HANDOVER REPORT FOR ASSIGNMENT-1: DATA LAKEHOUSE WITH SNOWFLAKE

ARJUN KHANIJAU
13478401
AUGUST 24, 2024

94693- BIG DATA ENGINEERING
MASTER OF DATA SCIENCE AND INNOVATION
UNIVERSITY OF TECHNOLOGY SYDNEY

TABLE OF CONTENTS

INTRODUCTION

This handover report outlines the analysis of a dataset consisting of CSV and JSON files using a Data Lakehouse architecture with Snowflake. The dataset contains daily records of top trending YouTube videos, sourced through the YouTube API and available on Kaggle, spanning from 2020-08-12 to 2024-04-15. The data covers 10 countries, with each country's data stored in separate files. The analysis involved uploading the data to cloud storage, ingesting it into Snowflake, performing data transformations, and examining the trends across these regions.

PART 1: DATA INGESTION

DATA FLOW DIAGRAM

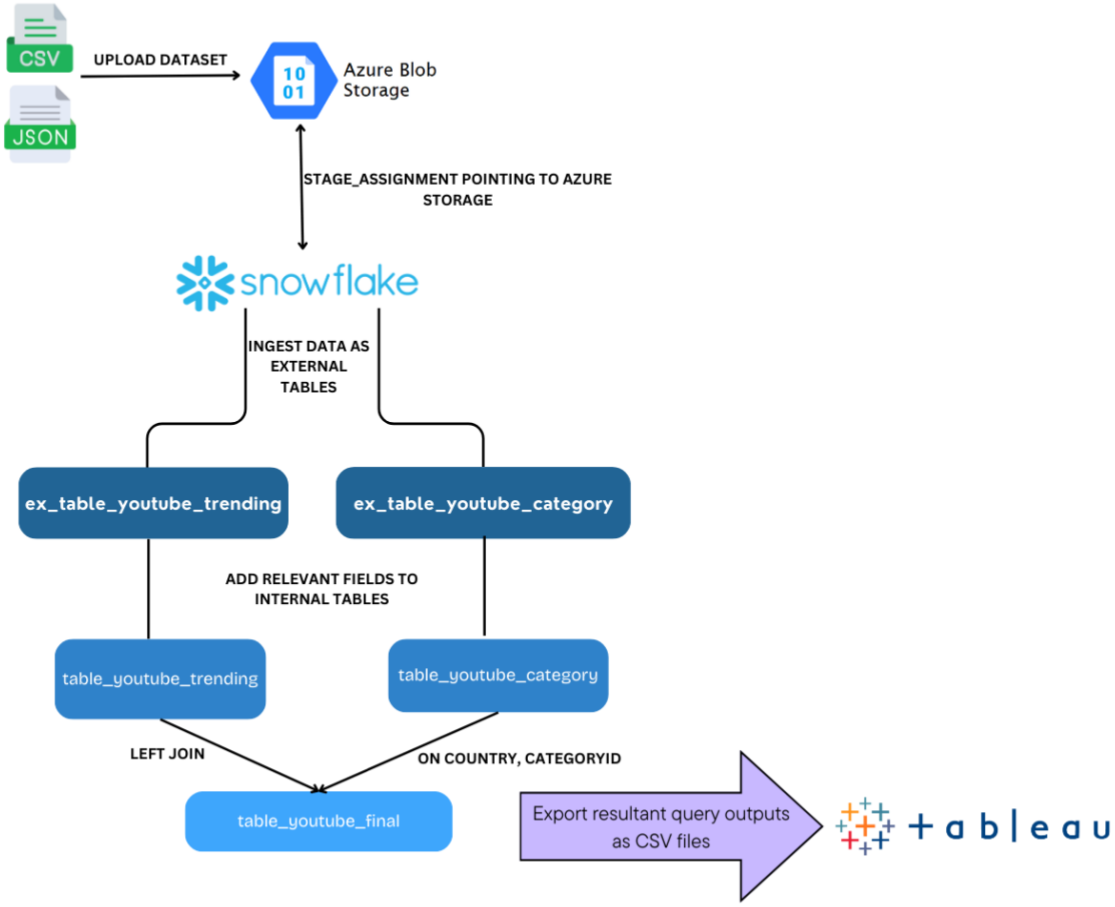

CSV
JSON
UPLOAD DATASET
10 01
Azure Blob Storage
STAGE_ASSIGNMENT POINTING TO AZURE STORAGE
snowflake
INGEST DATA AS EXTERNAL TABLES
ex_table_youtube_trending
ex_table_youtube_category
ADD RELEVANT FIELDS TO INTERNAL TABLES
table_youtube_trending
table_youtube_category
LEFT JOIN
ON COUNTRY, CATEGORYID
table_youtube_final
Export resultant query outputs as CSV files
tableau

Figure 1- DATA FLOW DIAGRAM

(Created using Canva)

1. Data Ingestion using Azure Blob Storage: Datasets are uploaded and stored in Azure Blob Storage, serving as the initial data repository.

2. Snowflake Staging: Data is accessed from Azure Blob Storage through a staging area in Snowflake, preparing it for further processing.

3. Data Processing:

- External Tables: Data is directly queried from Azure Blob Storage using external tables in Snowflake, allowing for efficient data access without moving data.
- Internal Table: Processed data is then stored in internal tables within Snowflake for further analysis and transformations.

4. Data Integration: Data from external and internal tables is joined to create a final output table.

5. Output and Visualization using Snowflake and Tableau: The final data is used for generating visualizations. Simple visualizations are created within Snowflake, while more complex geographical visualizations are handled in Tableau.

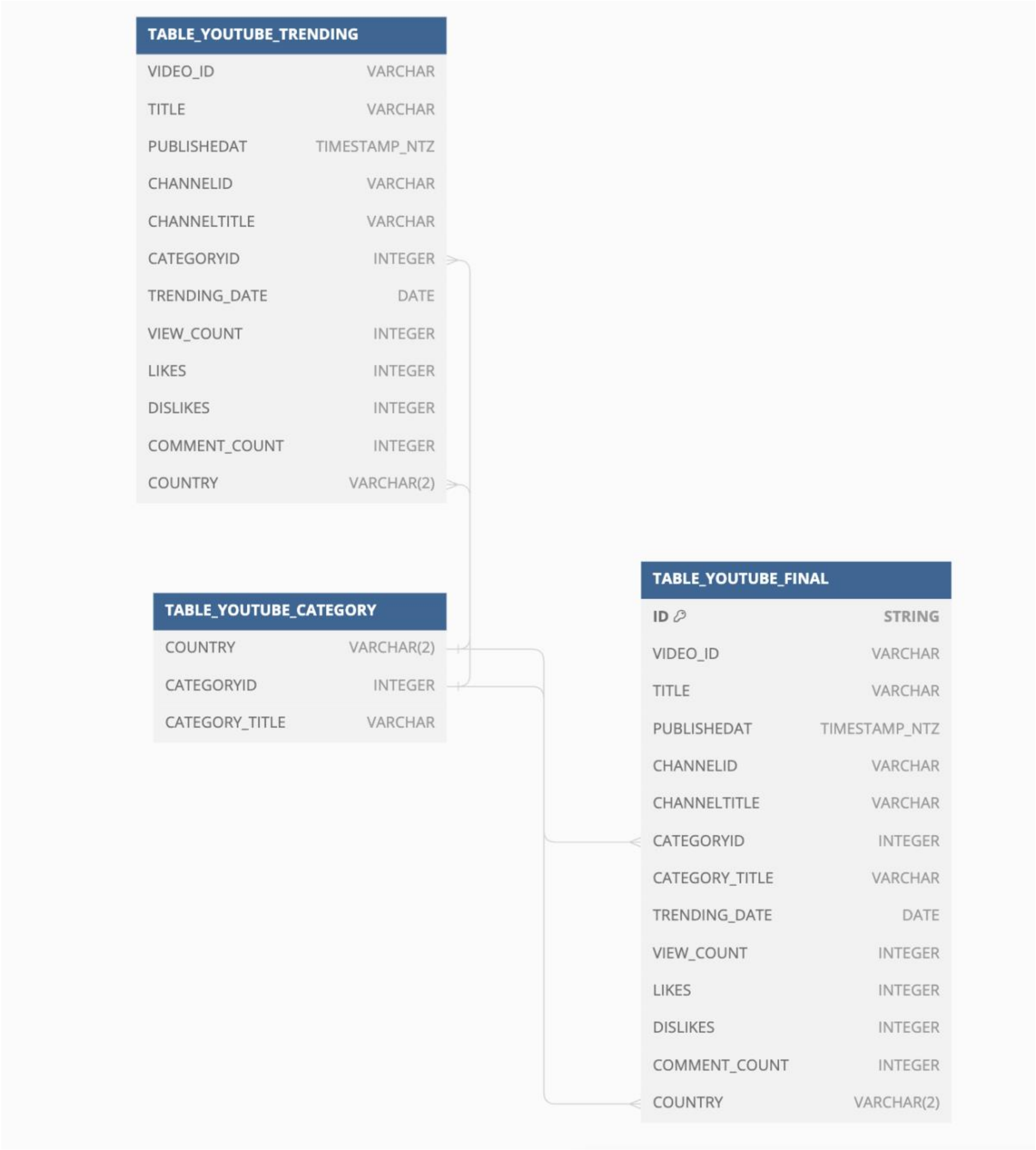
TOOLS USED

1. Azure Blob Storage: Used for storing and managing the project's datasets securely in the cloud.

2. Snowflake: Used for processing and querying data, as well as creating simple visualizations like bar charts.

3. Tableau: Utilized to generate geographical visualizations that couldn't be achieved directly within Snowflake, allowing for deeper insights into location-based data trends.

ENTITY-RELATIONSHIP DIAGRAM (ERD)



TABLE_YOUTUBE_TRENDING	
VIDEO_ID	VARCHAR
TITLE	VARCHAR
PUBLISHEDAT	TIMESTAMP_NTZ
CHANNELID	VARCHAR
CHANNELTITLE	VARCHAR
CATEGORYID	INTEGER
TRENDING_DATE	DATE
VIEW_COUNT	INTEGER
LIKES	INTEGER
DISLIKES	INTEGER
COMMENT_COUNT	INTEGER
COUNTRY	VARCHAR(2)

TABLE_YOUTUBE_CATEGORY	
COUNTRY	VARCHAR(2)
CATEGORYID	INTEGER
CATEGORY_TITLE	VARCHAR

TABLE_YOUTUBE_FINAL	
ID	**STRING**
VIDEO_ID	VARCHAR
TITLE	VARCHAR
PUBLISHEDAT	TIMESTAMP_NTZ
CHANNELID	VARCHAR
CHANNELTITLE	VARCHAR
CATEGORYID	INTEGER
CATEGORY_TITLE	VARCHAR
TRENDING_DATE	DATE
VIEW_COUNT	INTEGER
LIKES	INTEGER
DISLIKES	INTEGER
COMMENT_COUNT	INTEGER
COUNTRY	VARCHAR(2)

Figure 2- ERD showing the internal tables and its attributes

(Created using Dbdiagram.io)

- The above ERD displays the attributes and their data types for each table in the database.

- The ‘ID’ in ‘TABLE_YOUTUBE_FINAL’ is generated using the UUID_STRING() function, which creates a unique identifier for each record, serving as the primary key.

- Left join ‘TABLE_YOUTUBE_TRENDING’ with ‘TABLE_YOUTUBE_CATEGORY’ on ‘CATEGORYID’ and ‘COUNTRY’ to ensure every trending video has a valid category, even if some categories are not present in the category table.

Note: Use METADATA$FILENAME to extract the country code from the respective file names. For example, to extract country 'IN' from ‘youtube_trending/IN_youtube_trending_data.csv’, the index starts at 1, so 'IN' is located at position 18 with a length of 2 characters. This is done using ‘substring(METADATA$FILENAME, 18, 2)’.

PART 2: DATA CLEANING

- Resolve Duplicate Category Titles: Identify and handle duplicate ‘category_title’ in ‘table_youtube_category’.

- Fill Missing Category Titles: Identify and fill missing ‘category_title’ in ‘table_youtube_final’ using ‘CategoryID’.

- Replace Null Values: Update ‘NULL’ values in ‘category_title’ in ‘table_youtube_final’ with accurate data.

- Address Missing Channel Titles: Locate videos in ‘table_youtube_final’ that lack a ‘ChannelTitle’.

- Remove Invalid and Duplicate Records: Delete records with erroneous ‘video_ID’ values (‘#NAME?’) and handle duplicates by keeping the highest ‘view_count’ on that particular date.

- Isolate and Clean Duplicates: Create ‘table_youtube_duplicates’ to isolate and remove bad duplicates from ‘table_youtube_final’.

PART 3: DATA ANALYSIS

1. Top 3 Most Viewed Gaming Videos by Country on 1st April, 2024

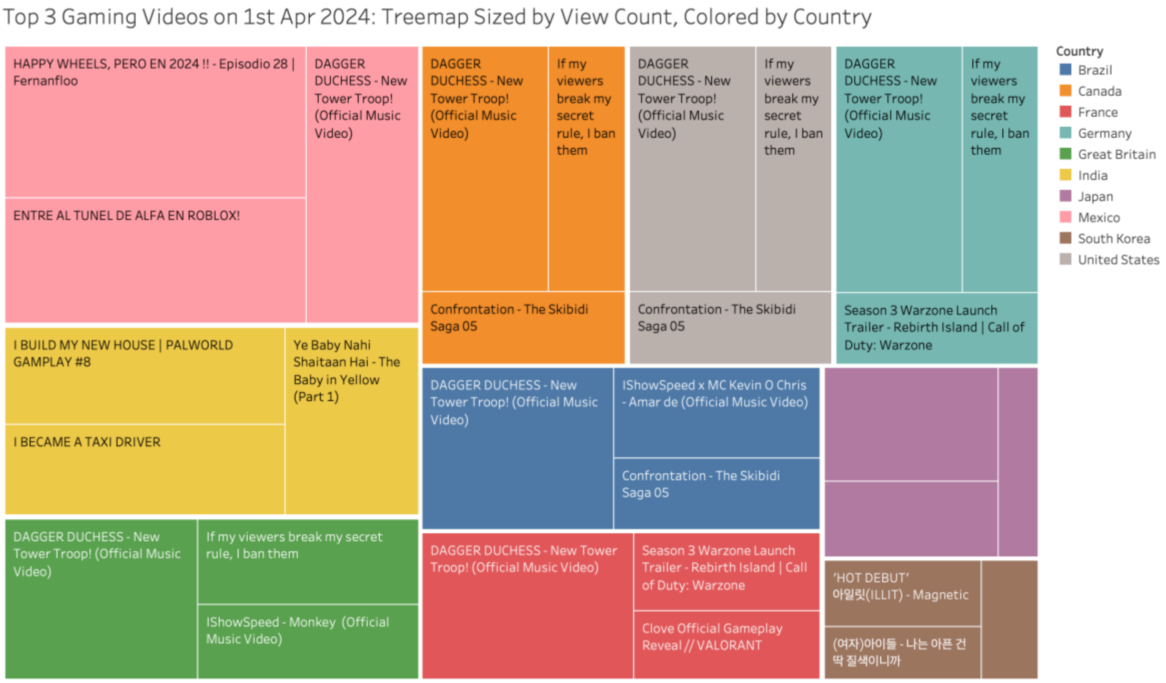


Figure 3- Treemap to visualize the top 3 gaming videos on 1st April, 2024

This treemap visualizes the top 3 most viewed gaming videos by country on 1st April 2024. The size of each tile represents the view count, with larger tiles indicating higher view counts. The color represents the country. Mexico has the largest tile, showing it had the highest view counts for gaming videos on this date, followed by India. South Korea has the smallest tile, indicating lower view counts.

Note: Due to size constraints, some tiles for Japan and one for South Korea do not display title names.

2. Count of Distinct Videos with 'BTS' in Title by Country

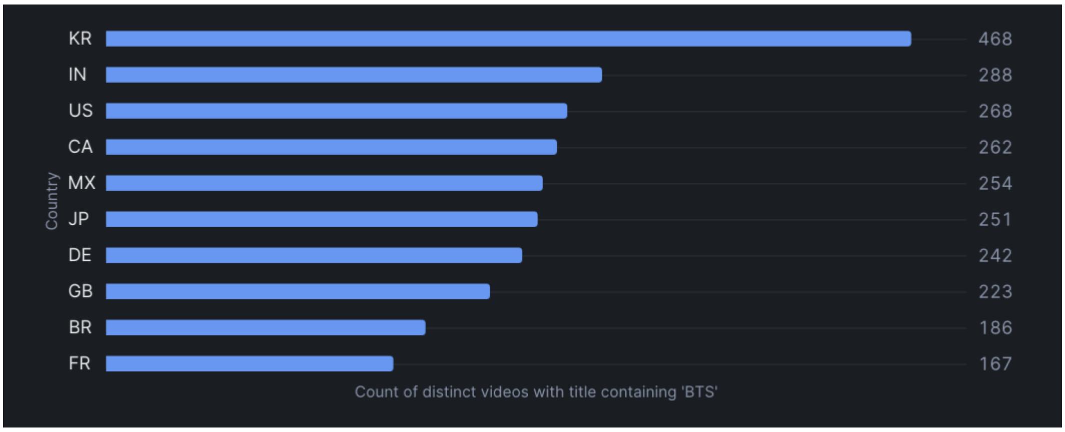


Figure 4- Bar chart showing the country-wise count in descending order

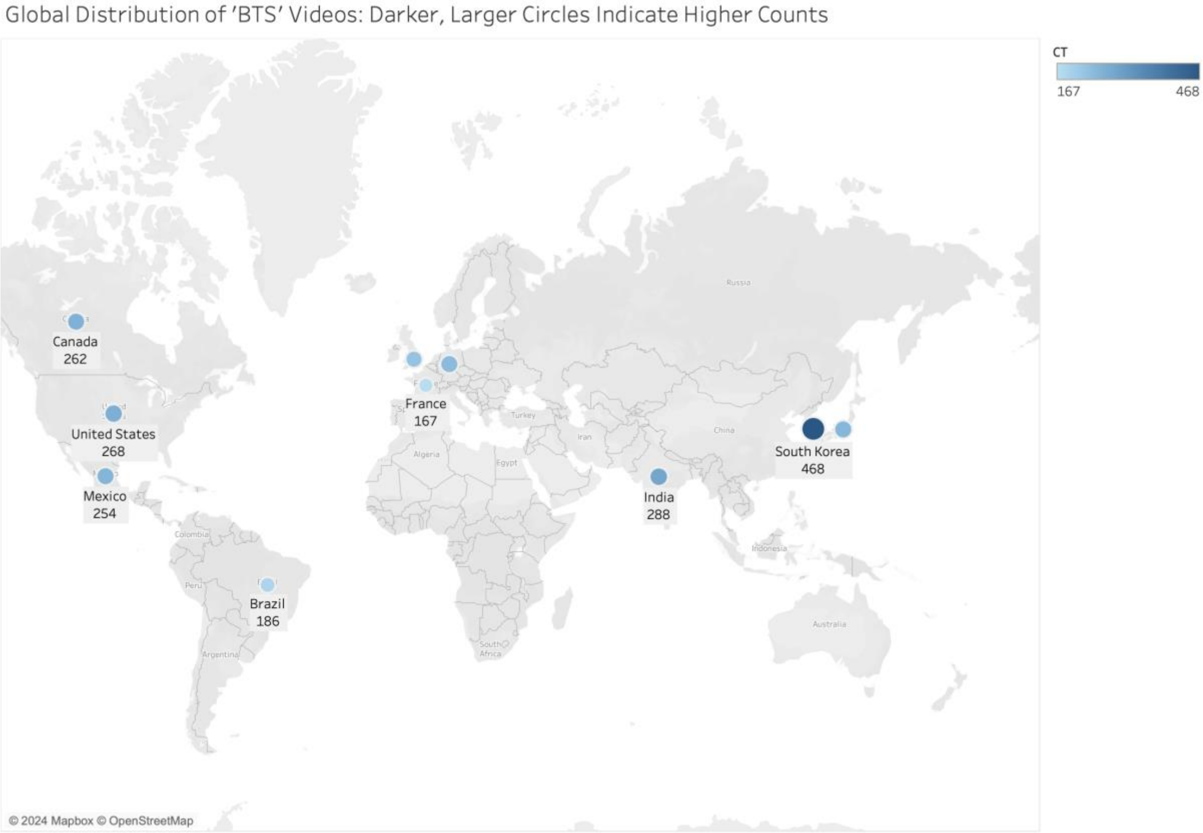


Figure 5- Map showing the global distribution of 'BTS' videos

This map visualizes the distribution of videos with 'BTS' in the title across countries. Darker and larger circles represent higher counts. South Korea has the highest count with 468 videos, followed by India and the USA. The lowest count is in France.

3. Top Viewed Video by Country and Month in 2024 with Likes Ratio

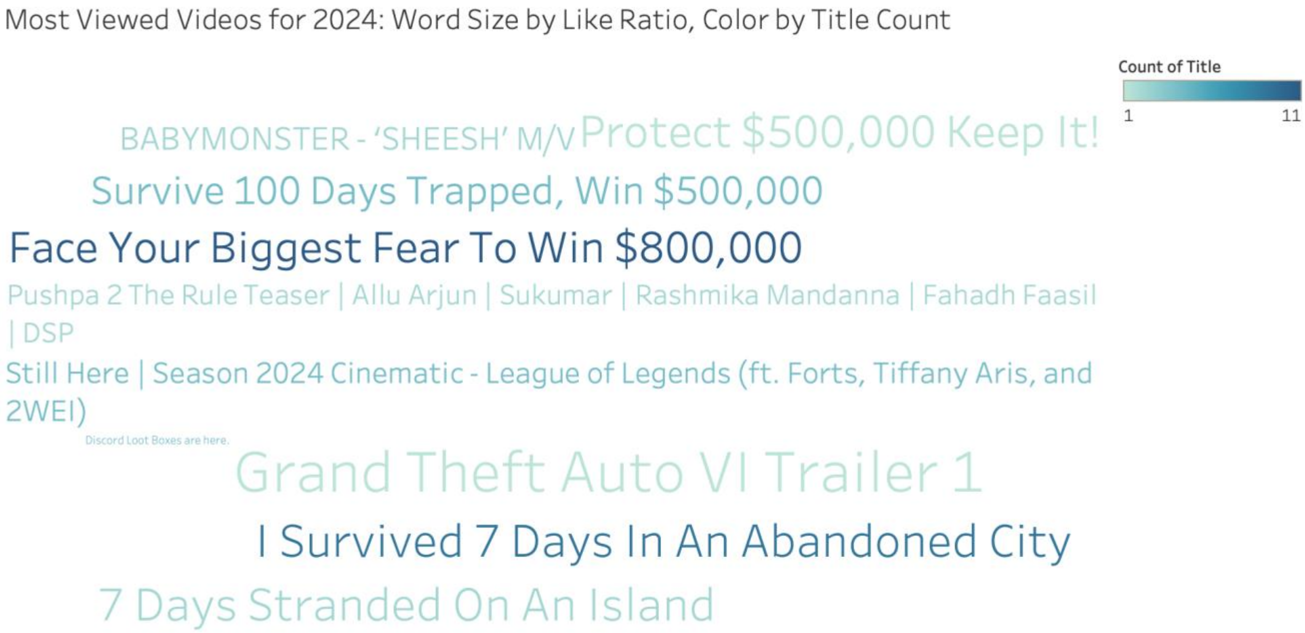
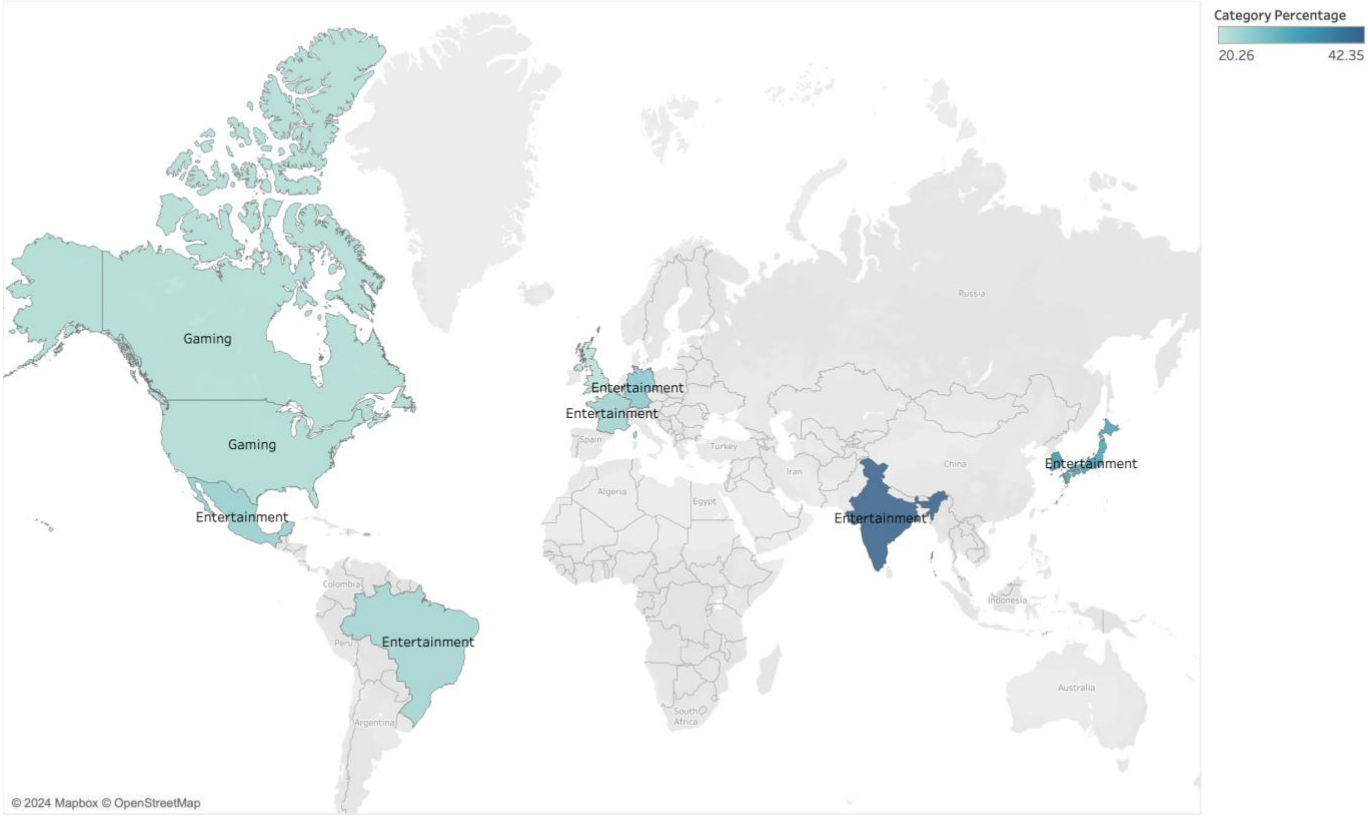


Figure 6- Word Cloud for most viewed videos in 2024

This word cloud visualizes the most viewed videos by country and month in 2024. Word size represents the maximum like ratio, with larger words indicating higher like ratios. Color intensity reflects the title count, with darker shades indicating more frequent titles. For example, 'Grand Theft Auto VI Trailer 1' is the largest word with a like ratio of 6.73, while 'Face Your Biggest Fear To Win $800,000' is the darkest word, indicating that the title appears 11 times for different countries and trending months.

4. Top Category by Distinct Videos and Percentage per Country (Data from 2022)

Top Video Category by Distinct Videos for Each Country



© 2024 Mapbox © OpenStreetMap

Figure 7- Map showing top video categories by distinct videos for each country, with its category percentage by color intensity

This map visualizes the top video categories by distinct video percentage for each country using data from 2022. Entertainment is the leading category in 8 countries, while Gaming dominates in the USA and Canada.

$$Category\ Percentage = \frac{Total\ Number\ of\ Distinct\ Videos\ in\ Category}{Total\ Number\ of\ Distinct\ Videos\ in\ Country} * 100$$

The darker the color, the higher the category percentage, with India showing the highest category percentage at 42.35%.

5. Top Channel by Distinct Video Count

Vijay Television has produced the most distinct videos, with a total of 2,049 videos.

PART 4: BUSINESS QUESTION

Which video category, excluding 'Music' and 'Entertainment', should be focused on when launching a new YouTube channel to maximize the chances of trending? Additionally, will this strategy be effective across different countries?

Logic to identify the category: Growth Rate

We follow these steps to answer the above business question:

1. Identify the Time Frame:

- Parameter: Quarter-wise growth rate in view count.

- Reason: Growth rate is a strong indicator of emerging trends, instead of simply considering average or total view count. Analyzing the last two quarters (Q4 2023 and Q1 2024) helps identify categories gaining traction in recent times, making it a reliable parameter for predicting future trends.

2. Determine the Last Trending Date in Dataset:

- Action: Found the maximum trending date to be 15th April 2024.

- Reason: Ensures the analysis captures the most recent data, reflecting the latest trends.

3. Account for Multiple View Counts:

- Challenges:
  - Trending on Multiple Days: The last entry has the maximum view count.
  - Trending in Multiple Countries: Metrics like view count and likes are the same across countries.

- Solution: Considered only the last view count for each video ID in that quarter. avoiding inflated data from multiple entries for the same video.

4. Calculate the Average View Count Across Countries:

- Action: Averaged the view count across countries to determine which category (excluding 'Music' and 'Entertainment') had the highest growth rate.

- Reason: By averaging across countries, the analysis accounts for global trends, offering a more balanced view of category performance.

5. Apply a Minimum Threshold of 100 Distinct Videos per Category:

- Reason: Setting a threshold of at least 100 distinct videos helps to mitigate the impact of skewed data. This ensures the growth rates are more representative and less susceptible to anomalies or outliers.

6. Evaluate the Strategy Across Countries:

- Action: For the identified top category, calculated country-wise growth rates.

- Reason: This step assesses whether the identified category performs well globally or if its success is confined to specific regions, helping determine if the strategy is universally applicable.

This approach, focusing on quarter-wise growth rate and considering a unique view count for each video in a specific quarter, ensures that the recommended category is not only trending but also has a strong likelihood of sustaining its performance across different markets.

SQL Subquery Explanations and Final Result Summary:

The concise explanations for each subquery and the final result are as follows:

- max_view_counts: Calculates the maximum view count for each video within Q4 2023 and Q1 2024, excluding "Music" and "Entertainment."

- quarterly_avg_view_count: Computes the average view count and the number of distinct videos for each category and quarter.

- filtered_categories: Filters categories with at least 100 distinct videos in both quarters and calculates the growth rate between Q4 2023 and Q1 2024.

- Final Select: Retrieves and orders the categories by growth rate, showing the average view counts and growth rate for each category.

Findings:

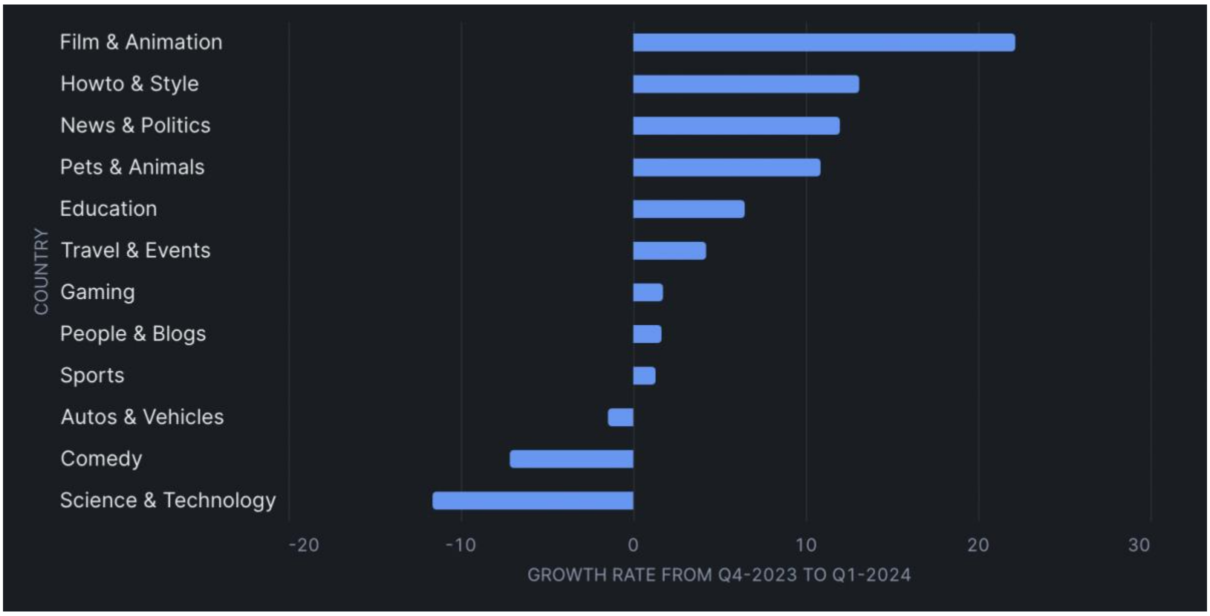


Figure 8- Category-wise growth rate from Q4 2023 to Q1 2024, averaged across countries

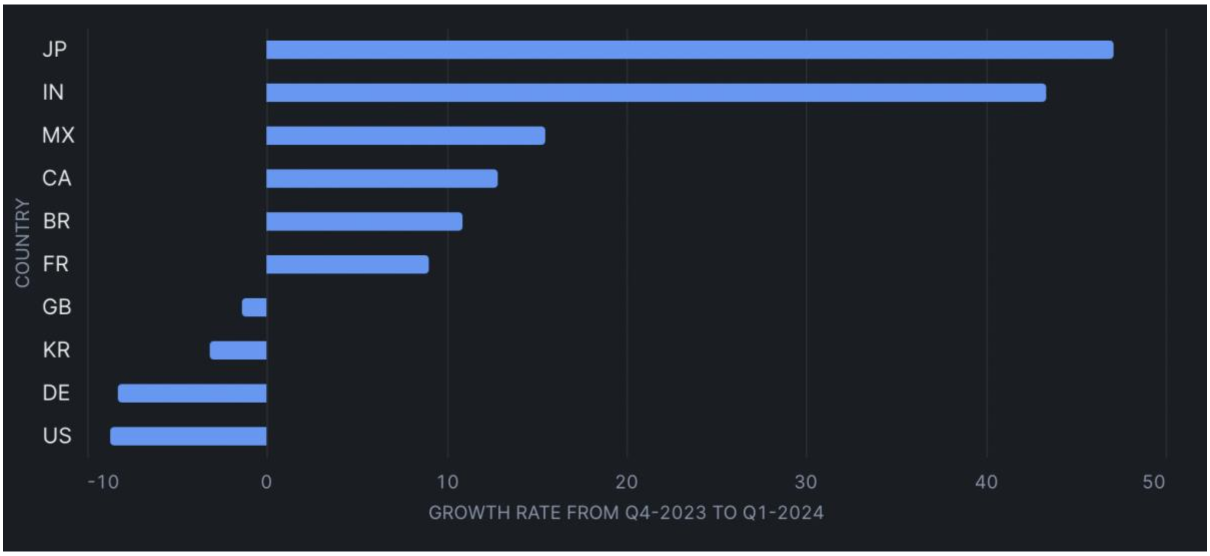


Figure 9- Country-wise growth rate from Q4 2023 to Q1 2024 for Film & Animation

FILM AND ANIMATION has the highest growth rate at 22.18%. Japan leads with a growth rate of 47%, followed by India at 43%. Four countries show positive growth rates ranging from 9% to 15%. However, four countries (Great Britain, South Korea, Germany, and the United States) have slightly negative growth rates, ranging between -1% and -9%.

Advantages of Focusing on Film and Animation:

- High Growth Rate: This category has demonstrated the highest overall growth rate at 22.18%, indicating strong and increasing viewer interest in this category.

- Market Potential in High Population Density Countries:
  - Japan: With a growth rate of 47%, Japan is leading the charge. Japan's high population density amplifies the impact, as content can quickly reach a large audience.

  - India: With a growth rate of 43%, India presents a massive market. The high population density means there's significant potential for viewership growth, especially as internet penetration continues to rise.

- Global Appeal: Film and Animation content often has universal themes that can cross cultural barriers, making it appealing across various countries and demographics.

Negatives of Focusing on Film and Animation:

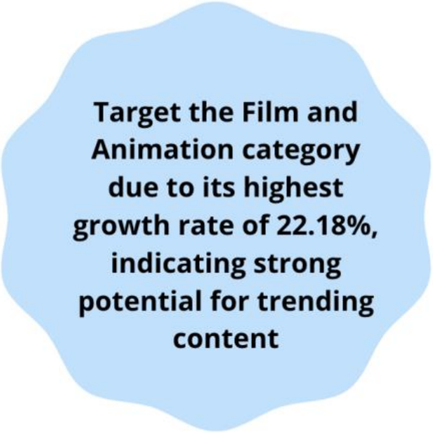
- Negative Growth in Some Key Markets: Great Britain, South Korea, Germany, and the United States- These countries have shown slightly negative growth rates ranging from -1% to -9%. This suggests that in some significant markets, the interest in Film and Animation may be reducing or becoming saturated.

- High Competition: The popularity of Film and Animation means there is also a high level of competition. Breaking into this category could be challenging, especially in markets with slower growth or negative trends.

- Cultural Sensitivity: While Film and Animation have global appeal, content that resonates well in Japan or India may not have the same impact in Western countries, leading to varied performance across regions.

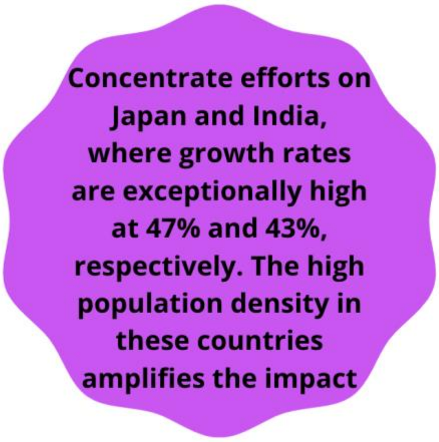
CONCLUSION

Based on the output of SQL queries and the analysis of growth rates, the following recommendations are provided to optimize video category selection and enhance visibility on YouTube:



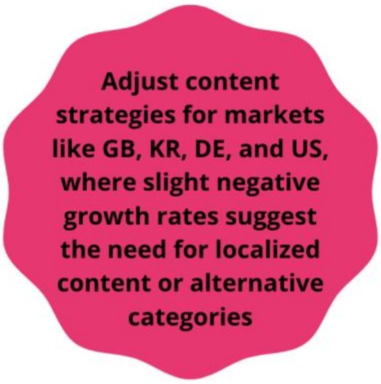
Target the Film and Animation category due to its highest growth rate of 22.18%, indicating strong potential for trending content

Recommendation 1- Focus on Film and Animation



Concentrate efforts on Japan and India, where growth rates are exceptionally high at 47% and 43%, respectively. The high population density in these countries amplifies the impact

Recommendation 2- Prioritize High-Potential Markets



Adjust content strategies for markets like GB, KR, DE, and US, where slight negative growth rates suggest the need for localized content or alternative categories

Recommendation 3- Adapt Strategy for Regional Variations

FUTURE WORK

1. Expand Data for Sentiment Analysis of Comments: The current dataset includes only comment counts. For deeper insights, request additional data sources that provide the actual content of comments. This will enable sentiment analysis with NLP techniques to better understand audience reactions and preferences.

2. Address Data Errors: The dataset contains errors such as '#REF!' and '#VALUE!' in the 'video_ID' field. Future work should focus on cleaning these data issues to ensure accuracy and consistency in the analysis.

3. Incorporate Engagement Metrics: Beyond view counts and likes, consider using additional data sources for adding other engagement metrics like watch time, share count, and subscriber growth to gain a more comprehensive understanding of video performance.