Data Lake house with Snowflake

Report

SUBMITTED BY:

HEMANG SHARMA
24695785

Table of Contents

1. Introduction

This report presents the steps and SQL queries used to analyze YouTube trending dataset in Snowflake. The goal was to transform raw data into a structured format, clean it, and perform various analyses to derive insights that can inform business strategies, such as the potential success of launching a YouTube channel in the "Gaming" category.

2. Project Objectives

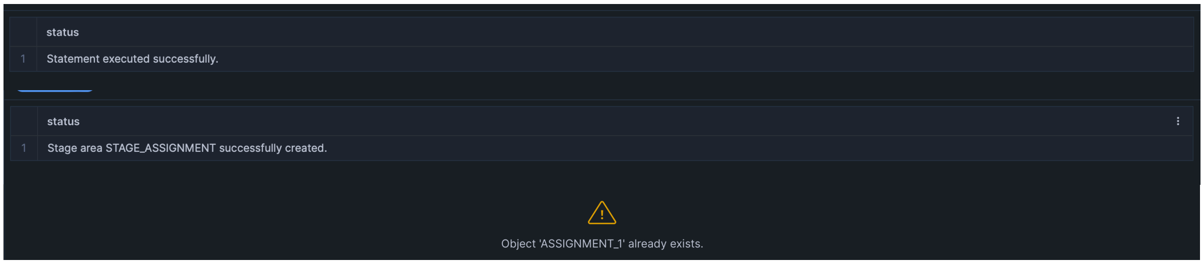
The primary objectives of this project are:

A. **Data Ingestion:** Efficiently load and structure YouTube trending data from external Azure storage into Snowflake.
B. **Data Cleaning:** Identify and rectify data inconsistencies, such as duplicates and missing values, to ensure data integrity.
C. **Data Analysis:** Conduct a series of analyses to extract meaningful insights from the data.
D. **Business Insights:** Provide recommendations based on the analysis to inform strategic decisions for content creation on YouTube.

3. Data Ingestion

Database and Stage Creation

A new Snowflake database, assignment_1, was created to house all project-related tables. An external stage named stage_assignment was set up to connect Snowflake to Azure Blob Storage, where the raw dataset was stored. This stage allowed Snowflake to access and load data from the storage into external tables.

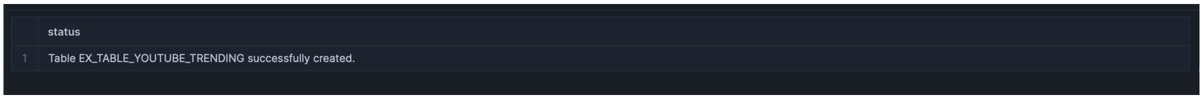


```
   status
1  Statement executed successfully.

   status
1  Stage area STAGE_ASSIGNMENT successfully created.

Object 'ASSIGNMENT_1' already exists.
```

External Table Creation

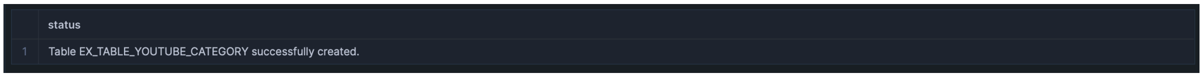
Two external tables were created to hold the YouTube trending data and the category data:

- **ex_table_youtube_trending:** Ingested the CSV files containing YouTube trending video data.



```
   status
1  Table EX_TABLE_YOUTUBE_TRENDING successfully created.
```

- **ex_table_youtube_category:** Ingested the JSON files containing category metadata.

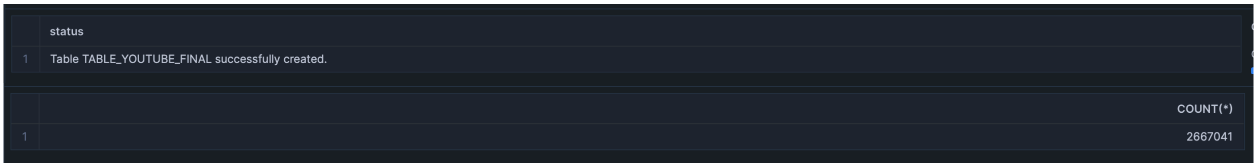


```
   status
1  Table EX_TABLE_YOUTUBE_CATEGORY successfully created.
```

These external tables were transformed into internal tables (table_youtube_trending and table_youtube_category) to facilitate efficient querying and analysis.

Final Table Creation

A final table, table_youtube_final, was created by joining the trending data with the category data on country and category_id. This table included a unique identifier for each record and combined all relevant data fields into a single table.



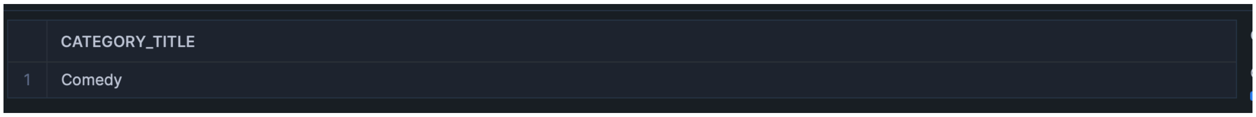
	status
1	Table TABLE_YOUTUBE_FINAL successfully created.

	COUNT(*)
1	2667041

4. Data Validation and Cleaning

Category Consistency Check

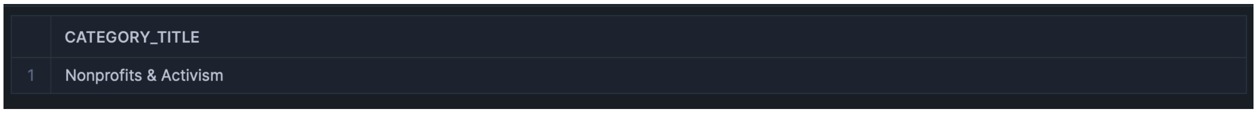
A query was run to identify categories with multiple CATEGORY_IDs, ensuring that each category is uniquely identified across different countries.



	CATEGORY_TITLE
1	Comedy

Single Country Category Check

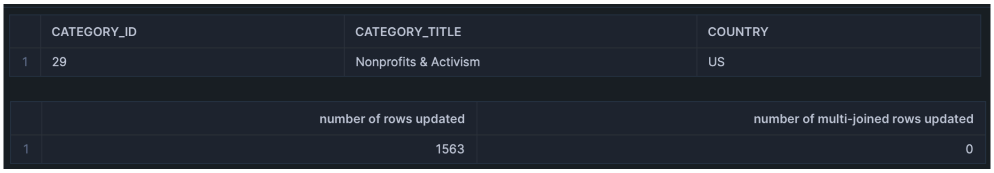
The query identified categories that appear in only one country.



	CATEGORY_TITLE
1	Nonprofits & Activism

Missing Category Titles

A query was executed to find records in the final table where the CATEGORY_TITLE was missing, and then these missing titles were updated based on the category_id.

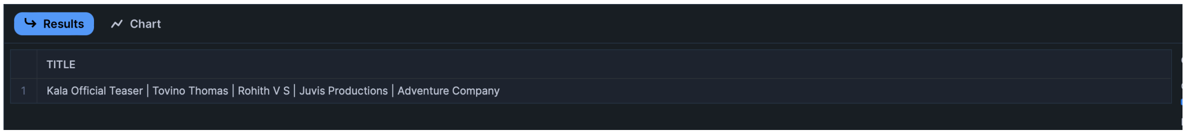


	CATEGORY_ID	CATEGORY_TITLE	COUNTRY
1	29	Nonprofits & Activism	US

	number of rows updated	number of multi-joined rows updated
1	1563	0

Null Channel Titles

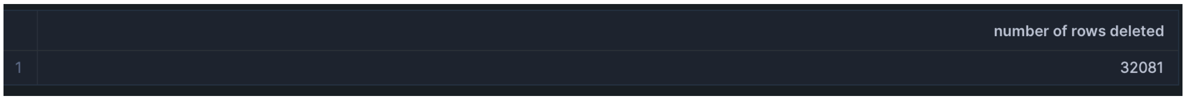
Records with missing channel_title were identified.



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

Invalid Records Removal

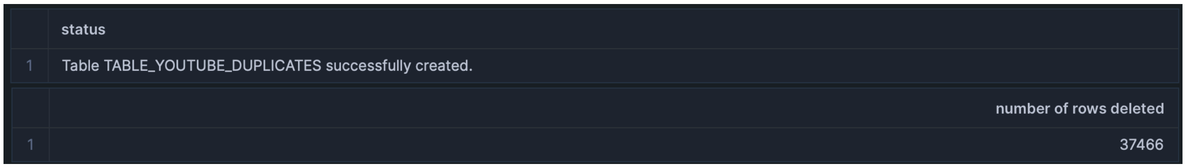
A query was executed to remove invalid records with a specific video_id.



	number of rows deleted
1	32081

Duplicate Records Handling

Duplicates were identified and removed, keeping the record with the highest view_count.

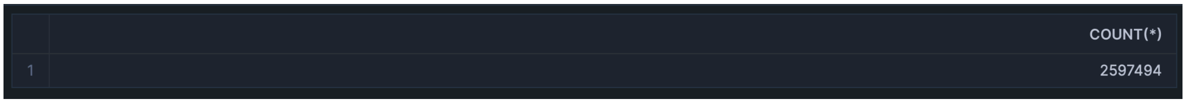


	status
1	Table TABLE_YOUTUBE_DUPLICATES successfully created.

	number of rows deleted
1	37466

Final Record Count

The final count of records in table_youtube_final was checked.

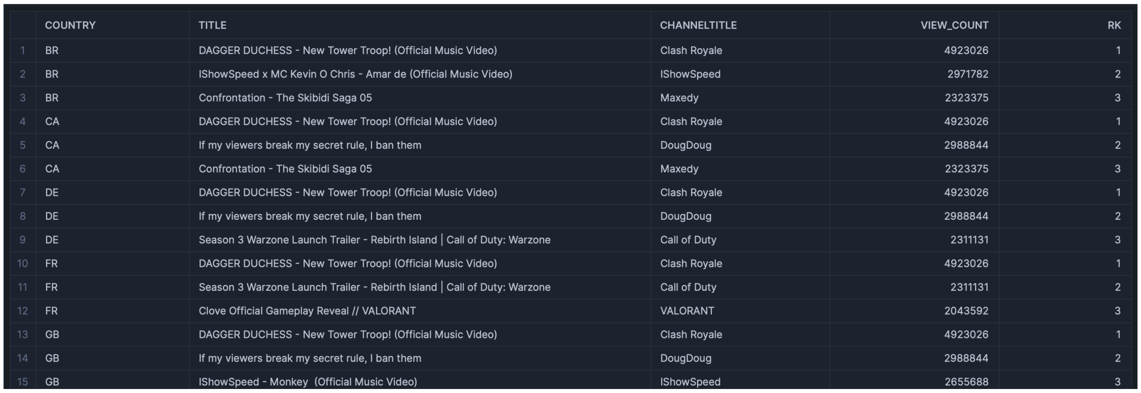


	COUNT(*)
1	2597494

5. Data Analysis

Top Viewed Videos in Gaming

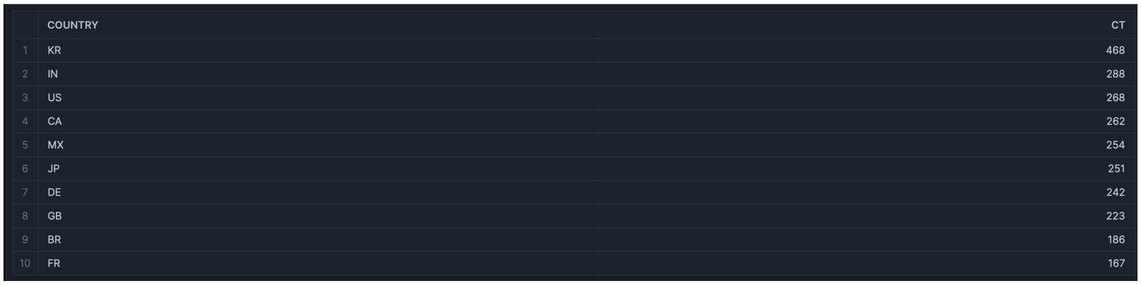
To identify the top 3 most viewed Gaming videos for each country on a specific date (2024-04-01), a ranking analysis was performed. The results showcased the top-performing gaming content across different regions, highlighting the popularity of the gaming category globally.



	COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
1	BR	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
2	BR	IShowSpeed x MC Kevin O Chris - Amar de (Official Music Video)	IShowSpeed	2971782	2
3	BR	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
4	CA	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
5	CA	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
6	CA	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
7	DE	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
8	DE	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
9	DE	Season 3 Warzone Launch Trailer - Rebirth Island | Call of Duty: Warzone	Call of Duty	2311131	3
10	FR	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
11	FR	Season 3 Warzone Launch Trailer - Rebirth Island | Call of Duty: Warzone	Call of Duty	2311131	2
12	FR	Clove Official Gameplay Reveal // VALORANT	VALORANT	2043592	3
13	GB	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
14	GB	If my viewers break my secret rule, I ban them	DougDoug	2988844	2
15	GB	IShowSpeed - Monkey (Official Music Video)	IShowSpeed	2655688	3

Analysis of BTS Mentions Across Countries

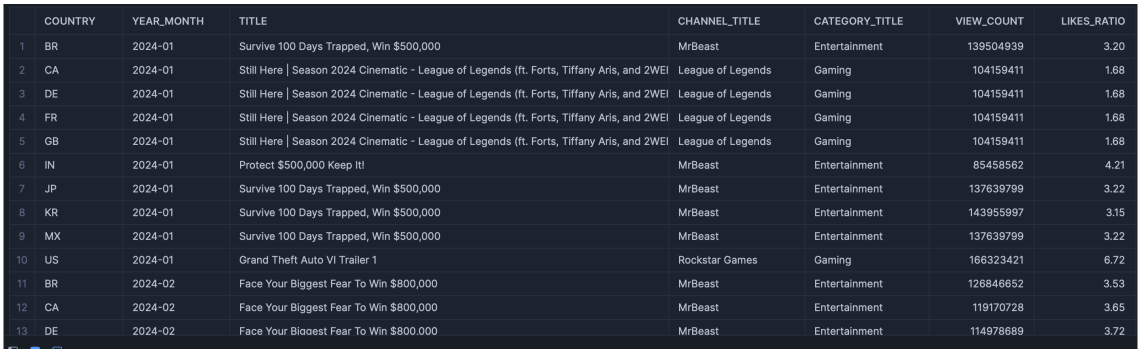
The number of distinct videos mentioning "BTS" in their title was counted for each country. This analysis provided insights into the global reach and influence of BTS.



	COUNTRY	CT
1	KR	468
2	IN	288
3	US	268
4	CA	262
5	MX	254
6	JP	251
7	DE	242
8	GB	223
9	BR	186
10	FR	167

Monthly Most Viewed Videos and Like Ratios

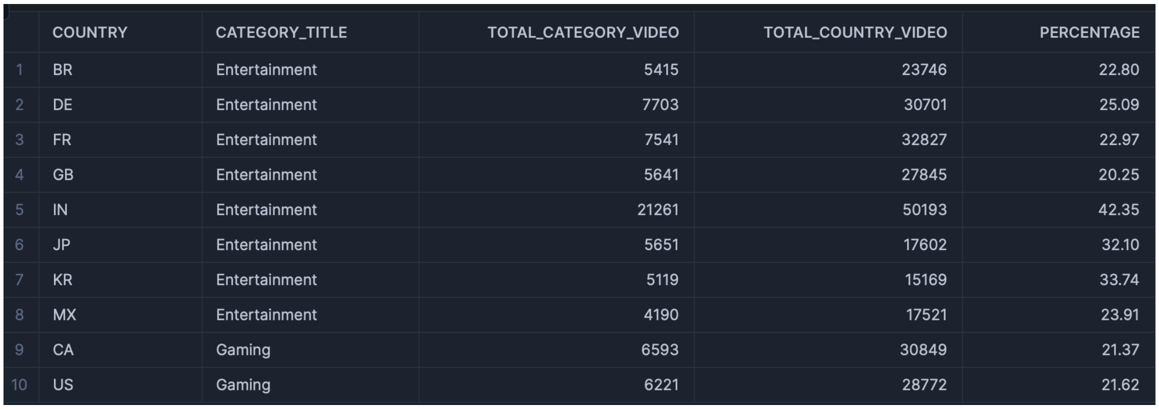
For each country and month in 2024, the most viewed video was identified, along with its like-to-view ratio. This analysis offered a perspective on the most engaging content over time and helped identify trends in viewer preferences.



	COUNTRY	YEAR_MONTH	TITLE	CHANNEL_TITLE	CATEGORY_TITLE	VIEW_COUNT	LIKES_RATIO
1	BR	2024-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	139504939	3.20
2	CA	2024-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2WEI	League of Legends	Gaming	104159411	1.68
3	DE	2024-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2WEI	League of Legends	Gaming	104159411	1.68
4	FR	2024-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2WEI	League of Legends	Gaming	104159411	1.68
5	GB	2024-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2WEI	League of Legends	Gaming	104159411	1.68
6	IN	2024-01	Protect $500,000 Keep It!	MrBeast	Entertainment	85458562	4.21
7	JP	2024-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	137639799	3.22
8	KR	2024-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	143955997	3.15
9	MX	2024-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	137639799	3.22
10	US	2024-01	Grand Theft Auto VI Trailer 1	Rockstar Games	Gaming	166323421	6.72
11	BR	2024-02	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	126846652	3.53
12	CA	2024-02	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	119170728	3.65
13	DE	2024-02	Face Your Biggest Fear To Win $800,000	MrBeast	Entertainment	114978689	3.72

Analysis of Most Popular Categories After 2022

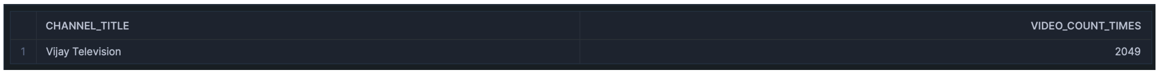
This analysis determined the category with the most distinct videos before 2022 and its percentage of the total distinct videos for each country. The results highlighted historical trends in content popularity.



	COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
1	BR	Entertainment	5415	23746	22.80
2	DE	Entertainment	7703	30701	25.09
3	FR	Entertainment	7541	32827	22.97
4	GB	Entertainment	5641	27845	20.25
5	IN	Entertainment	21261	50193	42.35
6	JP	Entertainment	5651	17602	32.10
7	KR	Entertainment	5119	15169	33.74
8	MX	Entertainment	4190	17521	23.91
9	CA	Gaming	6593	30849	21.37
10	US	Gaming	6221	28772	21.62

Most Distinct Videos by Channel

The channel with the highest number of distinct videos was identified.



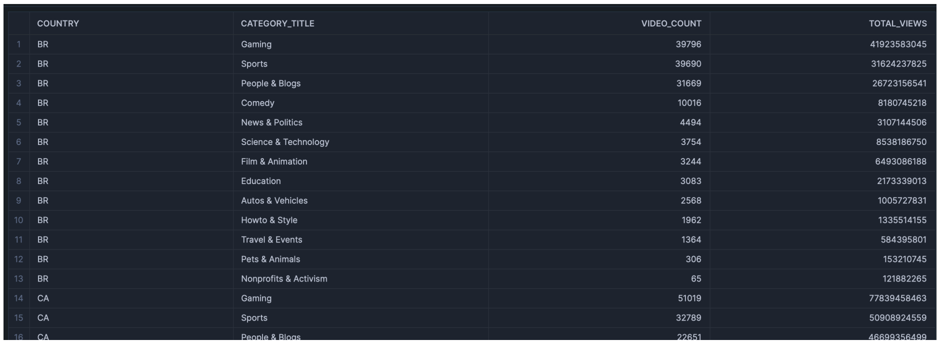
	CHANNEL_TITLE	VIDEO_COUNT_TIMES
1	Vijay Television	2049

6. Business Insights

Popular Categories Excluding Music and Entertainment

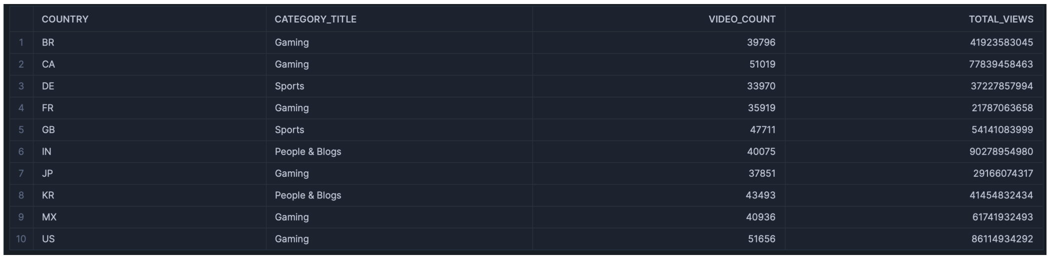
Several analyses were performed to determine the most popular categories excluding 'Music' and 'Entertainment' across various dimensions.

1. **Video Count and Total Views by Category:** Identified the category with the most videos and total views in each country.

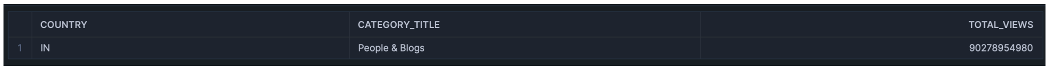


	COUNTRY	CATEGORY_TITLE	VIDEO_COUNT	TOTAL_VIEWS
1	BR	Gaming	39796	41923583045
2	BR	Sports	39690	31624237825
3	BR	People & Blogs	31669	26723156541
4	BR	Comedy	10016	8180745218
5	BR	News & Politics	4494	3107144506
6	BR	Science & Technology	3754	8538186750
7	BR	Film & Animation	3244	6493086188
8	BR	Education	3083	2173339013
9	BR	Autos & Vehicles	2568	1005727831
10	BR	Howto & Style	1962	1335514155
11	BR	Travel & Events	1364	584395801
12	BR	Pets & Animals	306	153210745
13	BR	Nonprofits & Activism	65	121882265
14	CA	Gaming	51019	77839458463
15	CA	Sports	32789	50908924559
16	CA	People & Blogs	22651	46699356499

2. **Most Viewed Category:** The category with the highest total views among the most popular categories was identified.

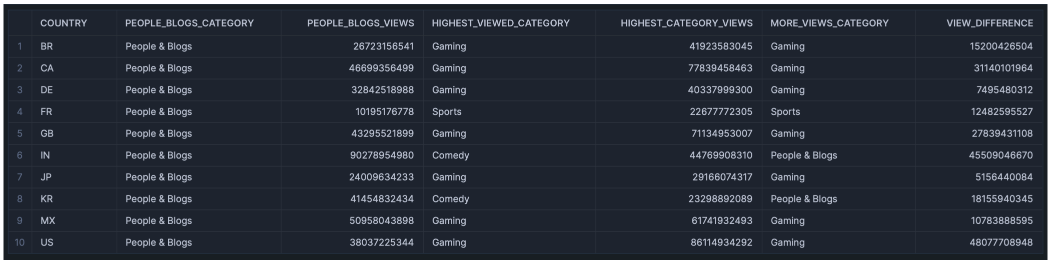


	COUNTRY	CATEGORY_TITLE	VIDEO_COUNT	TOTAL_VIEWS
1	BR	Gaming	39796	41923583045
2	CA	Gaming	51019	77839458463
3	DE	Sports	33970	37227857994
4	FR	Gaming	35919	21787063658
5	GB	Sports	47711	54141083999
6	IN	People & Blogs	40075	90278954980
7	JP	Gaming	37851	29166074317
8	KR	People & Blogs	43493	41454832434
9	MX	Gaming	40936	61741932493
10	US	Gaming	51656	86114934292



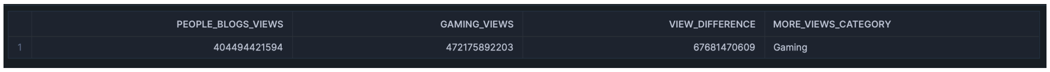
	COUNTRY	CATEGORY_TITLE	TOTAL_VIEWS
1	IN	People & Blogs	90278954980

3. **Views Comparison for 'People & Blogs':** Compared the views of 'People & Blogs' to other categories to identify which category has more views in each country.



	COUNTRY	PEOPLE_BLOGS_CATEGORY	PEOPLE_BLOGS_VIEWS	HIGHEST_VIEWED_CATEGORY	HIGHEST_CATEGORY_VIEWS	MORE_VIEWS_CATEGORY	VIEW_DIFFERENCE
1	BR	People & Blogs	26723156541	Gaming	41923583045	Gaming	15200426504
2	CA	People & Blogs	46699356499	Gaming	77839458463	Gaming	31140101964
3	DE	People & Blogs	32842518988	Gaming	40337999300	Gaming	7495480312
4	FR	People & Blogs	10195176778	Sports	22677772305	Sports	12482595527
5	GB	People & Blogs	43295521899	Gaming	71134953007	Gaming	27839431108
6	IN	People & Blogs	90278954980	Comedy	44769908310	People & Blogs	45509046670
7	JP	People & Blogs	24009634233	Gaming	29166074317	Gaming	5156440084
8	KR	People & Blogs	41454832434	Comedy	23298892089	People & Blogs	18155940345
9	MX	People & Blogs	50958043898	Gaming	61741932493	Gaming	10783888595
10	US	People & Blogs	38037225344	Gaming	86114934292	Gaming	48077708948

4. **Views Comparison Between 'People & Blogs' and 'Gaming':** Compared total views between 'People & Blogs' and 'Gaming' categories.



	PEOPLE_BLOGS_VIEWS	GAMING_VIEWS	VIEW_DIFFERENCE	MORE_VIEWS_CATEGORY
1	404494421594	472175892203	67681470609	Gaming

YouTube channel in the "Gaming" category

The "Gaming" category has significantly more total views (472.18 billion) compared to "People & Blogs" (404.49 billion). The difference of 67.68 billion views indicates that gaming content is more popular on YouTube overall. This suggests that there is a larger audience and higher engagement potential in the gaming category.

Strategy for Different Countries

While the strategy of focusing on "Gaming" generally has broad appeal, it's essential to adapt content to regional tastes. For example:

- **In Countries like the US and Canada:** The gaming category shows clear dominance, making it a strong choice.
- **In Markets where "People & Blogs" is competitive (e.g., India):** Although "Gaming" leads overall, there may be substantial opportunities in "People & Blogs" as well, especially if content is tailored to local preferences.

Starting a YouTube channel in the "Gaming" category is a smart move due to its high viewership and broad appeal. By integrating vlog-style storytelling into my gameplay, including personal experiences and behind-the-scenes content, I can connect with viewers who enjoy vlogs. This strategy lays a solid foundation for growth, though regional content tweaks might be necessary for greater impact.

7. Conclusion

This project successfully implemented a Data Lakehouse architecture in Snowflake to process and analyze YouTube trending data across multiple countries. The data was ingested, cleaned, and analyzed, leading to actionable insights, such as identifying the most promising YouTube content categories for potential expansion. Despite challenges such as handling duplicate and missing data, the project met its objectives and provided a strong foundation for further analysis or business decision-making.