Data Lakehouse Project Summary

UTS 94693 Big Data Engineering

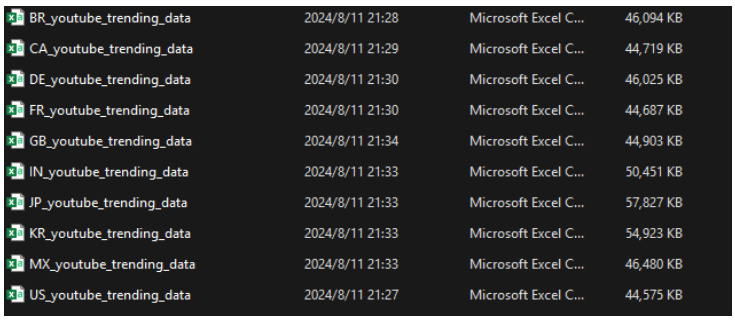
Sibo Zhang SID: 25520735

1 Project Overview

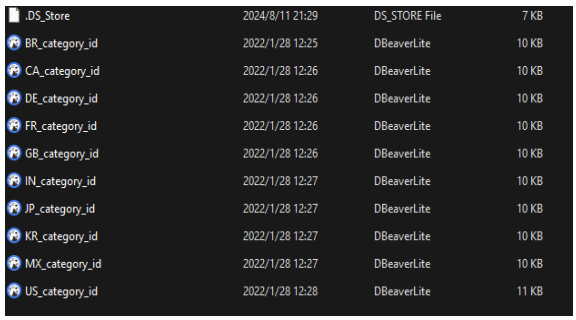
The goal of this project is to construct a Data Lakehouse using Microsoft Azure and Snowflake and use it for analyzing a dataset on YouTube Top Trending Videos.

1.1 Data Source

The dataset is from Kaggle website with daily updates(https://www.kaggle.com/rsrishav/youtube-trending-video-dataset). It consists of several years of data on daily trending Youtube videos in ten countries. It is made up of ten csv files and ten corresponding category json files as category ids vary with different countries. A screenshot of the dataset is as follows:



BR_youtube_trending_data	2024/8/11 21:28	Microsoft Excel C...	46,094 KB
CA_youtube_trending_data	2024/8/11 21:29	Microsoft Excel C...	44,719 KB
DE_youtube_trending_data	2024/8/11 21:30	Microsoft Excel C...	46,025 KB
FR_youtube_trending_data	2024/8/11 21:30	Microsoft Excel C...	44,687 KB
GB_youtube_trending_data	2024/8/11 21:34	Microsoft Excel C...	44,903 KB
IN_youtube_trending_data	2024/8/11 21:33	Microsoft Excel C...	50,451 KB
JP_youtube_trending_data	2024/8/11 21:33	Microsoft Excel C...	57,827 KB
KR_youtube_trending_data	2024/8/11 21:33	Microsoft Excel C...	54,923 KB
MX_youtube_trending_data	2024/8/11 21:33	Microsoft Excel C...	46,480 KB
US_youtube_trending_data	2024/8/11 21:27	Microsoft Excel C...	44,575 KB



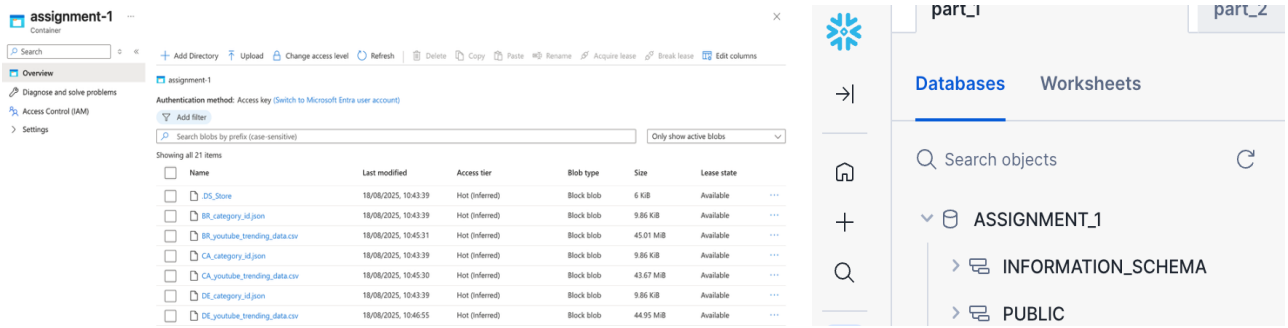
.DS_Store	2024/8/11 21:29	DS_STORE File	7 KB
BR_category_id	2022/1/28 12:25	DBeaverLite	10 KB
CA_category_id	2022/1/28 12:26	DBeaverLite	10 KB
DE_category_id	2022/1/28 12:26	DBeaverLite	10 KB
FR_category_id	2022/1/28 12:26	DBeaverLite	10 KB
GB_category_id	2022/1/28 12:26	DBeaverLite	10 KB
IN_category_id	2022/1/28 12:27	DBeaverLite	10 KB
JP_category_id	2022/1/28 12:27	DBeaverLite	10 KB
KR_category_id	2022/1/28 12:27	DBeaverLite	10 KB
MX_category_id	2022/1/28 12:27	DBeaverLite	10 KB
US_category_id	2022/1/28 12:28	DBeaverLite	11 KB

1.2 Data Engineering Lifecycle

The concept of Data Lakehouse and its implementation is at the core of a data engineering lifecycle. This project aims to set up a Data Lakehouse via Microsoft Azure cloud storage and Snowflake, providing data streams for analytics and other purposes. As such, it will follow the same general data engineering workflow: Data Ingestion, Data Transformation, and Data Servin. Since we will also be doing data analysis, the serving part will be omitted.

2 Data Ingestion

A Microsoft Azure and Snowflake account are needed for this project. After registration and uploading the dataset to Azure cloud storage and creating a database in Snowflake,



We need to establish a link between these two for loading data from Azure to Snowflake. This is done with the external stage feature in Snowflake.

```
CREATE OR REPLACE STAGE stage_assignment
    URL = 'azure://williamz25.blob.core.windows.net/assignment-1'
    CREDENTIALS = (AZURE_SAS_TOKEN = 'sv=2024-11-04&ss=bfqt&srt=sco&sp=rwdlacupiytfx&se=2025-08-25T03:31:43Z&st=2025-08-
20T19:16:43Z&spr=https&sig=P1Wf7Itd2ivLkbx0grUdKFCh%2BPVqBAL6ev1%2FXRsAahk%3D');
```

To get private access to the cloud storage space a SAS token from Azure portal is needed.

Next we create an external table, a feature used to query data from external stage as if they are stored in Snowflake. Note that as we have multiple files, bulking loading is used during this process. Two external tables were created, one for trending video datasets, and another for category_id json files. Finally we will merge these two tables on country and categoryid columns. The end table looked like follows:

COUNTRY	CATEGORYID	VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE	TRENDING_DATE	VIEW_COUN
FR	27	RmazDEQC4g8	Et si la Chine S'Écroulait ? (Anticipation)	2023-01-13	UCacgofzftbbkUhcftPvUzQg	AlterHis	2023-01-17	973
FR	23	0GhJ6oh3wQw	Coraline - Le Pire Éditeur au Monde	2023-01-13	UCfXXAQ-mp1uUcvSpvMcAAtv	LinksTheSun	2023-01-17	984
FR	10	QtZgGy8XZBI	Black M & Amir - Grandir (CLIP OFFICIEL)	2023-01-13	UCa_m88y-YP42wOdoYalk6og	Black M	2023-01-17	3577
FR	10	1dfAzhLkOto	DTF - G.A.G [Clip Officiel]	2023-01-12	UC1Ke8iR_NXio8H7GaB0S9pw	DTF	2023-01-17	4488
FR	24	ywE3B_wMWhQ	SFAM / HUBSIDE des employés craquent	2023-01-13	UCnC8KCcp2-DYfPY4_bKlc0A	Sandoz	2023-01-17	860
FR	1	PnIH2SaSrd4	AVATAR 2 le fameux bide ! Chrono-Critiqu	2023-01-13	UC3SLk50bvlivTtnFZqk-bHQ	benzaieTV	2023-01-17	1220
FR	20	uXNYs9YU7tI	J'ai utilisé un ITEM INTERDIT pour FINIR C	2023-01-13	UCfx-zJb69563vi-yLGhKwew	RAYTONIA	2023-01-17	609
FR	26	SwgA65BcFgY	J'ai acheté les 7 premiers produits recom	2023-01-13	UC38_-W5RqzyakXeb_5cpGCv	Laurette	2023-01-17	925
FR	27	nX4_R1UbzHg	La Ferme d'Alexandre de Prodealcenter -	2023-01-13	UCyEWw_87YOcFw9RoU00sW	Alexandre de Prodeal	2023-01-17	604
FR	25	igMD2Hb57mg	La Russie, les pays baltes et l'OTAN | ART	2023-01-12	UCwI-JbGNsojunnHbFAc0M4Q	ARTE	2023-01-17	4038

3 Data Transformation

Several null values existed in this dataset. In category_title column, the null values were replaced with the category_id value. Any row with video_id = "#Name?" were removed from the dataset.

There also exists records with the same video_id and trending_date in the same country, but with different metrics. This were viewed as duplicates and also removed from the dataset.

4 Data Analysis

Q1: For the top 3 most viewed video on trending_date 2024-04-01 in each country, two videos showed up in multiple countries:

DAGGER DUCHESS - New Tower Troop! (Official Music Video)
Confrontation - The Skibidi Saga 05
If my viewers break my secret rule, I ban them

Indonesia, Japan and Korea showed with different preferences compared to other countries.

Q2:
Korea, US ranked top on the number of distinct videos with BTS/bts in the title.

	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

Q3:

The most viewed video from Jan to April in 2024 were mostly produced by the famous vlogger MrBeast. The rest were largely gaming videos.

Q4:

The most viewed video category in 2022 in different countries is as follows:

	COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
1	BR	Entertainment	2670	11141	23.97
2	DE	Entertainment	3319	13473	24.63
3	FR	Entertainment	3621	15164	23.88
4	GB	Entertainment	2543	12187	20.87
5	IN	Entertainment	8521	22680	37.57
6	JP	Entertainment	2585	8119	31.84
7	KR	Entertainment	2360	6722	35.11
8	MX	Entertainment	1963	8303	23.64
9	CA	Gaming	3353	14581	23.00
10	US	Gaming	3119	13768	22.65

Q5:

Vijay Television is the channel with most distinct videos(2049) at the time recorded by the dataset.

5 Business Question

Let's start with the most basic strategy: betting on the category with the largest number of trendy videos.

	CATEGORY_TITLE	TOTAL_VIDEO
1	Sports	43323
2	People & Blogs	42876
3	Gaming	41615

Obviously sports is the hottest category according to data, followed by People & Blogs by a small margin. Next we can dig into that number by checking how concentrated it is. The rationale is that a small number of channels can produce most of the trendy videos within a category.

	CATEGORY_TITLE	AVG_VIDEO_PER_CATEGORY	MEDIAN_VIDEO_PER_CATEGORY	AVG_MEDIAN_DIFF
1	Nonprofits & Activism	5.2	1.0	4.2
2	People & Blogs	8.8	3.0	5.8
3	Pets & Animals	7.8	2.0	5.8

We can use the difference between the average and median number of trendy videos per channel to grasp the "winner-take-all" effect in each category. Unsurprisingly, categories that rank high in the previous chart rank turn out to be highly concentrated.

Channels like ESPN, Genshin Impact take a large slice of the pie. The notable exception is "People & Blogs" Category, ranking second as the hottest and fairest video category. Therefore a good strategy to enter the scene would be with this category to ensure a higher success rate at yielding trendy videos.

The intuition behind this strategy is that trending videos in this category are typically vlogs produced by single individuals/small studios with low costs. It is more accessible to most content creators compared to videos on sports or gaming events.

COUNTRY	CATEGORY_TITLE	RANK_BY_TOTAL_VIDEO
CA	People & Blogs	3
MX	People & Blogs	2
US	People & Blogs	3
DE	People & Blogs	3
JP	People & Blogs	2
IN	People & Blogs	1
KR	People & Blogs	1
FR	People & Blogs	4
BR	People & Blogs	3
GB	People & Blogs	3

A natural question is to ask if this strategy would be effective across all ten countries available in the dataset. Graph above shows the ranking of "People & Blogs" category by total distinct videos in each nation. In general

the strategy is stable across each nation, with lowest ranking appeared in French and highest one in Korea and Indonesia.

6 Bugs&Fixes

Overall the part with most bugs appeared in the data ingestion phase. The most notable one is that when bulk loading, snowflake automatically recognized all commas in the file as delimiters. But in dataset there were records with video titles containing commas. Functionality FIELD_OPTIONALLY_ENCLOSED_BY came to the rescue and solve the problem.

Another one appeared in the nested structure of json files when bulk loading. The solution is to combine the LATERAL_FLATTEN function with the subselect technique in SQL to a nested json structure.

Another issue faced when finding the most viewed videos in each category was the inability to put a window function inside another window function. But we need a window function for rank like RANK(() and another function for sum like COUNT DISTINCT. Again subqueries are needed to solve this problem.