Introduction  
Azure

Microsoft's public cloud computing platform is called Microsoft Azure, formerly known as Windows Azure. Numerous cloud services are offered by it, such as computation, analytics, storage, and networking. These services are available for users to choose from while creating and scaling new apps or using the public cloud to operate already existing applications.

Customers that have an Azure subscription get access to all of the services offered through the Azure site. These services allow subscribers to build cloud-based resources like databases and virtual machines (VM’s). In this assessment task we use Azure as a cloud storage wherein we dump all the data, which comprises of CSV and Json files.

Snowflake

Snowflake makes it possible for data storage, processing, and analytical solutions that is far quicker, simpler to use, and far more versatile than existing ones. The Snowflake data platform is not based on any "big data" software frameworks like Hadoop or current database technologies. Instead, Snowflake blends a brand-new SQL query engine with a cutting-edge architecture specifically created for the cloud. Snowflake offers all of the features and functionality of an enterprise analytic database to the user, in addition to a large number of extra features and distinctive capabilities. Using SQL queries we perform Data cleaning and analysis on our dataset.

Data Understanding

YouTube is a well-known platform for streaming videos around the world. It is in charge of keeping track of the most popular trending video on its platform across various nations. To gauge how viewers interacted with the video, a variety of data including views, trending date, posting date, likes, dislikes, comment availability, etc. were recorded. The dataset includes all of these features that were taken from the YouTube database for the nations of India, Great Britain, Canada, France, Mexico, Japan, USA, Germany, Russia, Brazil, and South Korea. The dataset is divided into two categories: Eleven files with all the main video features and Eleven JSON files with the regions, categoryId, and category title in them.

Key Features of CSV files are as follows:

1. Video\_id - The id given by YouTube to a particular video
2. Title - The Title of the Video
3. publishedAt - The Date when the video was Published/Uploaded by the creator
4. channelId - The ID given by YouTube to the Channel
5. channelTitle - The title/name of the YouTube channel.
6. categoryId- ID assigned to a particular category by YouTube.
7. trending\_date - Date when video was trending on YouTube
8. view\_count - views on the YouTube video
9. likes - Likes on the video
10. dislikes - Dislikes on the video
11. comment\_count - number of comments on a particular video
12. comments\_disabled - to check comments are enabled or no

Whereas, Key features of Json files are their categoryIds, and the titles of the respective categories. Note: categoryId is not the same for different countries.

Part 01 - Data Ingestion

After the data i.e. CSV and JSON files were dumped inside a Azures container `bdeat1`. We then go ahead and setup storage integration between Azure and Snowflake. In order to achieve this, we first create a database and warehouse on snowflake called `bdeat1`. Then we setup storage integration between Azure and snowflake using Azure’s Directory ID and the Blob link. In Azure, we have to go in access control to assign role to the appropriate Multi-tenant app names.

In order to do a one-off dump of JSON data into snowflake, we create a snowpipe event in Azure. For which again we had to go in access control to assign role to the appropriate Multi-tenant app names.

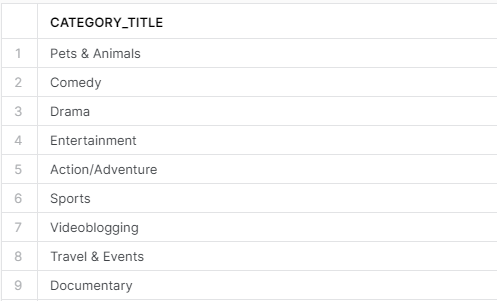
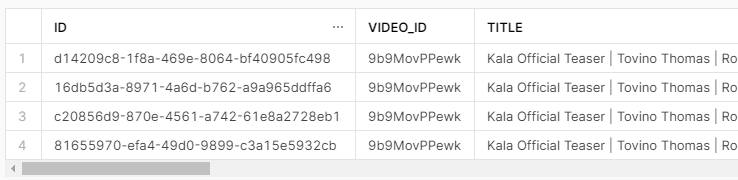
Now, the data can be loaded into Snowflake as external tables, where various operations on the datasets can be carried out. The JSON files were imported into a different external table and the CSV files were put in a different external table. The JSON files' important features were extracted, and two external tables i.e CSV and JSON, were combined with the following columns:

1. ID - ID assigned to every row
2. Video\_id – ID of the video assigned by YouTube
3. Title – Videos Title
4. publishedAt – The date the video was published or uploaded by the creator
5. ChannelID – ID given to the channel by YouTube
6. CategoryID - The ID of the category given by YouTube, we used CategoryID to merge CSV and JSON external tables
7. Category\_Title - The name of the category to which the video belongs
8. Trending\_date – The date when the video was trending on YouTube
9. View\_Count – Number of views the video has received
10. Likes - Number of likes the video has received
11. Dislikes - Number of dislikes the video has received
12. Comment\_Count - Number of Comments on the video
13. Comments\_Disabled - to check comments are enabled or no
14. Country - The name of the country from where the video was uploaded

Upon the merged table we will now perform Data cleaning and analysis.

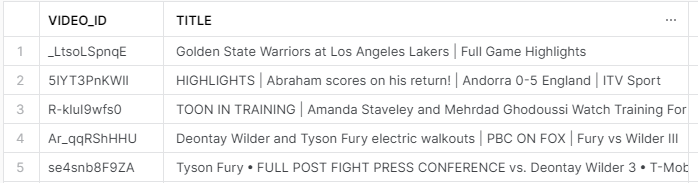
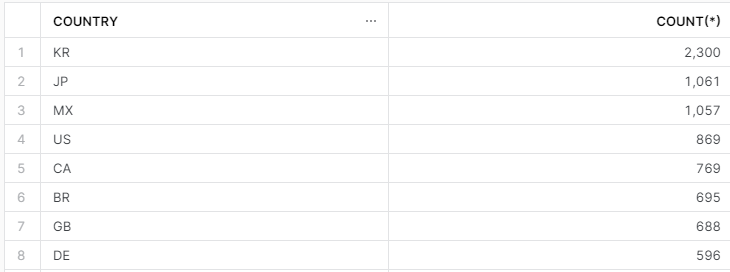
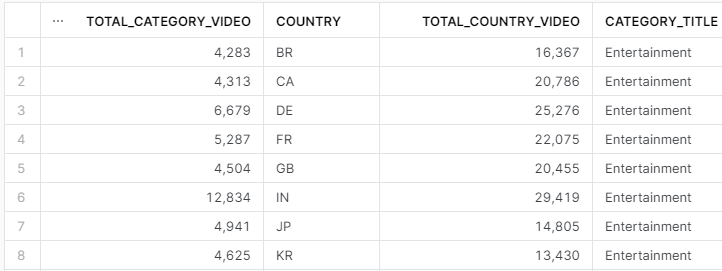
Part 02 - Data Cleaning

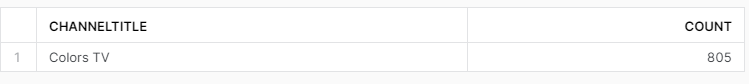
In order to clean the data, we did the following:

1. We tried to spot duplicates in the category\_title column from the JSON external table, table\_youtube\_category by using the function having count(). Here are some:  
   
2. In the external table holding all the JSON files, we tried to identify which category\_title is listed only in one country.  
     
   Category\_Title that appeared only in one country was **Nonprofit & Activism**
3. We then tried to identify what is the category\_id of the missing category\_title in the final merged table. Here are some:  
   
4. We then used the Update query on table\_youtube\_final, to replace the Null values of category\_title with category\_id
5. We then went on to identify which video doesn’t have channeltitle in the merged table\_youtube\_final. Here is a list of those videos.  
   
6. We furthermore, went on to delete rows from the merged table\_youtube\_final that had #NAME? as their Video ID  
   We deleted 58,296 such rows.
7. We identify duplicate values in video\_id, trending\_date and country from table\_youtube\_final.
8. We then delete those duplicate values. Over 3.5 million such rows were deleted.
9. We then count the number of records left in the final table after cleaning the dataset. There were 1119987 records left.

Part 03 - Data Analysis

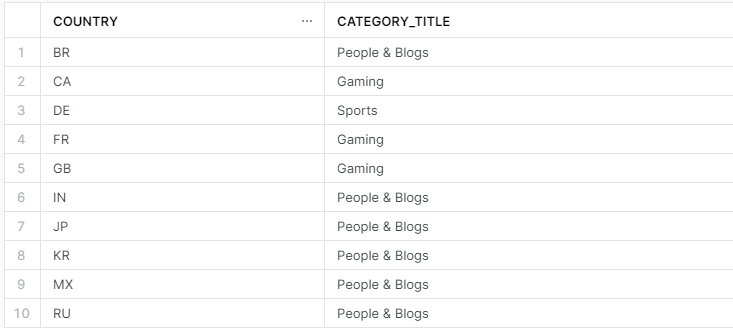
After cleaning the dataset thoroughly we do the following analysis

1. In the sports category, the top three videos for each nation were determined for trending date = 2021-10-17. The SQL function rank() was utilised for this. 
2. The number of videos using the word "BTS" were then counted for each nation. The reult were then put into descending order.  
   Korea had the most videos using the word BTS, followed by Japan and then Mexico3. 
3. Each month's most popular video for each nation was calculated. The results were arranged according to the country, month, and year. Additionally, a ratio was generated to determine the number of viewers and likers of the video. Here are the Results  
   
4. We attempted to identify the category with the most distinctive videos and its percentage (upto 2 decimals) of the total number of distinct videos in that country. The results were then arranged by category name and country.  
   Here are the Results for this analysis:  
   
5. In our final analysis we tried spot which channel had the most distinct videos.  
   The channel Colors TV had the most, which stood at 805 distinct videos.



Part 04

If I was to launch a YouTube channel, our analysis would be very crucial to help me determine which popular category should my videos target, in order to manifest quick growth. So, in order to answer this question, The video category with the most views were selected. It’s known that the most popular list on YouTube usually includes the video with the most views. Hence, this metric was therefore taken into account. Videos belonging to the music or entertainment category were ignored. It was determined that 'People and Blogs' received the most views. Gaming came in Second.



Ergo, in order to make it to the trending page, I should release Videos in the People & Blogs category, Unless if I am uploading video from France or Canada then I should aim my videos for the Gaming category. As seen in the above table each country has their own Interests