ‘List of YT video categories’ from Techpostplus.com

APIs (YouTube Data API)

* We used YouTube scrapping code (available online) with our API key to extract daily data on trending videos
* The Script is written to output the extracted data using API in to CSV file
* One CSV file was created from 16 countries after merging all by command line

CSV files

* We used CSV file from ‘Trending YouTube Video Statistics’ Dataset on Kaggle
* Datasets were available for multiple countries but for the simplicity and limit the amount of data, we used dataset of Canada

**TRANSFORM:**

We used two csv files, one from Kaggle and the one we created using API. Below are the steps we used to organize the data for production database

* Dropped unnecessary or redundant columns
* Renamed and reordered columns
* Added country code column
* Used concat to merge our international data with Canadian data
* Used split to remove unwanted elements in web scrapped data (HTML) from Wikipedia
* Joined all tables with channel titles with an outer join to create a dataframe of all Youtubers represented in the data (to be used as a primary key)

**LOAD:**

We used PostgreSQL because all of our datasets are very structured in a way so that we can use primary key and foreign key constraints to get effective query results. We created five tables, which are listed below.

* categories
* trending\_youtube
* most\_viewed\_youtubers
* most\_subscribed\_channels
* channel\_title

channel\_title is the unique identifier for any youtube channel and category\_id is also the unique number hence we selected them as primary keys to map them with other tables. The ERD of table structure is shown below.

