Data Lakehouse with Snowflake

YouTube Trending Data Analysis

Big Data Engineering

TABLE OF CONTENTS

1. Introduction

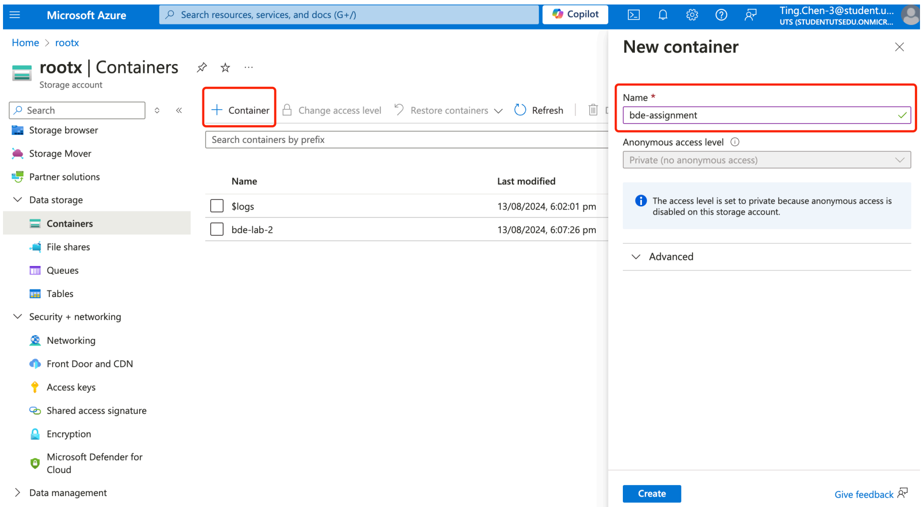
Digitalization has reshaped the entertainment industry, with YouTube leading the way as a video sharing platform for content creators to share creativity, while monetizing their work. The goal of this project is to build a data lakehouse, based on YouTube's trending data using Azure for cloud storage and Snowflake for data processing and data analysis. Additionally, address a key business question via data lakehouse regarding the best channel category to launch, along with relevant recommendations. The dataset includes daily trending YouTube videos from 2020-08-12 to 2024-04-15, across 10 major countries.

2. Environment Setup

The project starts with environment setup. The data was stored in Azure's container and it was connected to Snowflake via a Shared Access Signature (SAS) token. This approach allows us to save physical data in an Azure container and access it directly from Snowflake. Below are the steps to set up on both Azure and Snowflake.

2.1 Azure

1. Downloaded and unzipped the required data files in computer: CSV files for YouTube trending data and Json files for YouTube category data.
2. Created a new container called "bde-assignment" on Azure.



Name	Last modified
$logs	13/08/2024, 6:02:01 pm
bde-lab-2	13/08/2024, 6:07:26 pm

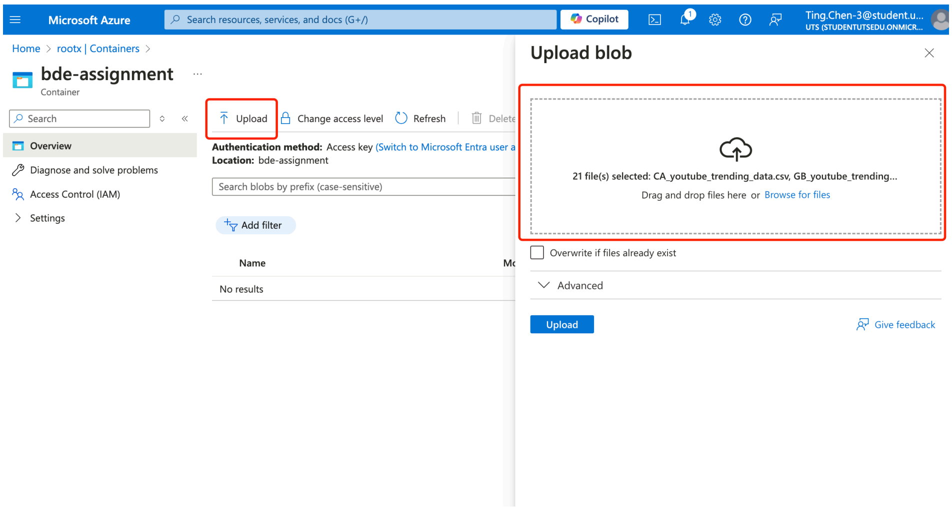
Name *: bde-assignment

Anonymous access level: Private (no anonymous access)

The access level is set to private because anonymous access is disabled on this storage account.

Fig 1. Create Azure container

3. Uploaded the downloaded datasets into "bde-assignment1" container.



Authentication method: Access key (Switch to Microsoft Entra user a

Location: bde-assignment

Search blobs by prefix (case-sensitive)

No results

21 file(s) selected: CA_youtube_trending_data.csv, GB_youtube_trending...

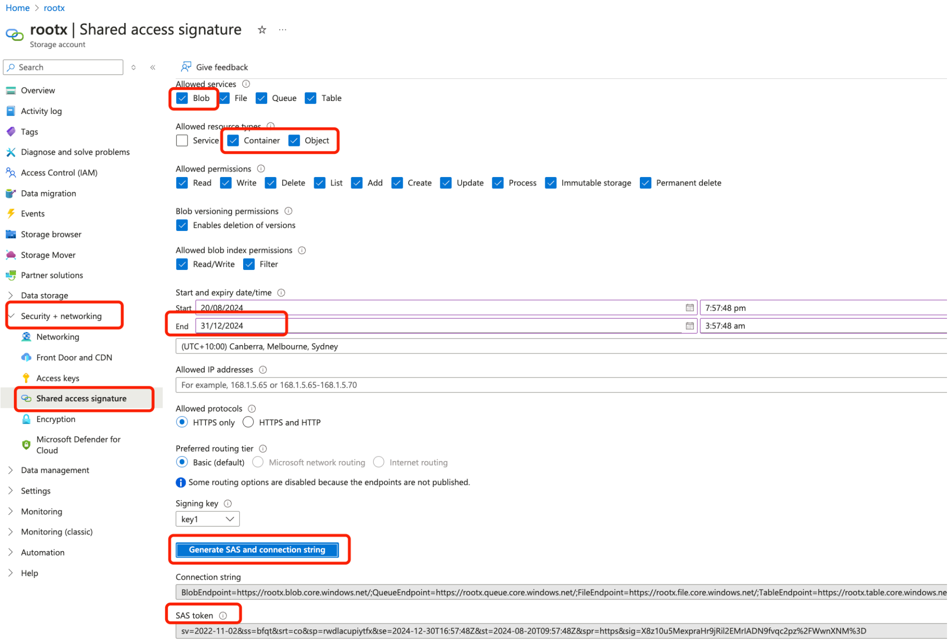
Drag and drop files here or Browse for files

□ Overwrite if files already exist

Advanced

Fig 2. Upload files into Azure container

4. Generate SAS token of the container by setting Shared access signature under Security + networking, allowing Blob services and Container and Object resource types. Set a appropriate expiry date.



Allowed services
☑ Blob ☑ File ☑ Queue ☑ Table

Allowed resource types
□ Service ☑ Container ☑ Object

Allowed permissions
☑ Read ☑ Write ☑ Delete ☑ List ☑ Add ☑ Create ☑ Update ☑ Process ☑ Immutable storage ☑ Permanent delete

Blob versioning permissions
☑ Enables deletion of versions

Allowed blob index permissions
☑ Read/Write ☑ Filter

Start and expiry date/time
Start 20/08/2024 7:57:48 pm
End 31/12/2024 3:57:48 am
(UTC+10:00) Canberra, Melbourne, Sydney

Allowed IP addresses
For example, 168.1.5.65 or 168.1.5.65-168.1.5.70

Allowed protocols
● HTTPS only ○ HTTPS and HTTP

Preferred routing tier
● Basic (default) ○ Microsoft network routing ○ Internet routing

Some routing options are disabled because the endpoints are not published.

Signing key
key1

Connection string
BlobEndpoint=https://rootx.blob.core.windows.net/;QueueEndpoint=https://rootx.queue.core.windows.net/;FileEndpoint=https://rootx.file.core.windows.net/;TableEndpoint=https://rootx.table.core.windows.ne

SAS token
sv=2022-11-02&ss=bfqt&srt=co&sp=rwdlacupiytfx&se=2024-12-30T16:57:48Z&st=2024-08-20T09:57:48Z&spr=https&sig=X8z10u5MexpraHr9jRil2EMrIADN9fvqc2pz%2FWwnXNM%3D

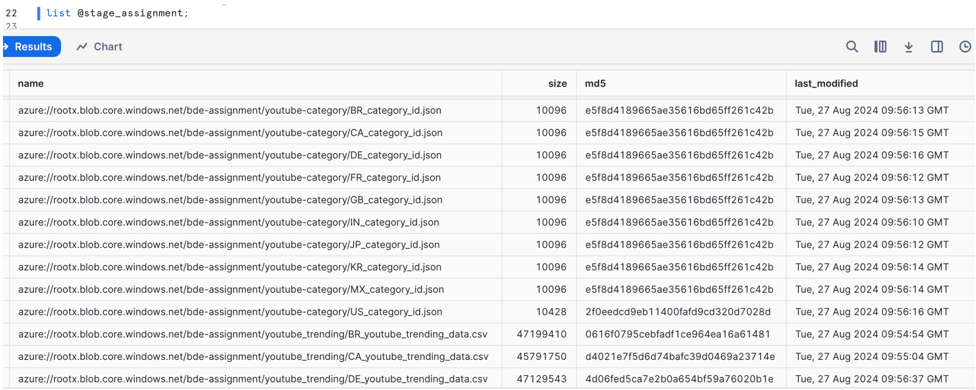
Fig 3. Generate SAS token from Azure container

2.2 Snowflake

Corresponding code file: part_1.sql (Q1 - Q3)

1. Created a database “assignment_1” in a worksheet on Snowflake, and switched to the database created.

2. Created a stage “stage_assignment”, connecting to the Azure storage, with URL='azure:// <*storage_account_name*>.blob.core.windows.net/bde-assignment' and CREDENTIALS=(AZURE_SAS_TOKEN= '?<*SAS token*>').
3. Listed all the files inside the stage_assignment to check if the stage between Azure and Snowflake is set up successfully.



```
list @stage_assignment;
```

name	size	md5	last_modified
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/BR_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:13 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/CA_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:15 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/DE_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:16 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/FR_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:12 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/GB_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:13 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/IN_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:10 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/JP_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:12 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/KR_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:14 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/MX_category_id.json	10096	e5f8d4189665ae35616bd65ff261c42b	Tue, 27 Aug 2024 09:56:14 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube-category/US_category_id.json	10428	2f0eedcd9eb11400fafd9cd320d7028d	Tue, 27 Aug 2024 09:56:16 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube_trending/BR_youtube_trending_data.csv	47199410	0616f0795cebfadf1ce964ea16a61481	Tue, 27 Aug 2024 09:54:54 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube_trending/CA_youtube_trending_data.csv	45791750	d4021e7f5d6d74bafc39d0469a23714e	Tue, 27 Aug 2024 09:55:04 GMT
azure://rootx.blob.core.windows.net/bde-assignment/youtube_trending/DE_youtube_trending_data.csv	47129543	4d06fed5ca7e2b0a654bf59a76020b1e	Tue, 27 Aug 2024 09:56:37 GMT

Fig 4. File lists on stage_assignment

3. *Data Preparation*

After setting up the environment on Azure and Snowflake, the data stored in Azure’s container was ingested into Snowflake. Two external tables were created to store the data in its original form. Then necessary transformations was performed to get tabular tables in Snowflake. Once finishing the ingestion, data cleaning process was conducted, which involved identifying problematic records and implementing cleaning procedures to ensure data consistency and accuracy.

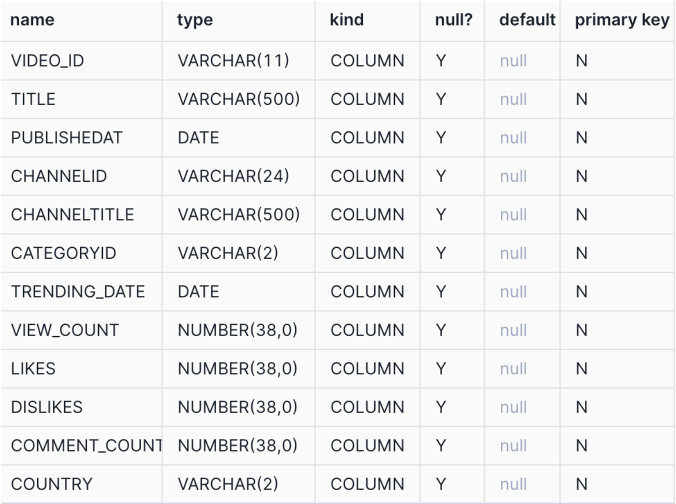
3.1 Data Ingestion

Corresponding code file: part_1.sql (Q4 – Q6)

1. Created two external tables to keep the original form of the datasets.
   a. Created file format file_format_csv to skip header, before parsing the CSV files
   b. Created external table ex_table_youtube_trending with any files ending with ‘.csv’ on the stage, specifying the file format to file_format_csv.
   c. Created external table ex_table_youtube_category with any files ending with ‘.json’ on the stage.

2. Transferred the data from external tables into tables with appropriate schemas.

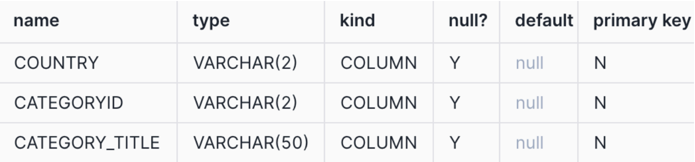
   a. Created table_youtube_trending with proper data types, and pull country from CSV file name by calling Metadata$Filename in split_part function. Part of the schema shown on Fig5:



name	type	kind	null?	default	primary key
VIDEO_ID	VARCHAR(11)	COLUMN	Y	null	N
TITLE	VARCHAR(500)	COLUMN	Y	null	N
PUBLISHEDAT	DATE	COLUMN	Y	null	N
CHANNELID	VARCHAR(24)	COLUMN	Y	null	N
CHANNELTITLE	VARCHAR(500)	COLUMN	Y	null	N
CATEGORYID	VARCHAR(2)	COLUMN	Y	null	N
TRENDING_DATE	DATE	COLUMN	Y	null	N
VIEW_COUNT	NUMBER(38,0)	COLUMN	Y	null	N
LIKES	NUMBER(38,0)	COLUMN	Y	null	N
DISLIKES	NUMBER(38,0)	COLUMN	Y	null	N
COMMENT_COUNT	NUMBER(38,0)	COLUMN	Y	null	N
COUNTRY	VARCHAR(2)	COLUMN	Y	null	N

Fig 5. Schema of table_youtube_trending

   b. Created table_youtube_category; pull country from Json file name by calling Metadata$Filename in split_part function; Flatten Json array with Lateral Flatten, and retrieve fields of interest, assign them proper data types. Part of the schema shown on Fig6:



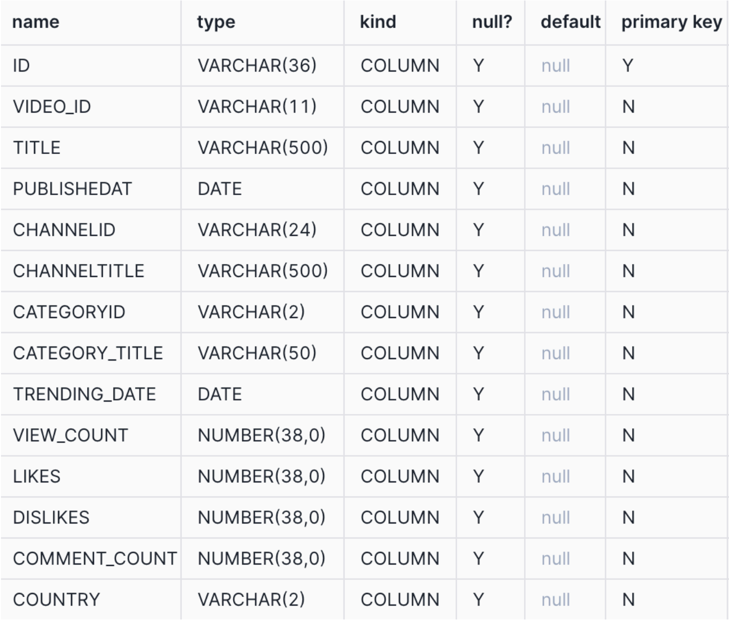
name	type	kind	null?	default	primary key
COUNTRY	VARCHAR(2)	COLUMN	Y	null	N
CATEGORYID	VARCHAR(2)	COLUMN	Y	null	N
CATEGORY_TITLE	VARCHAR(50)	COLUMN	Y	null	N

Fig 6. Schema of table_youtube_category

3. Merged tables into a final table for data analysis:

    a. Created a table called "table_youtube_final", using join to combine all columns from tables created previously on the fields of country and categoryid.

    b. Generated a new field id by using the "UUID_STRING()" function and set it as primary key. Part of the schema shown on Fig7:



name	type	kind	null?	default	primary key
ID	VARCHAR(36)	COLUMN	Y	null	Y
VIDEO_ID	VARCHAR(11)	COLUMN	Y	null	N
TITLE	VARCHAR(500)	COLUMN	Y	null	N
PUBLISHEDAT	DATE	COLUMN	Y	null	N
CHANNELID	VARCHAR(24)	COLUMN	Y	null	N
CHANNELTITLE	VARCHAR(500)	COLUMN	Y	null	N
CATEGORYID	VARCHAR(2)	COLUMN	Y	null	N
CATEGORY_TITLE	VARCHAR(50)	COLUMN	Y	null	N
TRENDING_DATE	DATE	COLUMN	Y	null	N
VIEW_COUNT	NUMBER(38,0)	COLUMN	Y	null	N
LIKES	NUMBER(38,0)	COLUMN	Y	null	N
DISLIKES	NUMBER(38,0)	COLUMN	Y	null	N
COMMENT_COUNT	NUMBER(38,0)	COLUMN	Y	null	N
COUNTRY	VARCHAR(2)	COLUMN	Y	null	N

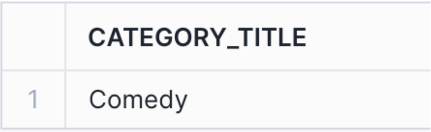
Fig 7. Schema of table_youtube_final

3.2 Data Cleaning

Corresponding code file: part_2.sql

Once the final table is prepared, it is essential to check for invalid, duplicated, or missing records of the dataset before starting data analysis and answering business questions. Below are the potential issues found and cleaning procedures implemented:

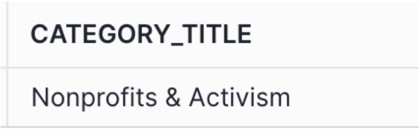
1. Found category_title “Comedy” has multiple categoryid for each country in table_youtube_category. For clarification, category_title instead of categoryid might be a better choice for data analysis.



	CATEGORY_TITLE
1	Comedy

Fig 8. Duplicated category_title

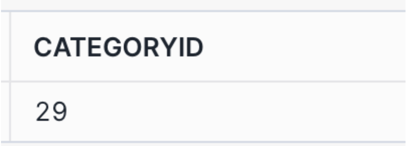
2. Found that category_title “Nonprofits & Activism” only appears in the US. There may be missing values in this categorytitle. Further check for missing values in category_title was required.



CATEGORY_TITLE
Nonprofits & Activism

Fig 9. Category_title only appears in one country

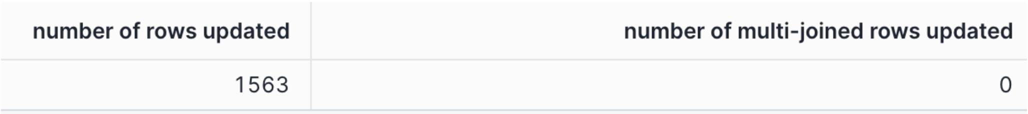
3. Found missing values in category_title, where categoryid is 29.



CATEGORYID
29

Fig 10. Categoryid of the missing category_titles

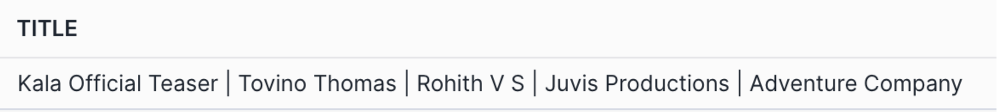
4. Fixed category_title in table_youtube_final by filling the null values with “Nonprofits & Activism”, using update and where clause.



number of rows updated	number of multi-joined rows updated
1563	0

Fig 11. Update the null values in category_title

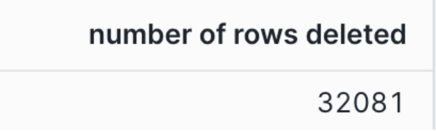
5. Found that a video record has a null channeltitle in table_youtube_final. There might be some duplicates to be handled.



TITLE
Kala Official Teaser | Tovino Thomas | Rohith V S | Juvis Productions | Adventure Company

Fig 12. Video without channeltitile

6. Removed invalid video records in table_youtube_final, where the video_id is '#NAME?'. Deleted 32081 rows.

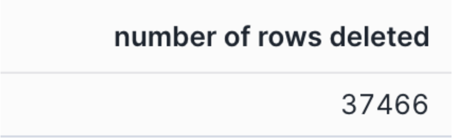


number of rows deleted
32081

Fig 13. Remove invalid data

7. Kept track of duplicates in final table: any record with the same video_id, country and trending_date were considered as duplicate, and the video record with the highest view_count was kept. Other records would be saved into the newly created table table_youtube_duplicates.

8. Handled duplicates in final table: deleted duplicated video records in table_youtube_final, by specifying where the video_id was in table_youtube_duplicates. 37466 duplicates removed.



number of rows deleted
37466

Fig 14. Remove duplicates

9. Checked the rows of table_youtube_final: 2597494.

RECORDS
2597494

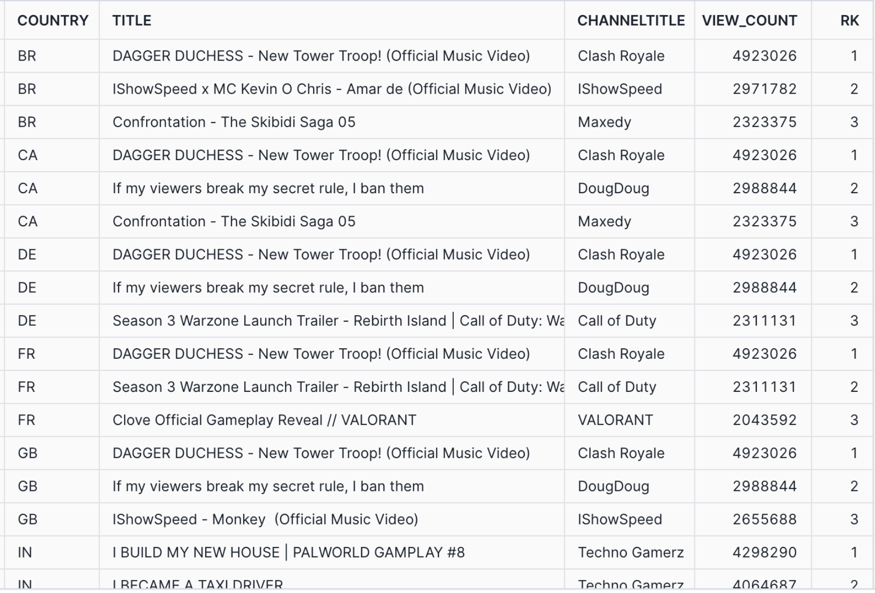
Fig 15. Number of rows of table_youtube_final

4. Data Analysis

Corresponding code file: part_3.sql

In this section, query questions was addressed in order to gain a comprehensive understanding of the dataset. These questions focus on metrics such as most viewed video, number of videos with specific title, and dominant category regarding different channels, categories and countries. Here are the detailed queries:

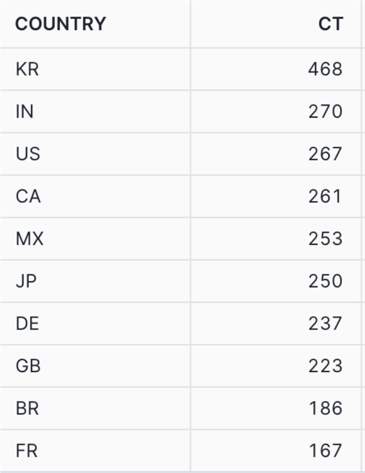
1. Queried the most viewed trending videos in each country in the category of Gaming and in the trending date of '2024-04-01'. Noticed that there was a video appeared in 7 countries in the same day.



COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
BR	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
BR	IShowSpeed x MC Kevin O Chris - Amar de (Official Music Video)	IShowSpeed	2971782	2
BR	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
CA	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
CA	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
CA	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
DE	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
DE	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
DE	Season 3 Warzone Launch Trailer - Rebirth Island | Call of Duty: Wa	Call of Duty	2311131	3
FR	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
FR	Season 3 Warzone Launch Trailer - Rebirth Island | Call of Duty: Wa	Call of Duty	2311131	2
FR	Clove Official Gameplay Reveal // VALORANT	VALORANT	2043592	3
GB	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
GB	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
GB	IShowSpeed - Monkey (Official Music Video)	IShowSpeed	2655688	3
IN	I BUILD MY NEW HOUSE | PALWORLD GAMPLAY #8	Techno Gamerz	4298290	1
IN	I BECAME A TAXI DRIVER	Techno Gamerz	4064687	2

Fig 15. Result of most 3 viewed trending videos

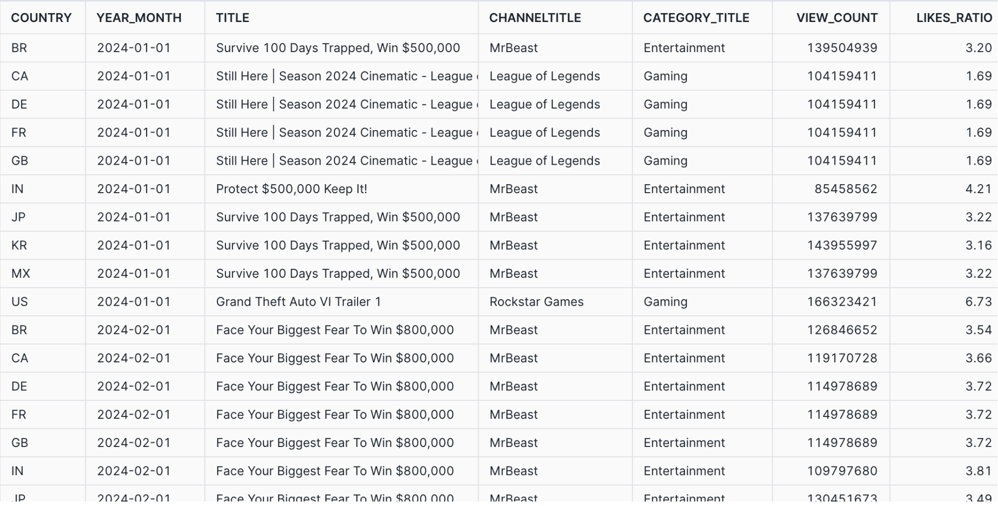
2. Checked that if K-pop elements such as “BTS” is popular in some specific countries. It turned out that Korea (KR) has the most trending video regarding ‘BTS’.



COUNTRY	CT
KR	468
IN	270
US	267
CA	261
MX	253
JP	250
DE	237
GB	223
BR	186
FR	167

Fig 16. Result of the count of ‘BTS’ trending videos

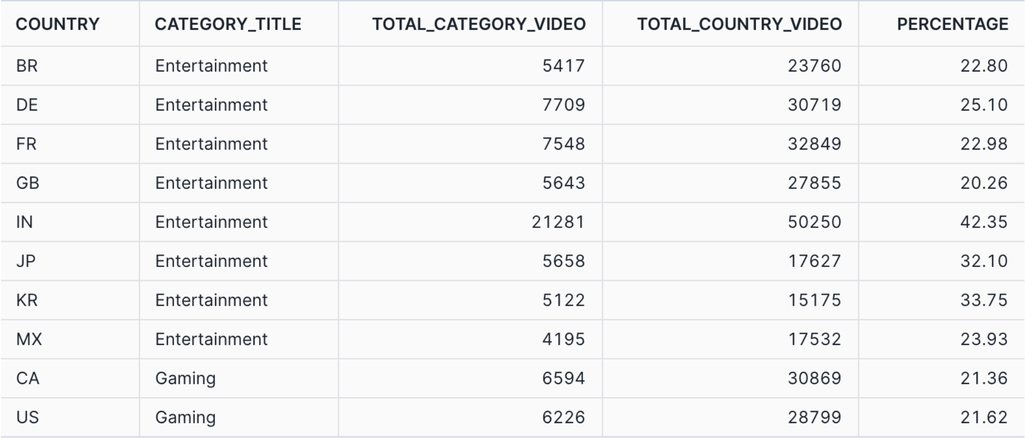
3. Queried the most viewed video and its corresponding like_ratio (likes / view_count) in each month of 2024 in each country. Notice that entertainment is dominating in nearly every country, and Gaming also has a broad coverage of viewers.



COUNTRY	YEAR_MONTH	TITLE	CHANNELTITLE	CATEGORY_TITLE	VIEW_COUNT	LIKES_RATIO
BR	2024-01-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	139504939	3.20
CA	2024-01-01	Still Here | Season 2024 Cinematic - League	League of Legends	Gaming	104159411	1.69
DE	2024-01-01	Still Here | Season 2024 Cinematic - League	League of Legends	Gaming	104159411	1.69
FR	2024-01-01	Still Here | Season 2024 Cinematic - League	League of Legends	Gaming	104159411	1.69
GB	2024-01-01	Still Here | Season 2024 Cinematic - League	League of Legends	Gaming	104159411	1.69
IN	2024-01-01	Protect $500,000 Keep It!	MrBeast	Entertainment	85458562	4.21
JP	2024-01-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	137639799	3.22
KR	2024-01-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	143955997	3.16
MX	2024-01-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	137639799	3.22
US	2024-01-01	Grand Theft Auto VI Trailer 1	Rockstar Games	Gaming	166323421	6.73
BR	2024-02-01	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	126846652	3.54
CA	2024-02-01	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	119170728	3.66
DE	2024-02-01	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	114978689	3.72
FR	2024-02-01	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	114978689	3.72
GB	2024-02-01	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	114978689	3.72
IN	2024-02-01	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	109797680	3.81
JP	2024-02-01	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	130451673	3.49

Fig 17. Result of the most viewed videos in 2024 per month

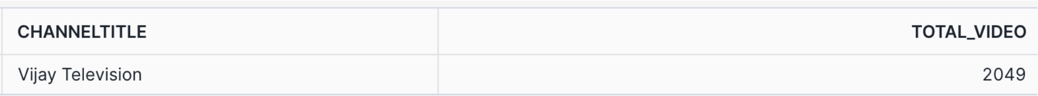
4. Retrieved category_title with the most distinct trending videos for each country and the proportion of the country.



COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
BR	Entertainment	5417	23760	22.80
DE	Entertainment	7709	30719	25.10
FR	Entertainment	7548	32849	22.98
GB	Entertainment	5643	27855	20.26
IN	Entertainment	21281	50250	42.35
JP	Entertainment	5658	17627	32.10
KR	Entertainment	5122	15175	33.75
MX	Entertainment	4195	17532	23.93
CA	Gaming	6594	30869	21.36
US	Gaming	6226	28799	21.62

Fig 18. Result of the category_title with the most distinct trending videos in each country

5. Checked channeltitle with the most distinct videos. "Vijay Television" has the most distinct trending videos with 2049 of them.



CHANNELTITLE	TOTAL_VIDEO
Vijay Television	2049

Fig 19. Result of the channeltitile with the most distinct trending videos

5. *Business Question*

Corresponding code file: part_4.sql

In this section, a business question was set to be addressed: Which channel category (excluding "Music" and "Entertainment") of video may be the best choice when considering to start a new channel in order to have it appeared on YouTube top trending videos.

5.1 Get Data Ready

To make a more representative decision, two rules were set before retrieving data:

1. Derived business insights from data starting in 2023, considering the fast-changing nature of trends, and to get a dataset that is closer to real life as the post-pandemic era.
2. Considered only the first instance of a video appearing in the trending list. This means each trending video was counted only once, on the first trending day it appeared YouTube top trending videos for the purpose of data simplicity and consistency.

5.2 Methodology

The purpose of this part is to transform business question into mathematical question. Assumed that there are three aspects needed to be considered when deciding a channel category to launch: audience Interest and reach, channel competitiveness within category, and audience engagement. Here are the hypothesis and the key metrics for answering the business question:

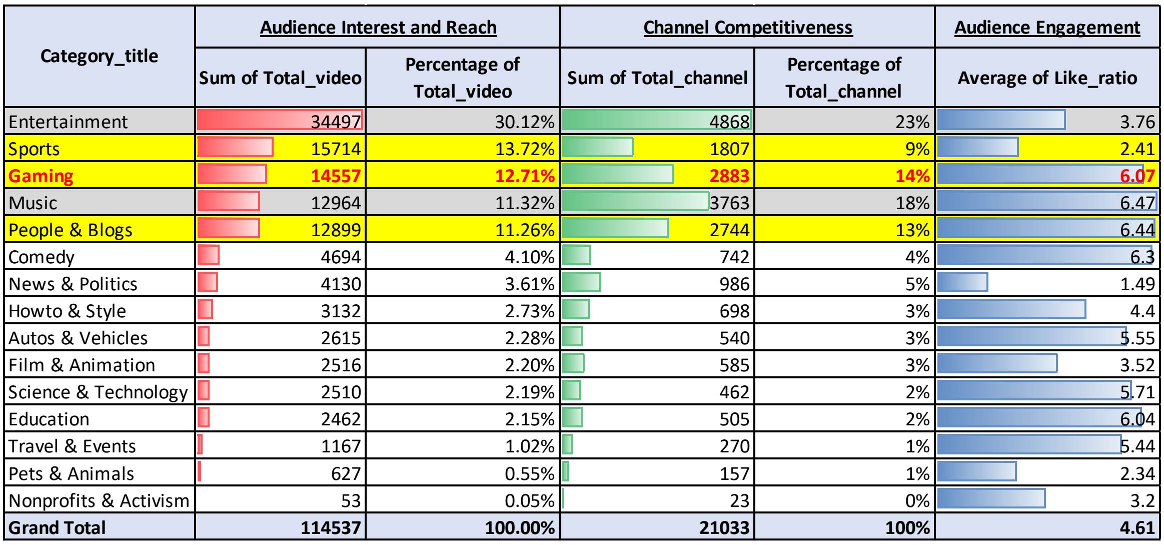
1. Hypothesis: categories with a high audience interest and reach, high engagement and low competitiveness are more likely to get videos on the trending list.
2. Key metrics to be considered:
   a. Audience Interest and Reach: Total number of distinct trending videos within the category. It is considered as the most fundamental metric out of three.
   b. Channel Competitiveness: Total number of distinct channels within the category.
   c. Audience Engagement: Likes ratio (the percentage of likes against view_count) within the category.

5.3 Data Exploration

1. Overall performance

Data was retrieved under the requirements outlined above, and key metrics were calculated.

Fig 20 provides the result for each category:



Category_title	Audience Interest and Reach		Channel Competitiveness		Audience Engagement
	Sum of Total_video	Percentage of Total_video	Sum of Total_channel	Percentage of Total_channel	Average of Like_ratio
Entertainment	34497	30.12%	4868	23%	3.76
Sports	15714	13.72%	1807	9%	2.41
Gaming	**14557**	**12.71%**	**2883**	**14%**	**6.07**
Music	12964	11.32%	3763	18%	6.47
People & Blogs	12899	11.26%	2744	13%	6.44
Comedy	4694	4.10%	742	4%	6.3
News & Politics	4130	3.61%	986	5%	1.49
Howto & Style	3132	2.73%	698	3%	4.4
Autos & Vehicles	2615	2.28%	540	3%	5.55
Film & Animation	2516	2.20%	585	3%	3.52
Science & Technology	2510	2.19%	462	2%	5.71
Education	2462	2.15%	505	2%	6.04
Travel & Events	1167	1.02%	270	1%	5.44
Pets & Animals	627	0.55%	157	1%	2.34
Nonprofits & Activism	53	0.05%	23	0%	3.2
Grand Total	**114537**	**100.00%**	**21033**	**100%**	**4.61**

Fig 20. Result of the key metrics for each category

According to Fig 20, ‘Sports’, ‘Gaming’ and ‘People & Blogs’ appeared as the categories of interest, showing the most audience interest and reach apart from ‘Entertainment’ and ‘Music’, with proportions ranging from 11% to 14%.

In terms of channel competitiveness, ‘Gaming’ and ‘People & Blogs’ had the most distinct channels, indicating strong competitiveness, while ‘Sports’ was less competitive.

Regarding audience engagement, ‘Sports’ was far less than the other two and even below average. Fewer than 3 views out of 100 gave likes in this category. However, ‘Gaming’ and ‘People & Blogs’ both received high engagement from viewers, with over 6 out of 100 views liking the video in these categories.

In conclusion, ‘Gaming’, ‘People & Blogs’ and ‘Sports’ are the strongest category candidates for launching a YouTube channel. Comprehensively, ‘Gaming’ stands out as the best option, though ‘Sports’ may be preferable if competitiveness is prioritized.

2. Country differences

The focus of this part is to figure out that if the overall pattern in the previous analysis also apply to different country. Fig 21 – 23 persent the results for each category across different countries.

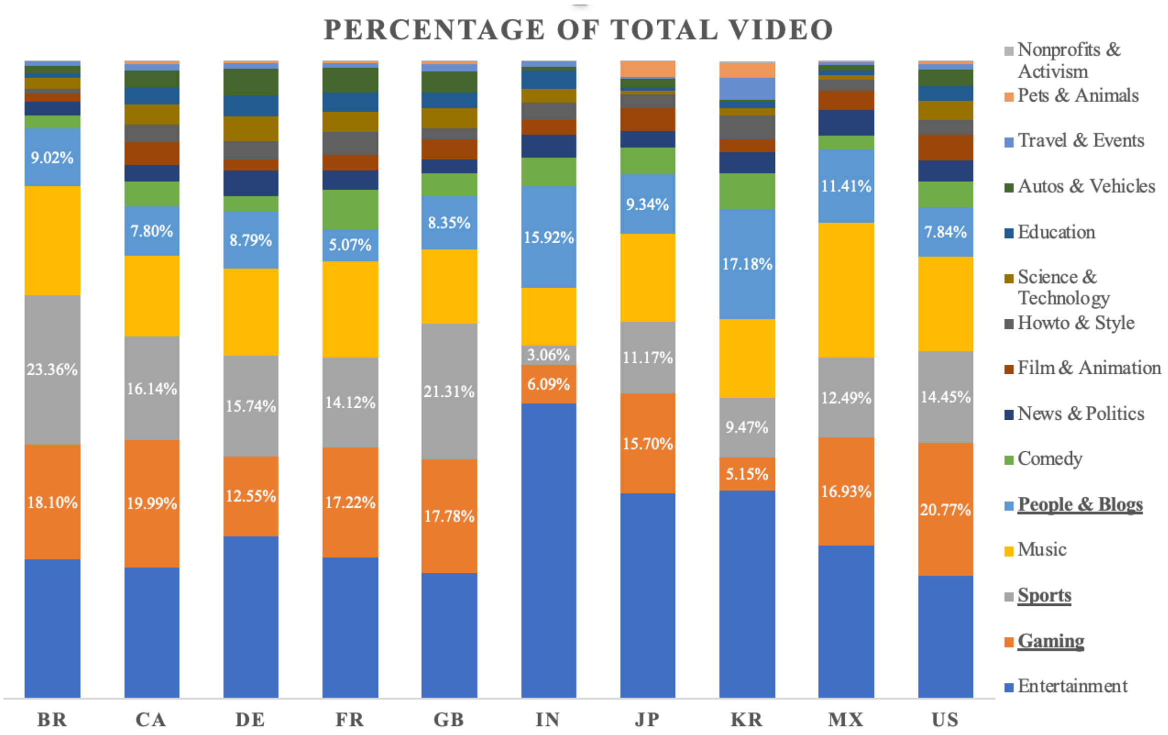


Fig 21. Result of the distinct trending videos

Looking at Fig 21, three countries presented significantly different patterns comparing to the overall one. Firstly, ‘People & Blogs’ in India and Korea had much higher audience interest and reach than the other two categories of interest, with more than 15% respectively. In Great Britain, instead of ‘Entertainment’, ‘Sports’ was the most favoured category, with a proportion of 21%.

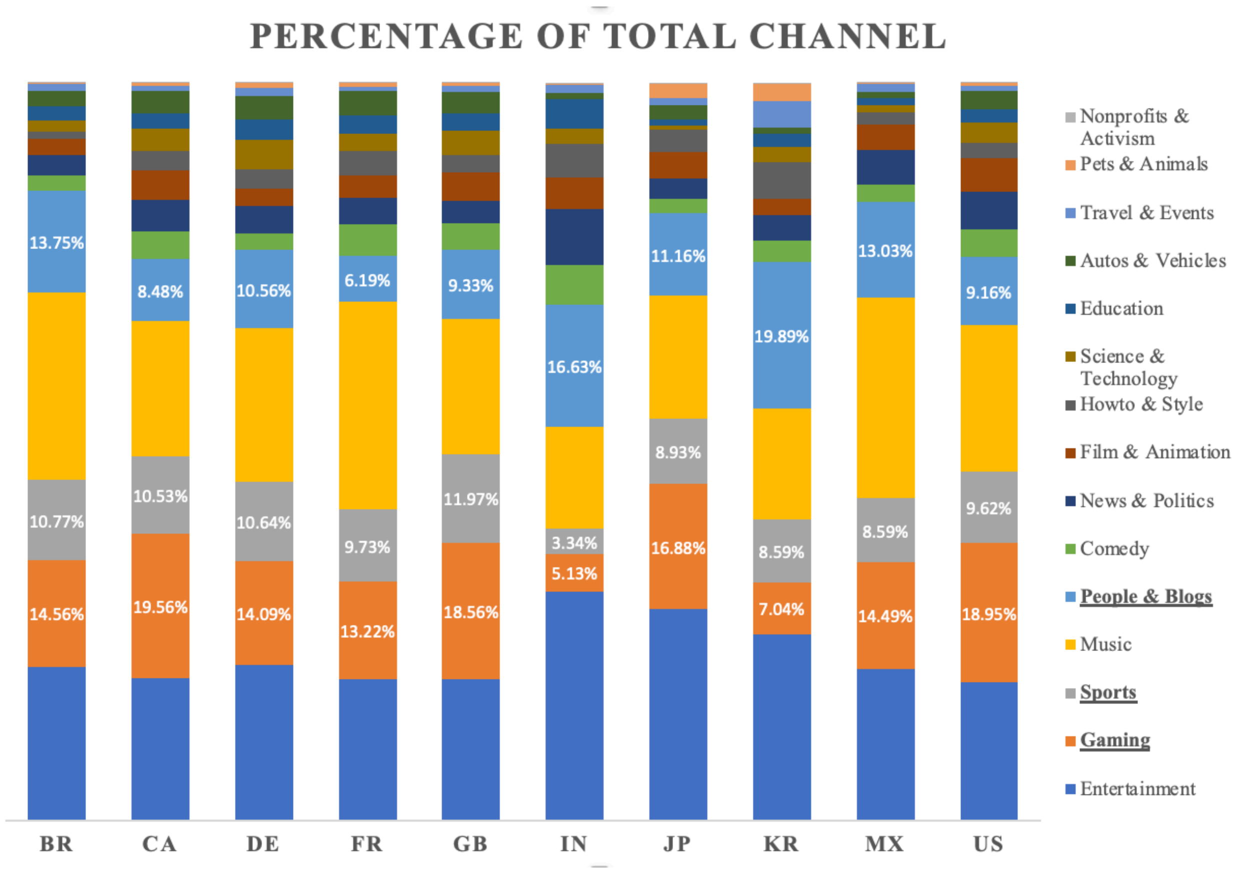


Fig 22. Result of the distinct channel

According to Fig 22, apart from IN, KR and GB, a pattern similar to the overall analysis was observed, with ‘Sports’, being the less competitive category.

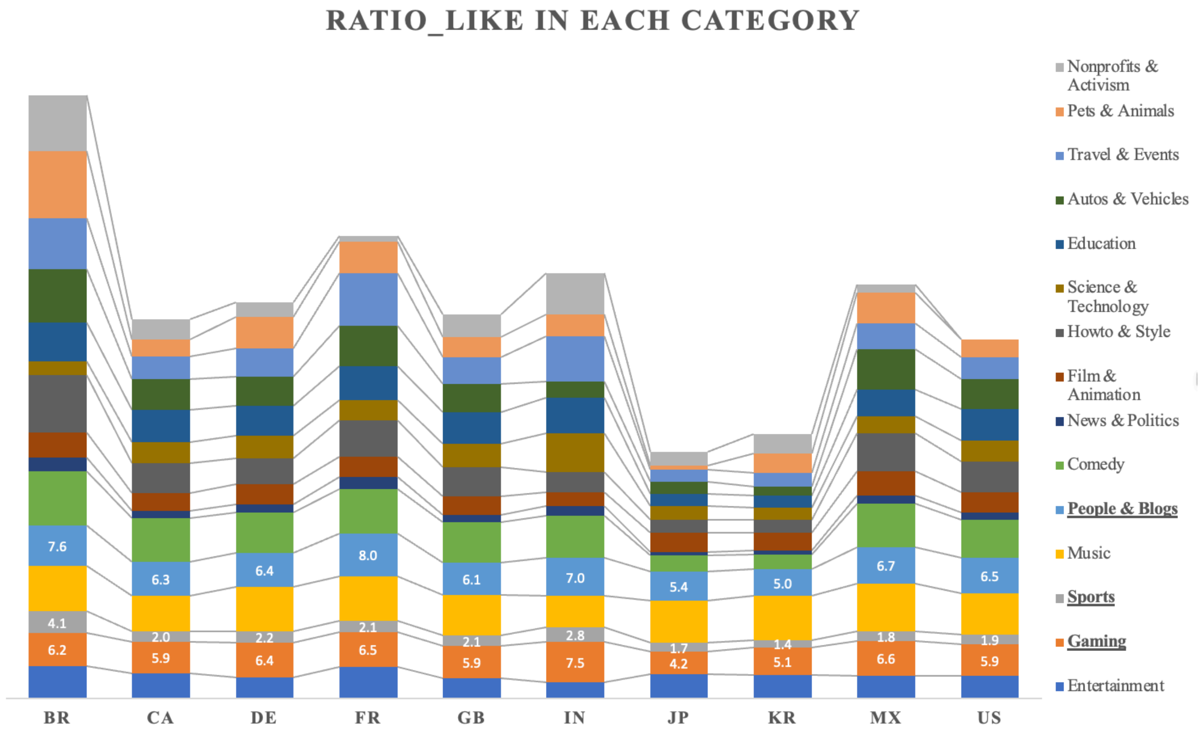


Fig 23. Result of the ratio_like

In terms of audience engagement as shown in Fig 23, the data was similar to the overall pattern, where ‘Gaming’ and ‘People & Blogs’ had fairly high engagement while ‘Sports’ behaving low.

5.4 Recommendation

In general, ‘Gaming’ category is the best category for launching a YouTube channel to have video appeared in trending videos. However, the best option can be varied based on different situations. ‘Sports’ will be a better choice when the business focus is on less competitiveness, and reconsiderations should be taken when the business focus is on countries like India, Korea and Great Britain.

6. Conclusion

This project has illustrated how data can be stored into cloud computing platform Azure and then transferred to the data platform Snowflake in its original form. Data engineers can then perform ETL processes within Snowflake to create well-structure datasets suitable for client-side data analysis. This approach provides an efficient, flexible and scalable for data processing, including data extraction, loading, transfer, cleaning, and transformation. Additionally, it shows strong capabilities in facilitating data analysis and addressing business questions.

7. Appendix

1. Assignment brief and source files used in this project are part of the learning materials from Big Data Engineering of University of Technology Sydney (UTS).
2. Dataset used in this project: Trending data, Category data.
3. Country code dictionary:

Country ID	Country Name
IN	India
US	USA
GB	Great Britain
DE	Germany
CA	Canada
FR	France
BR	Brizil
MX	Mexico
KR	Korea
JP	Japan