Assignment 1: Data Lake with Snowflake

Simon Lim

24661225

Big Data Engineering

# CONTEXTS AND OBJECTIVES OF THE PROJECT

YouTube is the worldwide video sharing platform, which share various categories and top trending videos on the platform. Top trending videos are evaluated based on video’s number of views, shares, likes and comments etc. A dataset with a daily record of the top popular and trending videos has been extracted through YouTube API. In this regard, the objective of the project is to analyse the dataset, which consists of CSVs and Jsons files, by using a Data Lakehouse with Snowflake. The procedure of the project includes data ingestion, data cleaning, data analysis and answering business questions. Snowflake is a main data warehouse and SQL platform used in this project, in order to load and analyse data.

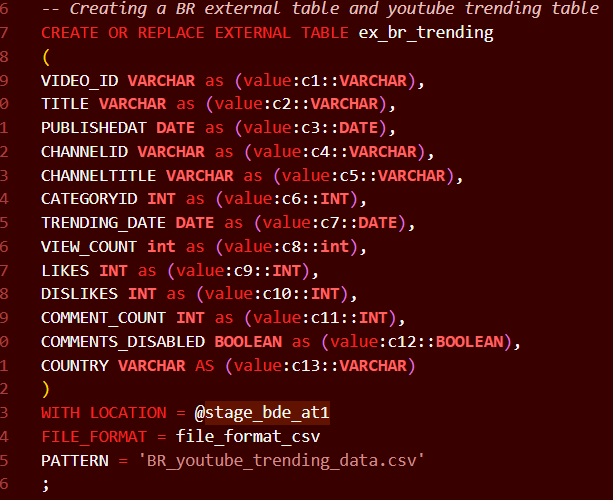
# PRESENTATION OF DATASET

There are two types of dataset files, one with CSV file and the other with Json file. For CSV file, dataset includes several months of daily trending YouTube video records for 11 different countries, including India, USA, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea and Japan. Hence, there are 11 CSVs files representing top trending videos of each country. Various data of daily video records are included in dataset, such as video\_id, title, published date, category id, number of views, likes, dislikes, channel id, comment counts and trending date etc.

For Json file, dataset also include category id, title and channel id for each country.

# PIPELINE OF THE PROJECT WITH ISSUES

1. Before ingesting data, there are two necessary processes to be set up. The first step was to set up a cloud storage account on Microsoft Azure and snowflake account was a necessary process. Subsequently, it was also necessary to set up a Storage Integration between Azure and Snowflake.
2. **Data Ingestion** 
   1. Stage, named ‘stage\_bde\_at1’, was created to store, load and unload data files.
   2. For CSVs files, external tables for each country were created, adding the new column ‘country ’. (11 External Tables). Columns in external tables include video\_id, title, publish date, channel id, channel title, category id, trending date, view count, likes, dislikes, comment count, comment disabled and country.



**Figure 1. An External Table for Brazil**

Then, a new table, ‘table\_youtube\_trending’ was created to store all data received from external tables. Subsequently, null values in the new column ‘country’ were updated after data for each country were inserted. External tables with same columns for other countries were all inserted into the table ‘table\_youtube\_trending’, in conjunction with updates of ‘country’ column.

* 1. For Json files, I manually added a country column and each corresponding country value. The rationale behind it was that it was possible to load 11 Json files as each row in an external table, helping to simplify code. Hence, an external table for 11 countries was created. Then, a new table ‘table\_youtube\_category’ was created to store data from the external table, including three columns, ‘country’, ‘category id’ and ‘category title’.
  2. The last stage in data ingestion was to create a final table by combining two tables. Before combining them, it was necessary to set up primary keys and foreign keys (category id, country) for tables, as those two columns are common columns two tables included.
  3. Finally, two tables were combined using left join on country and category id and created a final youtube table, “table\_youtube\_final”.



**Figure 2. Combining two tables**

1. **Data Cleaning**

In this stage, null values and duplicate values were cleaned in table\_youtube\_final table.

* 1. In “table\_youtube\_category”, only “Comedy” has duplicates if we do not take into account the categoryid.
  2. In “table\_youtube\_category”, only “Nonprofit & Activism” appears in one country.
  3. Categoryid, which has missing values in category\_title, is ‘29’.
  4. Null values (3161 rows) in category\_title were filled with category id ‘29’.
  5. There is one video, which has a null value for channeltitle.

A screenshot of a computer

Description automatically generated

* 1. All records, containing ‘video\_id’ of “#NAME?” were all deleted (15831 rows were deleted).
  2. A duplicate table, “table\_youtube\_duplicates”, containing duplicates of a combination of video\_id, trending\_date and country, was created.

A screen shot of a computer

Description automatically generated

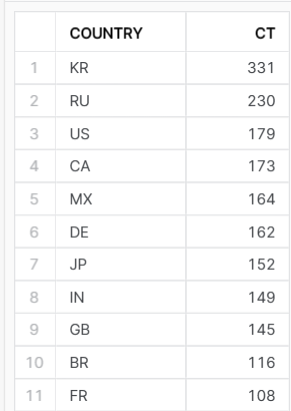
* 1. Finally, the duplicates in “table\_youtube\_final” were deleted using “table\_youtube\_duplicates”.
  2. Final number of rows in table\_youtube\_final were 1123017.

1. **Data Analysis**
   1. In “table\_youtube\_final”, three most viewed videos in the “Sports” category and for the trending\_date = “2021-10-17” for each country were analyzed and displayed.

A screenshot of a computer

Description automatically generated

* 1. The number of distinct video with a title containing the word “BTS” for each county was displayed.



* 1. The most viewed videos and their like\_ratio were displayed with their year and month and for each country.

A screenshot of a computer

Description automatically generated

* 1. For each cointry, videos with a category\_title, that has the most distinct videos and their percentage out of the total distinct number of videos of that country was displayed.

A screenshot of a graph

Description automatically generated

* 1. Channeltitle, that has the most distinct videos and its number were displayed.

A screenshot of a calendar

Description automatically generated

# ANSWERING BUSINESS QUESTIONS

1. If I was to launch a new YouTube channel, which category (excluding “Music” and “Entertainment”) of video would I be trying to create to have them appear in the top trend of YouTube?

* I would create ‘Gaming’ category, as it has the highest total\_view\_count (195569682751) out of all categories.

A screenshot of a computer

Description automatically generated

1. Will this strategy work in every country?

* No, unfortunately it does not work in every country. There are many countries, in which “People & Blog” have higher total\_view\_count than “Gaming” category, including Japan, Korea, India, Russia, Brazil etc. Therefore, it is important to create a category of video, depending on country and other potential variables, such as age and gender.

A screenshot of a table

Description automatically generated

# REFERENCE

*YouTube Trending Video Dataset (updated daily)*. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset

‌