# **ETL Project**

### **What is ETL?**

ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL full-form is Extract, Transform and Load.

Flat files

CSV/JSON

Staging Area

Transform

PostgreSQL

ETL Process

Teradata

SQL Server

Oracle

**Data Description:**

YouTube 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).

This dataset is a daily record of the top trending YouTube videos.

**Data Content:**

This dataset includes several months of data on daily trending YouTube videos. Data is included for the USA and Canada regions with up to 200 listed trending videos per day.

Each region’s data is in a separate csv and json file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

The data also includes a category\_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON.

**Extract:**

**Extract flat files from Kaggle.com**

* US Video csv
* Canada Video csv
* US Video Category json

**Transform:**

* Joining Datasets:
* Add Country column to differentiate data between US and Canada
* Merge US and Canada video datasets
* Data Cleaning:
* Filter error videos from merged dataset
* Create new dataset with the specific columns

**Load:**

* Create “youtube\_db” database on PostgreSQL database
* Create “youtube\_data” and “youtube\_category” tables in “youtube\_db” database
  + Define table column and datatype
* Load data from dataframe to PostgreSQL database