An Analysis of YouTube Trending Data With snowflake

By Esther Csoke 24542312

Project Overview

YouTube

Youtube is a free video-sharing platform created in 2005 and is the top-visited website on the web, with an average of 8.2 billion monthly visits. The trending page of Youtube page considers multiple factors when determining what is selected to be on the trending page. Factors such as likes, comments, views and shares that YouTube collects quantify user interaction.

Tools
Snowflake, SQL, Microsoft Azure

Data Types
CSV, JSON

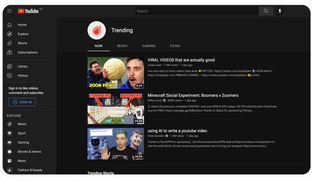

Figure 1. Screenshot of the YouTube trending page

The project aim is to use several months of Youtube trending data **from 12-08-2020** to the current date to set up a data lakehouse using Snowflake, clean the data and extract valuable insights. And to ultimately analyse what category (excluding Music and Entertainment) of YouTube video is best to create that will appear on the trending page.

The Data

Data includes several months from the 12-08-2020 to current date.

It includes values from the following 11 countries:

- 1. India (IN)
- 2. United States of America (US)
- 3. Great Britain (GB)
- 4. Germany (DE)
- 5. Canada (CA)
- 6. France (FR)
- 7. Russia (RU)
- 8. Brazil (BR)
- 9. Mexico (MX)
- 10. South Korea (KR)
- 11. Japan (JP)

Each region is uploaded into a separate file.

There are 2 different types of data files:

1. YouTube Trending Date

This data is in 11 seperated CSV files and contains the following values:

- `video_id` – An ID for the YouTube video (VARCHAR)
- `title` – Title of the YouTube video (VARCHAR)
- `publishedat` -The date when the YouTube video was published (DATETIME)
- `channelid` - The YouTube's channel ID (VARCHAR)
- `channeltitle`- The YouTube channel title (VARCHAR)
- `categoryid`- The Id of the category the YouTube video is associated with (INT)
- `trending_date` – The date of the video being on the trending page (DATE)
- `view_count` – The amount of views the video has. (INT)
- `likes` – The amount of likes the video has (INT)
- `dislikes`- The amount of dislikes of the video (INT)
- `comment_count`- The amount of comments on the video (INT)
- `comments_disabled` – Whether comments were enabled on the video (BOOL)

2. YouTube Category Data

The data is in 11 JSON files seperated by country and features the following values needed:

- `categoryid` - An id of assosiated with a category title (INT).
- `category_title` - A YouTube category title (VARCHAR).

Overall Architecture

Azure Storage: Used to store data outside the Snowflake environment and is connected via a storage integration

Snowflake: Used to create a data lakehouse, referencing data from Azure Storage cloud via storage integration.

SQL Used in Snowflake to query and perform analysis.

A basic flow of the process is shown below:

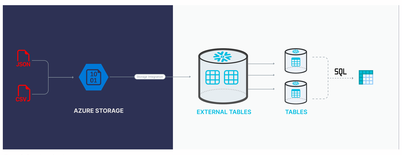

Figure 2. Overview of project

Data Ingestion

Downloading files to Azure Storage

The data sets were retrieved via Google drive (links are accessible in Appendix 1) and uploaded to a newly created container in Azure storage called utscontainer. A personal Azure directory Id was acquired to begin storage integration between Azure and Snowflake.

The data consisted of:

- 11 CSV files for the trending data consisted of YouTube's main video factors.

- 11 JSON files for the categories YouTube uses on the trending page.

Storage integration between Azure and Snowflake

In Snowflake, a new database was created and the storage integration was set up with Azure by generating an Azure directory id. The storage integration initialised the retrieve Azure consent URL to accept permissions. Afterwards, a new role assignment called 'Storage Blob Data Owner' was assigned in Azure with Snowflakes generated Azure multi-tenant app name.

Loading Data from Azure to Snowflake

A new stage was set up using the storage integration setup. The **external tables** were set up for both the CSV trending data and the JSON categories data. Both external table setups brought some challenges that are listed below.

Trending Data

The challenge when ingesting the external table called `ex_table_youtube_trending` for the trending data was that initially, the external table had no exact column names, as shown below:

Figure 3. Example of intial trending external table output

Therefore, a **file format** was created to help load the CSV files. Additionally, the values had to be renamed. Extra challenges were faced in the overall syntax of creating the table and the order of declaring the values and transforming the external table using the file format.

The values did not include a country column from the individual country files. Another column was created with the metadata of the file names using `metadata$filename`, and a new table was created from the external table called `table_youtube_trending`.

Category Data

An external table `ex_table_youtube_category` was that initially the external table had no column names and values that were needed were in nested json.

Figure 4. Example of nested JSON in category external table

Creating the external table was a challenge with flattening the data. Upon reading the documentation, the table had to be flattened twice using the argument `LATERAL FLATTEN` and the keys' selectors to acquire the data needed. This table was called `table_youtube_category`.

Finally, `table_youtube_category` and `table_youtube_trending` were joined together using a LEFT JOIN on their `country` and `categoryid` columns. While also creating an `id` column for all rows.

Data Cleaning

The rest of this handover focuses on various SQL queries performed for data cleaning and analysis in Snowflake.

`1.table_youtube_category` was checked for duplicates in the `category_title` without accounting for `categoryid`. The category title of 'Comedy' had duplicates in all countries.

2.Secondly, `table_youtube_trending` was checked for category titles only appearing in one `country`. One category of **Nonprofits and Activism** only appeared in the country of the **US**.

`3. table_youtube_final` was checked if any missing `categoryid`**'s** had a missing `category_title`. The category of **29** had the category of NULL.

4. The NULLS of `category_title` from `categoryid` 29 were replaced with the value of 29 for data hygiene. 3162 rows were updated.

5. One video was found to have a missing `channeltitle`. The video has the `ID` of 596a98d8-4a78-4d75-b31d-e82d461f2ec7

ID	VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE
596a98d8-4a78-4d75-b31d-e82d461f2ec7	9b9MovPPewk	Kala Official Teaser | Tovino Thomas | Rohith V S | Juvis	2021-01-21T12:30:29Z	UCOQtOy-FCJLRdwfiNLiqPFZA	NULL

Table 1. The output for the missing channeltitle

A large number of `video_id`'s had the value of **$NAME?** in table_youtube_final. 14619 rows were deleted containing this `video_id`.

Dealing duplicates in TABLE_YOUTUBE_FINAL

`table_youtube_final` appeared to have many duplicates with the same `video_id, country and trending_date`. The duplicates were ranked by highest `view_count`. A new table, `table_youtube_duplicates` was created to contain duplicates that weren't ranked 1. The rows in the `table_youtube_duplicates` were 37842. The rows that were present in `table_youtube_duplicates` were **removed** from `table_youtube_final`. Lastly `table_youtube_final` was checked to confirm it's row number and equaled to 1,123,017 rows.

Data Analysis of table_youtube_final

1. Videos in the Sports category for the date of '2021-10-17'

- The 3 moved viewed videos for this category and date were found by their country and rank.

2. Distinct BTS videos

- The table below shows the count of **distinct** videos with a `title` containing the word 'BTS'. These results are ordered by descending order.

Country	CT
KR	331
RU	230
US	179
CA	173
MX	164
DE	162
JP	152
IN	149
GB	145
BR	116
FR	108

Table 2. Output for distinct videos featuring the title BTS

3. Top viewed videos and likes ratio

It was checked what were the top videos by `country, year and month`. An example of the output is below:

COUNTRY	YEAR_MONTH	TITLE	CHANNELTITLE	CATEGORY_TITLE	VIEW_COUNT	LIKES_RATIO
BR	2020-08-01		Big Hit Labels	Music	2445507902	6.52
CA	2020-08-01		Big Hit Labels	Music	2326649205	6.76
DE	2020-08-01		Big Hit Labels	Music	2191104491	7.06

Table 3. Output for top videos by country, year and month

4. Percentage of Distinct Videos

- For each `country` it was queried what `category_title` has the most distinct videos and another column was created based on the percentage of that most distinct video `category_title` by **total distinct videos**. The results are displayed in the table below.

COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
BR	Entertainment	4293	16371	26.22
CA	Entertainment	4313	20807	20.73
DE	Entertainment	6679	25299	26.4
FR	Entertainment	5297	22096	23.97
GB	Entertainment	4511	20472	22.03
IN	Entertainment	12839	29431	43.63
JP	Entertainment	4945	14816	33.38
KR	Entertainment	4625	13457	34.37
MX	Entertainment	3628	15347	23.64
RU	Entertainment	10400	63877	16.28
US	Entertainment	3812	19130	19.93

Table 4. Output showing percent of distinct videos by country in highest category

5. Most Distinct Videos

It was checked what channeltitle has the most distinct videos and what was the number. The results below show the query result.

CHANNELTITLE	DISTINCT_COUNT
Colors TV	805

Table 5. channeltitle with most distinct videos

Business Question

What Category of Video To Create?

The main goal is to analyse what video category is optimal to create to get into the top trending page of YouTube. The categories Music and Entertainment are excluded from the analysis. While YouTube's trending algorithm is far more complex than the scope of a SQL analysis, however, a few analyses were conducted. It was first checked by what category had the highest viewer engagement on a video's first trending day.

The category with the highest average likes on the first day of trending was Non-profits and Activism as shown below:

CATEGORY_TITLE	AVG_LIKES
Nonprofits & Activism	147023.117647

Table 6. Output for highest average likes on first day of trending

Additionally, the category with the greatest average `view_count` on the first `trending_date` was Nonprofits & Activism.

CATEGORY_TITLE	AVGVIEWS
Nonprofits & Activism	1350755.235294

Table 7. Output for highest average view_count on first trending date

However, this category only appears in the US and be reflective of current political events. Therefore, the analysis went further. The count of how many times a category has been on the trending page is analysed. The count was analysed based on the YouTube trending algorithm has been argued to be complex. Therefore, viewer engagement metrics may have performed better in these categories to appear more frequently overall. The table below shows that the highest performing category is People & Blogs.

CATEGORY_TITLE	TRENDING_TIMES_COUNT
People & Blogs	134130
Gaming	122123
Sports	112916

Table 8. Top three highest appearing trending categories

As shown below, the results were grouped by country. The **People and Blogs** `category` has the most consistent number of the highest videos per country. Therefore, choosing this category may be the optimal solution. However, it does not work in every country and may be due to entertainment consumption and metrics differences.

Country	MOST_FEATURED_CATEGORY
US	GAMING
GB	SPORTS
CA	GAMING
RU	PEOPLE & BLOGS
KR	PEOPLE & BLOGS
JP	PEOPLE & BLOGS
BR	PEOPLE & BLOGS
MX	PEOPLE & BLOGS
IN	PEOPLE & BLOGS
DE	SPORTS
FR	GAMING

Table 9. Highest appearing trending categories by country

Appendix

Appendix 1:

Trending data:

https://drive.google.com/file/d/1bsRxgSTXenOhKCjN3nSqmisy9aMokdeW/vie w?usp=sharing

Category data:

https://drive.google.com/file/d/13818ZbLMSpCNHR9CO3Ecty7iv_-HEHhx/vie w?usp=sharing