Overview

YouTube is one of the most popular video-sharing platforms in the world. It maintains a list of the top trending videos on its website and updates it almost every 15 minutes. During each update, the videos move up, down or stay in the same position in the list. The top-trending videos are determined by a combination of the factors such as number of views, likes, shares and comments. The aim of this project is to take the dataset of the top trending YouTube videos and analyse it and finally make a recommendation about which category of video to create to make them appear in the top trending videos list on YouTube.

Data Understanding

The top trending YouTube videos dataset was downloaded from Kaggle, which had been originally retrieved using the YouTube API. The dataset contains daily trending videos data from 2020-08-12 to 2022-01-28.

The dataset has 11 separate csv files for 11 different regions. All the regions also have a corresponding category_id json file which has information about the video categories. Following table contains the name of the files for all the regions in the dataset.

Region Code	**Region**	**Trending data (.csv)**	**Category data (.json)**
BR	Brazil	BR_youtube_trending_data.csv	BR_category_id.json
CA	Canada	CA_youtube_trending_data.csv	CA_category_id.json
DE	Germany	DE_youtube_trending_data.csv	DE_category_id.json
FR	France	FR_youtube_trending_data.csv	FR_category_id.json
GB	Great Britain	GB_youtube_trending_data.csv	GB_category_id.json
IN	India	IN_youtube_trending_data.csv	IN_category_id.json
JP	Japan	JP_youtube_trending_data.csv	JP_category_id.json
KR	South Korea	KR_youtube_trending_data.csv	KR_category_id.json
MX	Mexico	MX_youtube_trending_data.csv	MX_category_id.json
RU	Russia	RU_youtube_trending_data.csv	RU_category_id.json
US	United States of America	US_youtube_trending_data.csv	US_category_id.json

Table 1: File names for all the 11 regions in the dataset

The trending data contains relevant information about the video including the video title and views. Following table shows the data dictionary

Field	Data Type	Description
video_id	String	ID of the video
title	String	Title of the video
publishedAt	Date	The date at which the video was uploaded
channelId	String	ID of the channel
channelTitle	String	Name of the channel
category_id	Integer	ID of the category of video
trending_date	Date	The data at which the video was trending
view_count	Integer	Number of views the video received
likes	Integer	Number of likes the video received
dislikes	Integer	Number of dislikes the video received
comment_count	Integer	Number of comments made on the video
comments_disabled	Boolean	TRUE if the comments were disabled, otherwise FALSE.

Table 2: Data Dictionary of trending data

The category data for each region contains the relevant information for all the categories including the category ID and category title.

Figure 1: Snippet of the category data

Overall Architecture

The overall architecture of the project is shown in Figure 2.

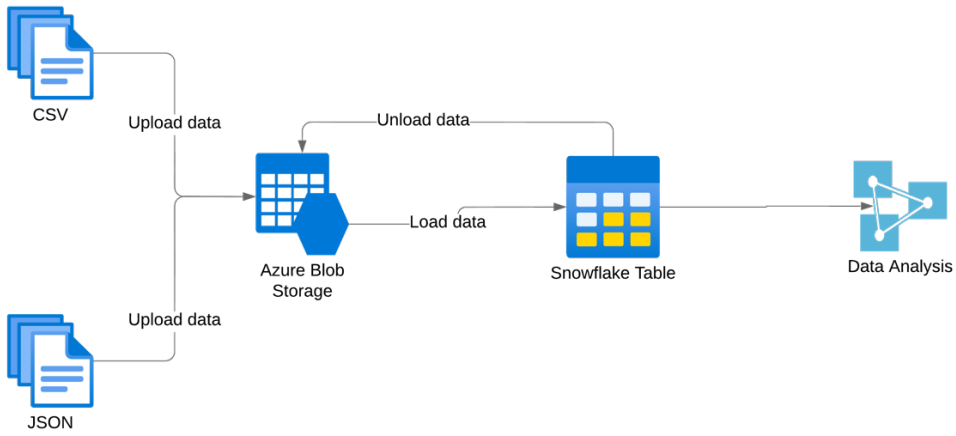

Figure 2: Overall Architecture

The overall architecture of the project can be described in 3 parts. They are:

1. Azure Blob Storage: At the start of the project, all the data downloaded from Kaggle is uploaded to the Azure Storage Container to be later loaded into Snowflake.
2. Snowflake Table: After the data has been uploaded to the Azure Storage Container, the data is at first transferred to an external table and then loaded into a (internal) Snowflake table.
3. Data Analysis: Once the data has been loaded into the Snowflake table, the data is cleaned, transformed, and prepared for analysis. The data is then examined to find valuable insights.

Setup and Data Ingestion

The project makes use of Azure Cloud Storage and Snowflake to ingest the data into a Data Lakehouse and then transform it and finally do some analysis on it.

Pre-requisites:

- Setup an account with Microsoft Azure
- Create a Storage account in Azure Portal
- Setup a Snowflake account

Data Ingestion:

- Download the YouTube trending and category data from Kaggle: YouTube Trending Video Dataset
- Create a new storage container in Azure Portal
- Upload all the downloaded .csv and .json files to the newly created storage container
- Move to Snowflake and create a database for the project
- Create a storage integration between Azure and Snowflake by using the appropriate AZURE_TENANT_ID, storage account name and storage container name
- For loading the data from Azure into Snowflake, create a stage using the newly created storage integration
- Insert the data into external tables on Snowflake so that metadata about the files, including the filename, can be accessed
- Retrieve the region/country codes from the .csv and .json filenames by making use of the external tables so that the data can be analysed by country
- Transfer the data, along with the region codes, from the external tables to (internal) tables on Snowflake so that the data can be transformed and analysed using Data Manipulation Language (DML). The two tables should be called 'table_youtube_trending' and 'table_youtube_category'

Data Preparation

The two tables: 'table_youtube_trending' and 'table_youtube_category' were joined on country and categoryid using left outer join so that none of the records were deleted from the 'table_youtube_trending' table. This combined table was named as 'table_youtube_final.'

The final table then went through some cleaning steps:

1. Duplicates for category_title was found out by finding those category titles that had multiple rows for each country in the 'table_youtube_category' table. 'Comedy' was found to be the category that had duplicates in the table.
2. The category title: 'Nonprofits & Activism' appeared only for one country in the 'table_youtube_category' table. This indicated that there were probably some missing values in the category_title column.
3. '29' was the category_id for the missing category_title values. Interestingly, the category_id, 29, corresponds to the 'Nonprofits & Activism' category_title.
4. The missing category_title values were then replaced with 'Nonprofits & Activism' as their category_id was 29. This was done with the help of a subquery. At first, the category_title corresponding to category_id of 29 was retrieved, which was 'Nonprofits & Activism.' Then the retrieved category_title was put in the SET clause, to replace the missing values.
5. The video with no channel title was discovered to be: '9b9MovPPewk.'
6. The video_id column had some garbage values with the value: '#NAME?'. These were dropped from the 'table_youtube_final' table.
7. The 'table_youtube_final' table contained a lot of duplicates with same video_id, country, and trending_date but they had different likes, dislikes, etc. So, it was decided to drop those

duplicates that had less view_count, i.e., the duplicate with the highest view_count would be retained.
To achieve this, the videos for each video_id, country, and trending_date were ranked in descending order using the row_number function. The duplicates that did not have the highest rank were then loaded into a table called: 'table_youtube_duplicates.' These were the bad duplicates that need to be dropped from the 'table_youtube_final' table.

8. The bad duplicates from the 'table_youtube_final' were deleted by matching the unique id of the 'table_youtube_final' with the unique id of 'table_youtube_duplicates.' This ensured that all the bad duplicates were deleted from 'table_youtube_final' table.
9. The final row count for 'table_youtube_final' table was 1,123,017.

Data Analysis

1. The 3 most viewed videos for each country in the 'Sports' category and trending_date: '2021-10-17' was determined in two steps. Firstly, the videos for each country in the 'Sports' category and trending_date: '2021-10-17' were ranked using the row_number function and stored temporarily using the WITH clause. Then, the top 3 ranked videos for each country were retrieved from the temporarily stored result.

A	B	C	D	E
COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
BR	BRASIL 4 X 1 URUGUAI | MELHORES MOMENTOS | 12¬™ RODADA ELI	ge	4562725	1
BR	MAIS TRväS GOLS DE CRISTIANO RONALDO! PORTUGAL 5 X 0 LUXEMI	TNT Sports Brasil	2053005	2
BR	‚ô´ NEYMAR TvÅ DE VOLTA!! E A DUPLA COM RAPHINHA DECOLOU! |	FutParv≥dias	814491	3
CA	Sore loser! An idiot! Tyson Fury reveals what was said between him &	BT Sport Boxing	6913800	1
CA	World's Smallest TV | OT 30	Dude Perfect	6222811	2
CA	Eliminatorias | Brasil 4-1 Uruguay | Fecha 12	CONMEBOL	4354963	3
DE	Eliminatorias | Brasil 4-1 Uruguay | Fecha 12	CONMEBOL	4354963	1
DE	Lesnar returns for the Universal Title Match Contract Signing with Reig	WWE	2872431	2
DE	Timo Werner schievüt DFB-Team zur WM: Nordmazedonien - Deutsch	DAZN LV§nderspiele	1793189	3
FR	Lesnar returns for the Universal Title Match Contract Signing with Reig	WWE	2872431	1
FR	Le film de la finale de l'UEFA Nations League, Equipe de France I FFF 2	Fv©dv©ration Franvßa	1504302	2
FR	Espagne 1-2 France, le reÃÅsumeÃÅ - Finale UEFA Nations League I FF	Fv©dv©ration Franvßa	1454288	3

2. The count of the number of distinct videos containing the word 'BTS' for each country was found out with the help of contains function. The result was then ordered in descending order.

A	B
COUNTRY	CT
KR	331
RU	230
US	179
CA	173
MX	164

3. The most viewed video and its likes_ratio for each country and year_month was found out in several steps. Firstly, the year_month column was retrieved from trending_date with the help of date_from_parts and date_part functions. Secondly, the video ranks for each country and

year_month were temporarily stored with the help of WITH clause. Then, the most viewed video for each country and year_month along with its likes_ratio was retrieved from the temporarily stored result.

A	B	C	D	E	F	G
COUNTRY	YEAR_MONT	TITLE	CHANNELTIT	CATEGORY_	VIEW_COUN	LIKES_RATIO
BR	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	244507902	6.52
CA	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	232649205	6.76
DE	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	219110491	7.06
FR	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	232649205	6.76
GB	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	208581468	7.31
IN	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	253995993	6.34
JP	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	262319276	6.2
KR	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	262319276	6.2
MX	1/8/2020	BTS (Î∞©ÌÉ	Big Hit Label	Music	253995993	6.34

4. The category_title having the most distinct videos for each country and its percentage out of the total distinct videos of that country was found out in 3 steps. Firstly, the category_title having the most distinct videos for each country was determined and loaded into an intermediary table called t1. Secondly, the total distinct videos for each country were determined and loaded into another intermediary table called t2. Then, these two tables were joined on country to get the final result.

A	B	C	D	E
COUNTRY	CATEGORY_TITLE	TOTAL_CATE	TOTAL_COUI	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	22.04
IN	Entertainment	12839	29431	43.62
JP	Entertainment	4945	14816	33.38
KR	Entertainment	4625	13457	34.37
MX	Entertainment	3628	15347	23.64

5. The channeltitle with the most distinct videos was determined by ranking the video count for each channel and then finding the channeltitle with the highest video count.

A	B
CHANNELTITLE	VIDEO_COUNT
Colors TV	805

Launching a new YouTube channel

To create top trending videos for the new YouTube channel, the data needs to be analysed to find which category of videos appear the most in the trending list.

The top 4 category of videos that appear the most in the trending data are: 'Music', 'Entertainment', 'People & Blogs', and 'Gaming.'

A	B
CATEGORY_TITLE	COUNT
Entertainment	294506
Music	180233
People & Blogs	134130
Gaming	122123
Sports	112916
Comedy	67079

The 'Music' and 'Entertainment' categories are not considered as they are way too common. So, both 'People & Blogs' and 'Gaming' seem like good candidates.

The top trending list on YouTube is determined by measuring the user's interactions including number of views, likes, dislikes, etc. So, the data is then analysed to find the categories with most views, likes and comments.

A	B	C	D	E
CATEGORY_TITLE	TOTAL_VIEWS	TOTAL_LIKES	TOTAL_DISLIKES	TOTAL_COMMENTS
Music	7.399E+11	52268363406	929227566	6320480751
Entertainment	5.19487E+11	28497215536	664670847	1678100432
Gaming	1.95818E+11	11542284329	198036479	793960547
People & Blogs	1.87418E+11	11133135822	262353145	607093594
Sports	1.33652E+11	3618342269	94999368	280758545
Comedy	1.08493E+11	7406938061	139239196	328883606
Science & Technology	61917522815	2810121892	71176606	168100428
Film & Animation	52847803701	2259479759	50014528	151725049
Howto & Style	34743662369	1648085570	46339391	92650234

Here, 'Gaming' and 'People & Blogs' have the 3rd and 4th most viewed videos in the trending data. So, they both still seem very good candidates.

The top trending list on YouTube also depends on which videos have the wider reach of viewers. So, the data is then analysed to find the 3rd ranked category_title for each country (the top 2 ranked category_title for each country is 'Music' and 'Entertainment').

A	B	C	D	E	F	G
COUNTRY	CATEGORY_TITLE	TOTAL_VIEWS	TOTAL_LIKES	TOTAL_DISLIKES	TOTAL_COMMENTS	RK
DE	People & Blogs	19436702702	982747391	29713741	46889358	3
JP	People & Blogs	9926658604	356374080	6587957	23040036	3
US	Gaming	38792528568	2133688005	36186120	168756436	3
GB	Gaming	31878388652	1817568703	30440427	135128343	3
KR	People & Blogs	13515625109	510986481	10345529	46806347	3
MX	Gaming	22537854653	1701393011	27652395	95092343	3
IN	People & Blogs	29447709750	1701580352	40224435	72419243	3
RU	People & Blogs	8012856672	614397280	18966570	50941699	3
BR	People & Blogs	13300888918	1219498587	22075278	45090514	3
CA	Gaming	36677437270	2054166219	33989411	153911978	3
FR	Sports	7866547047	200338481	5126913	16132182	3

Here, Gaming is the 3rd most viewed category for 4 countries, whereas 'People & Blogs' is the 3rd most viewed category for 6 countries. So, 'People & Blogs' seems to have a wider reach of viewers and therefore, the videos for the new YouTube channel should be created under the 'People & Blogs' category.

Even though the 'People & Blogs' category of videos seem to have a wider reach of viewers it is still unlikely that they would trend in every country. It is unlikely that they would trend in US, Great Britain, Canada, and France.

References

1. https://support.google.com/youtube/answer/7239739?hl=en
2. https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset?select=BR_youtube_trending_data.csv