TRENDING NOW: ON YOUTUBE

DATA LAKEHOUSE WITH SNOWFLAKE

WONGWARA WIJARA 14191732
BIG DATA ENGINEERING | AUTUMN 2023

UTS

PROJECT OVERVIEW

YouTube, sometimes called a platform where "cat videos and world-changing ideas collide," offers diverse content ranging from the funny to the informative. As we navigate the world of online video content, it's necessary to remember that while YouTube supports a wide range of genres, not all content providers make it to the top trending lists.

The question is, if one were to embark on the adventure of launching a new YouTube channel while avoiding the domains of "Music" and "Entertainment," what type of video category should be generated in order to win a coveted spot among the platform's top trends? Furthermore, does this method work across borders, overcoming geographical limits to reach audiences worldwide?
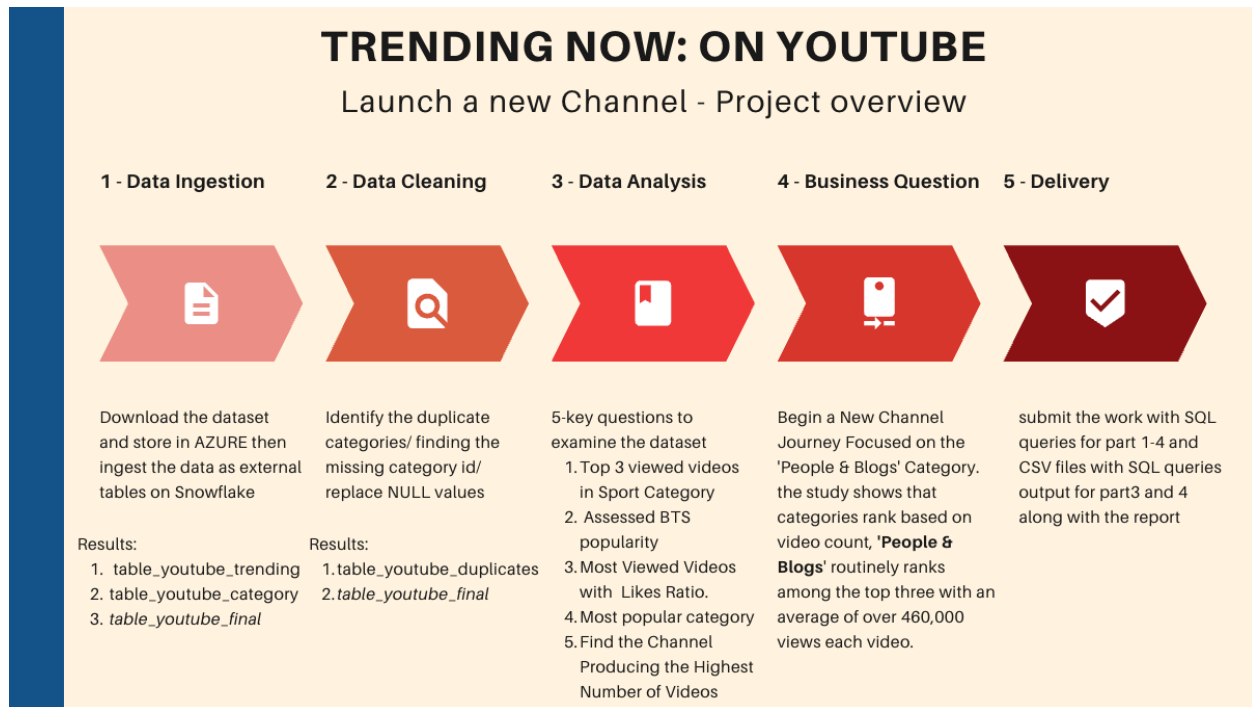

TRENDING NOW: ON YOUTUBE
Launch a new Channel - Project overview

1 - Data Ingestion

Download the dataset and store in AZURE then ingest the data as external tables on Snowflake

Results:
1. table_youtube_trending
2. table_youtube_category
3. *table_youtube_final*

2 - Data Cleaning

Identify the duplicate categories/ finding the missing category id/ replace NULL values

Results:
1. table_youtube_duplicates
2. *table_youtube_final*

3 - Data Analysis

5-key questions to examine the dataset
1. Top 3 viewed videos in Sport Category
2. Assessed BTS popularity
3. Most Viewed Videos with Likes Ratio.
4. Most popular category
5. Find the Channel Producing the Highest Number of Videos

4 - Business Question

Begin a New Channel Journey Focused on the 'People & Blogs' Category. the study shows that categories rank based on video count, **'People & Blogs'** routinely ranks among the top three with an average of over 460,000 views each video.

5 - Delivery

submit the work with SQL queries for part 1-4 and CSV files with SQL queries output for part3 and 4 along with the report

Figure 1: Project Overview

1. DATA SOURCE EXPLORATION

Figure 2: Kaggle data source - YouTube Trending Video Dataset.

The primary goal of this project is to analyse the daily trending videos on YouTube using a Data Lakehouse powered by Snowflake. We will extract data from a large Kaggle dataset, which includes both CSV and JSON files, and securely store it in Microsoft Azure.

This dataset contains daily records of trending YouTube videos in India, the USA, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea, and Japan. We intend to ingest, transform, and analyse this data using Snowflake's capabilities to acquire essential insights into YouTube's top-trending videos.

video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	view_count	likes	dislikes	comment_count	comments_disabled
s9FH4rDMvds	LEVEI UM FORA? FINGI EST	2020-08-11T22:21:49Z	UCGfBwrCoi9ZJjKiUd8N	Pietro Guedes	22	2020-08-12T00:00:	263835	85095	487	4500	FALSE
jbGRowa5tlk	ITZY ,ÁúNot Shy,Áù M/V TE/	2020-08-11T15:00:13Z	UCaO6TYtIC8U5ttz62hT	JYP Entertainment	10	2020-08-12T00:00:	6000070	714310	15176	31040	FALSE
3EfkCrXKZNs	Oh Juliana PARViDIA - MC N	2020-08-10T14:59:00Z	UCoXZmVma073v5G1c	As IrmVÉs Mota	22	2020-08-12T00:00:	2296748	39761	5484	0	TRUE

Figure 3: Sample data in CSV file.

```
{
    "kind": "youtube#videoCategoryListResponse",
    "etag": "kBCr3I9kLHHU79W4Ip5196LDptI",
    "items": [
        {
            "kind": "youtube#videoCategory",
            "etag": "IfWa37JGcqZs-jZeAyFGkbeh6bc",
            "id": "1",
            "snippet": {
                "title": "Film & Animation",
                "assignable": true,
                "channelId": "UCBR8-60-B28hp2BmDPdntcQ"
            }
        },
        {
            "kind": "youtube#videoCategory",
            "etag": "5XGylIs7zkjHh5940dsT5862m1Y",
            "id": "2",
            "snippet": {
                "title": "Autos & Vehicles",
                "assignable": true,
                "channelId": "UCBR8-60-B28hp2BmDPdntcQ"
            }
        }
```

Figure 4: Sample data in JSON file.

2. SETUP

Focus on data ingestion from Microsoft Azure into Snowflake; this project will be divided into four parts – Data Ingestion, Data Cleaning, Data Analysis and launch a new channel.

Setup Guide - Part 1: Data Ingestion

1. Download the dataset – from the google drive

    a. Trending data (year 2020-2022)

    b. category data

2. Upload the Dataset to Azure Blob Storage – Create a container to organise data.

Figure 5: Create a container called bde-at1 in Azure.

3. Ingest Data into Snowflake – Create an external stage, then transfer tables into trending table and category table to store the trending data and category data

Figure 6: Trending table after correcting the data type in Snowflake with 1,175,478 rows.

Figure 7: Category table from JSON file with 342 rows.

4. Create the final table by combining the two tables with 'id' to identify the video

VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE	CATEGORYID	TRENDING_DATE	VIEW_COUNT	LIKES	DISLIKES	COMMENT_COUNT	COMMENT_DISABLED	COUNTRY
uq5LClQN3cr	안녕하세요	2020-08-09T09:32:48	UCu9BCtGIE	보검 BK	24	12/8/2020	5947503	53326	105756	139946	FALSE	KR
l-ZbZCHsHDC	부락토스의	2020-08-12T09:00:08	UCRuSxVu4ir	총몇명	1	12/8/2020	963384	28244	494	3339	FALSE	KR
9d7jNUjBoss	평생 반성하	2020-08-10T09:54:13	UCMVC92EQ	양팡 YangPang	22	12/8/2020	2950885	17974	68898	50688	FALSE	KR
3pI_L3-sMVg	ㅇ ㅏ ㄴ ㅕ	2020-08-11T15:00:58	UCkQCwnkQ	꽈뚜룹 Quaddurup	24	12/8/2020	1743374	36893	1798	8751	FALSE	KR
zrsBjYukE8s	박진영 (J.Y.	2020-08-11T09:00:13	UCaO6TYtlC	JYP Entertainment	10	12/8/2020	3433885	353337	9763	23405	FALSE	KR
jbGRowa5tlh	ITZY "Not Sh	2020-08-11T15:00:13	UCaO6TYtlC	JYP Entertainment	10	12/8/2020	6000070	714301	15176	31040	FALSE	KR
X-TPQPEyRG	흑인 마동석	2020-08-10T09:37:33	UCpCiIDf9Ur	드림텔러(DreamTelle	1	12/8/2020	690331	6657	393	834	FALSE	KR
REUA4roJnd	진심으로 사	2020-08-11T14:00:01	UCwZTeeKyt	임다TV	24	12/8/2020	659743	13040	13337	18286	FALSE	KR
7FOi40uehC	집에 혼자 있	2020-08-12T09:00:02	UCCJkwrmhl	밍포발랄Mingggo	23	12/8/2020	340149	13166	338	5697	FALSE	KR
0dsnm3n6Zd	정윤호가 무	2020-08-11T09:30:00	UC0SoPwEHl	달라스튜디오	23	12/8/2020	351264	11125	103	2324	FALSE	KR

Figure 8: Final table with 1,175,478 rows

Setup Guide - Part 2: Data Quality Assessment and Cleaning

This data quality assessment and cleaning process aims to provide a more reliable dataset for analytical purposes. Further steps can be taken to address any remaining data quality issues and improve overall data integrity.

1. Identification of Duplicate Categories

We identified discrepancies within the "Comedy" category during the data quality assessment. It appeared 22 times when the "categoryId" element was disregarded for identifying duplicates. Further investigation may be necessary to ensure data consistency in this category.

2. Categories are only available in one country.

Our analysis revealed that the category "Non-profits & Activism" is present in only one country within the dataset. This finding can provide insights into the regional distribution of video categories.

3. Finding the Missing Category ID

Records in the final dataset that lacked a category title were allocated a categoryId value of 29, reflecting the category "Non-profits & Activism." This mapping was created by combining data from the trending database and the category table.

4. Replacing NULL Values

To enhance data completeness, missing category values were replaced with the corresponding category names from the category table. This corrective action impacted 3,162 rows in the dataset.

5. Videos Without Channel Titles

Within the dataset, a single video identified by the video ID "9b9MovPPewk" lacked a channel title. This observation may require further investigation and validation.

6. Deleting Specific Records

In the interest of data quality, 14,619 rows containing the video ID "#NAME?" were removed from the dataset.

7. Creating a Table for Bad Duplicates

A total of 37,842 rows were classified as "bad duplicates." These records shared identical video IDs, countries, and trending dates but displayed lower view counts than their counterparts.

8. Removing Duplicates

After deleting duplicates (37,842 rows) from the final dataset, we obtained a corrected dataset including 1,123,017 rows, ensuring data consistency and correctness for subsequent studies.

Setup Guide - Part 3: Issues

As shown in the table below, several key obstacles were faced and resolved when handling data and preparing it for analysis. The challenges ranged from data type issues to difficulties with Azure role assignment and dealing with duplicate records.

| Issue	Description	How to solve
1. Failed to copy the name from AZURE_CONSENT_URL for role assignment	My Snowflake instance cannot access or interact with my Azure storage resources without a role assignment.	login to My Apps dashboard to receive the name and add it as a role assignment for `Storage Blob Data Owner`
2. Data Type Mismatch	There was a data type mismatch in the column publishAt, trending_date from create a trending table.	Checking the data and ensuring that the correct data type was assigned to each column
3. Handling Duplicates	There are some rows missing while removing the duplicates eliminated twice as many records as expected (75,684 instead of 37,842).	Modifying the query to use the id column as the only identifier for duplicate elimination.

Table 1: Issues faced in the project.
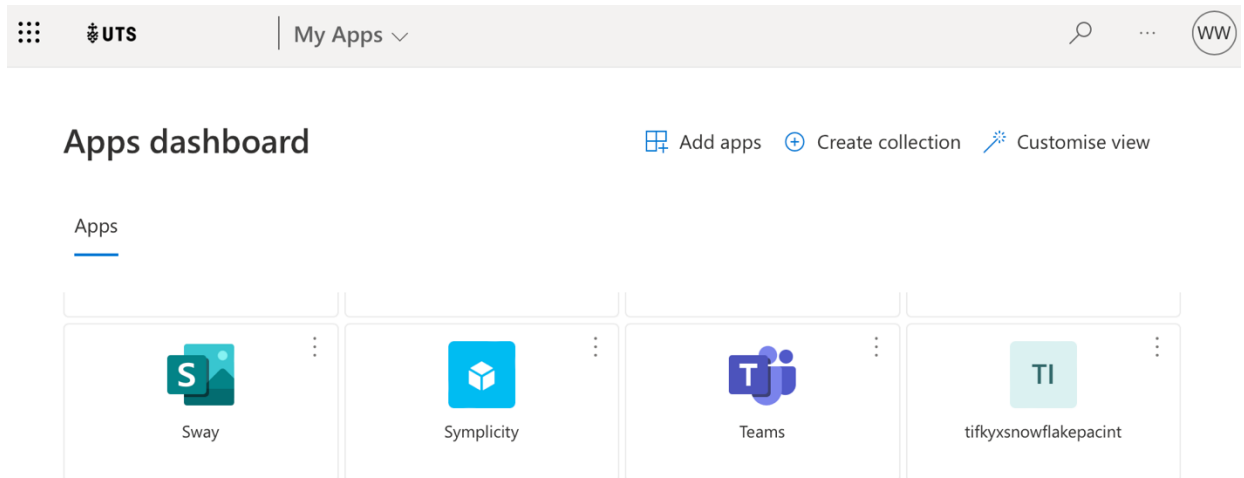
Figure 9: My Snowflake name in the Microsoft Apps dashboard (tifyxsnowflakepacint) for solve the first issue - My Apps dashboard.

3. YOUTUBE DATA ANALYSIS

To gather significant insights for the successful launch of a new YouTube channel, we have formulated five key questions to examine our dataset, with an emphasis on avoiding the "Music" and "Entertainment" domains:

Question 1: Discover the Top 3 Most Viewed Videos in the "Sports" Category.

To retrieve the top 3 most viewed videos, data from 17 October 2021 was chosen, with the "Sports" Category.

COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RANK
MX	Race Highlights | 2021 Turkish Grand Prix	FORMULA 1	5,165,706	1
MX	Eliminatorias | Brasil 4-1 Uruguay | Fecha 12	CONMEBOL	4,354,963	2
MX	Argentina 3-0 Uruguay I Eliminatorias a Catar 2022	TyC Sports	2,794,195	3
RU	Highlights: Watford 0-5 Liverpool | Mane, Salah & a Firmino hat-trick for se	Liverpool FC	1,705,824	1
RU	Т-34 VS боец UFC. Хадис VS Андрюшко. Назир. Армеец VS Нилов. Мар	HFC MMA	1,203,498	2
RU	Knockout Of The Year?! Superbon SHOCKS Giorgio Petrosyan 😱	ONE Championship	1,020,267	3
US	Dice Stacks from $1 to $100	That's Amazing Shorts	69,866,143	1
US	FIGHT HIGHLIGHTS | Anthony Joshua vs. Oleksandr Usyk	DAZN Boxing	8,639,065	2
US	Sore loser! An idiot! Tyson Fury reveals what was said between him & Deo	BT Sport Boxing	6,913,800	3

Figure 10: Results from the Top 3 Viewed videos in Mexico, Russia, and the United States

According to the findings, the United States had the most viewed video in the "Sports" category, with over 69 million views. Other nations' top three most viewed videos ranged from 1.7 million to 6.9 million views. This significant difference shows the importance of the most viewed sports video in the United States compared to other countries.

This analysis helps new content creators understand the significance of geographic differences in viewer preferences, the potential impact of a popular video in a certain category, and the competitive landscape.

Question 2: Assess BTS's Popularity Across Various Countries.

K-pop is one of the world's largest music industries, with BTS establishing itself as a legendary influence on worldwide pop music (MITCHELL, 2022). In our analysis, we attempt to quantify the popularity of well-known boybands such as BTS in different locations. To accomplish this, we count the number of separate videos using the word "BTS" in their names.

	COUNTRY	CT
1	KR	331
2	RU	230
3	US	179
4	CA	173
5	MX	164
6	DE	162
7	JP	152
8	IN	149
9	GB	145
10	BR	116
11	FR	108

Figure 11: BTS video in each country

Surprisingly, the dataset contains at least 100 videos with the word "BTS" in the title, highlighting the legendary boyband's strong popularity. South Korea has the most significant number of such videos, followed by Russia in second place and the United States in third place.

This also highlights the importance of geographical variations in BTS-related material, which can help content makers modify their strategy to target various regions.

Question 3: Analyse the Most Viewed Videos for Each Country and Month and the Likes Ratio.

Our analysis shows surprising conclusions after considering the country-specific likes ratios, measured as the proportion of likes to the view count, and considering the most recent year inside the dataset, 2022.

Among the most viewed videos, the "People & Blogs" category in the GB (Great Britain) region has the most excellent likes ratio. One outstanding example in this area received over 49 million views and an exceptional likes ratio of 5.82.

	COUNTRY	YEAR_MONTH	TITLE	CHANNELTITLE ⋯	CATEGORY_TITLE	VIEW_COUNT	LIKES_RATIO
188	BR	2022-01-01	The Call │ Season 2022 Cinematic - League	League of Legends	Gaming	51,873,027	1.17
189	CA	2022-01-01	DIY hands cast #Shorts	Wood Mood	Howto & Style	85,006,005	3.30
190	DE	2022-01-01	Balloon Pop Racing Is INTENSE!! 🔴	How Ridiculous	Sports	107,927,053	3.67
191	FR	2022-01-01	We Don't Talk About Bruno (From Encanto)	DisneyMusicVEVO	Music	33,862,057	1.96
192	GB	2022-01-01	Diver Cracks Egg at 45 ft Deep #Shorts	Shanger Danger	People & Blogs	47,397,141	5.82
193	IN	2022-01-01	Ghareeb Baap aur Khilone Wala ‖ Emotiona	Peep Peep	Entertainment	87,847,902	2.66
194	JP	2022-01-01	Kep1er 케플러 │ 'WA DA DA' M/V	Kep1er	People & Blogs	34,418,748	3.85
195	KR	2022-01-01	The Call │ Season 2022 Cinematic - League	League of Legends	Gaming	90,978,675	0.69
196	MX	2022-01-01	The Call │ Season 2022 Cinematic - League	League of Legends	Gaming	68,204,276	0.90
197	RU	2022-01-01	Balloon Pop Racing Is INTENSE!! 🔴	How Ridiculous	Sports	91,248,098	3.82
198	US	2022-01-01	He went against a pro and came up clutch	SportsNation	Sports	42,093,658	4.36

Figure 12: Most viewed video in the recent year with likes ratio.

However, it is worth noting that in another country, Brazil (BR), videos have received even more views in the "Gaming" category, topping 51 million. Despite having more views, Brazil's videos had a lower likes-to-view ratio of 1.17.

This information gives content creators the knowledge they need to begin their YouTube journey, ensuring that their chosen categories fit with audience interests while accommodating regional variances.

Question 4: Determine each country's most popular video category and its share of total video content.

According to the data, "Entertainment" appeals to a broad audience and consistently accounts for the highest percentage of videos out of all videos in each country, making it the most popular video category in most of those countries.

However, Russia has an intriguing exception, where the "People & Blogs" category has surpassed "Entertainment." While this category may be a safe pick for most people, examining categories such as "**People & Blogs**" in popular places can be a strategic move to acquire a particular audience.

	COUNTRY	CATEGORY_TITLE	...	TOTAL_CATEGORY_VIDEOS	TOTAL_COUNTRY_VIDEOS	PERCENTAGE
1	BR	Entertainment		4,293	16,378	26.21
2	CA	Entertainment		4,313	20,817	20.72
3	DE	Entertainment		6,679	25,320	26.38
4	FR	Entertainment		5,297	22,110	23.96
5	GB	Entertainment		4,511	20,482	22.02
6	IN	Entertainment		12,839	29,456	43.59
7	JP	Entertainment		4,945	14,831	33.34
8	KR	Entertainment		4,625	13,473	34.33
9	MX	Entertainment		3,628	15,361	23.62
10	RU	People & Blogs		10,400	63,901	16.28
11	US	Entertainment		3,812	19,142	19.91

Figure 13: The most popular category and its share.

Question 5: Identify the channel title that has produced the most distinct videos and calculate the total number of videos produced by that channel.

The "Colors TV" channel has created an outstanding 805 distinct videos. This indicates that they have successfully connected with their viewers. This stresses the importance of our new channel implementing a content strategy that allows frequent video uploads and connects with users to build a devoted following.

4. LAUNCH A NEW CHANNEL

Based on the insights gained, our strategy will focus on creating diverse content that caters to regional preferences, emphasising targeting popular categories in specific regions. We will implement a content strategy that ensures consistent and frequent video uploads to foster user engagement and build a dedicated following.

4.1 What is each country's most popular video category, excluding Music and Entertainment?

The strategic choice to launch a new YouTube channel should be to delve into the 'People & Blogs' category. This decision is rooted in data-driven findings that reveal 'People & Blogs' to consistently rank as the top category in six out of eleven countries, showcasing its universal appeal. This makes it a strong candidate for a new channel.

Furthermore, the "People & Blogs" category is highly popular, with videos receiving over 900,000 views in Brazil and over 2.3 million in India. Choosing this category ensures not just higher viewing but also increased audience interaction.

	COUNTRY	CATEGORY_TITLE	HIGHEST_VIEWS	...	AVG_VIEWS	TOTAL_VIEW_COUNT
1	BR	People & Blogs	89,688,590		939,117.36	13,756,191,042
2	CA	Gaming	73,728,043		2,070,150.61	38,401,293,785
3	DE	People & Blogs	93,336,756		1,674,861.08	20,431,630,297
4	FR	Sports	32,695,020		743621.00	8,173,882,046
5	GB	Gaming	72,681,293		2,080,557.51	33,409,592,539
6	IN	People & Blogs	141,191,928		2,324,643.29	31,073,506,920
7	JP	People & Blogs	35,541,885		666,080.15	10,242,980,487
8	KR	People & Blogs	34,418,748		878,079.37	13,927,216,898
9	MX	Gaming	68,204,276		1,753,579.19	23,499,714,721
10	RU	People & Blogs	48,914,469		462,899.23	8,266,917,414
11	US	Gaming	73,728,043		2,171,759.59	40,829,080,273

Figure 14: Category with the highest total view count in each country

4.2 The likes ratio in the "People & Blogs" category

An in-depth analysis shows that the 'People & Blogs' category has huge potential for success in every country. According to the information, this category consistently ranks among the top four in terms of view count across all nations, demonstrating its worldwide popularity.

	COUNTRY	CATEGORY_TITLE	VIDEO_COUNT	LIKES_RATIO	LIKES_RANK	CATEGORY_RANK
1	BR	People & Blogs	14,648	9.24	4	1
2	CA	People & Blogs	9,926	5.2	6	2
3	DE	People & Blogs	12,199	4.97	7	1
4	FR	People & Blogs	5,573	8.15	3	4
5	GB	People & Blogs	10,106	5.03	6	2
6	IN	People & Blogs	13,367	5.58	6	1
7	JP	People & Blogs	15,378	3.56	4	1
8	KR	People & Blogs	15,861	3.74	2	1
9	MX	People & Blogs	13,899	8.28	3	2
10	RU	People & Blogs	17,859	7.63	1	1
11	US	People & Blogs	9,493	5.83	5	3

Figure 15: Likes ratio rank in the category.

Furthermore, "People & Blogs" frequently scores in the top 7 like ranks, highlighting its interaction with viewers. These statistics significantly demonstrate the potential of introducing a new channel in the 'People & Blogs' category across borders, overcoming geographical boundaries to reach audiences globally.

5. CONCLUSION

The effortless integration of Azure Blob Storage with Snowflake enabled us to extract, store, and analyse massive volumes of data from many sources, maintaining data integrity and reliability. This method can reach a worldwide audience by crossing geographical borders. According to our statistics, the "People & Blogs" category constantly ranks among the top categories regarding view count across all nations. It also has a high likes-to-engagement ratio, making it an excellent choice for starting a new YouTube channel and following a content strategy that encourages regular video uploads and user involvement.

Finally, the data-driven insights offered in this study provide a great roadmap for future YouTube creators aiming to achieve trending status and engage with viewers worldwide. The collaboration between Azure Blob Storage and Snowflake was critical in our investigation, emphasising the necessity of using advanced data management and analytics tools to influence content production and channel strategy.

6. APPENDIX

- Data Dictionary

features	Data type	Description
1. id	varchar	Identifier
2. video_id	varchar	The ID assigned for video
3. title	varchar	Video's title
4. publishedAt	timestamp	When the video released
5. channelID	varchar	The ID assigned for channel
6. channelTitle	varchar	The name of the channel
7. categoryID	integer	The ID assigned for category
8. category_title	varchar	Title of the category
9. trending_date	date	Date when video started trending
10. view_count	integer	video views count
11. likes	varchar	likes to count for video
12. dislikes	varchar	dislikes count for video
13. comment_count	varchar	count the comment
14. comment_disable	boolean	comment disable
15. country	varchar	country of the video origin

Table 2: factors in the final table with its data type and description

7. REFERENCE

MITCHELL, A. (28 May 2022). *What's next for K-pop? A roundtable unpacks the genre's past, present and future.* GRAMMYS. https://www.grammy.com/news/future-of-kpop-roundtable-interview-twice-stray-kids-jakops-videos

Youtube trending video dataset (Updated daily). (2023.). Retrieved 4 September 2023, from https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset