Big Data Engineering

Assignment 1: Data Lakehouse with Snowflake

Aim:

The goal of this assignment is to analyse a dataset (made of CSVs and Jsons files) by using a Data Lakehouse with Snowflake. You will have to upload the data on a cloud storage, ingest the data into the Data Lakehouse, perform data transformation and finally analyse it.

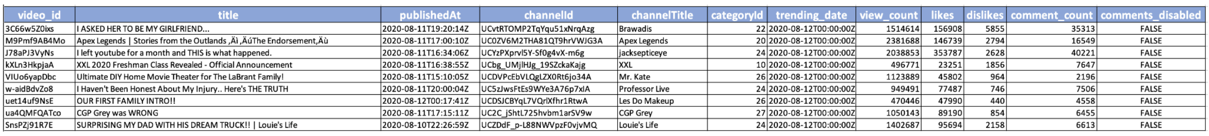
Introduction to the dataset

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users' interactions (e.g. number of views, shares, comments and likes).

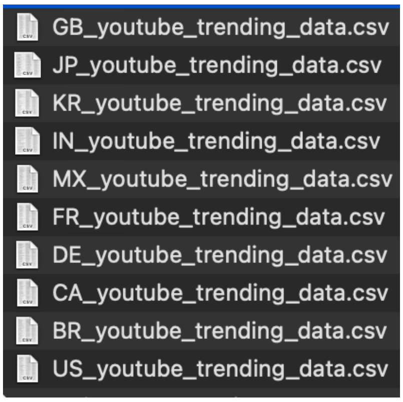
A dataset with a daily record of the top trending YouTube videos has been extracted through the Youtube API and made available on the Kaggle (https://www.kaggle.com/rsrishav/youtube-trending-video-dataset)

This dataset includes several months (from 2020-08-12 to 2024-04-15) of data of daily trending YouTube videos. Data is included for the IN, US, GB, DE, CA, FR, BR, MX, KR, and JP regions (India, USA, Great Britain, Germany, Canada, France, Brazil, Mexico, South Korea, and Japan respectively), with up to 200 listed trending videos per day.

Each region’s data is in a separate file. Data includes the video title, channel title, published time, views, likes and dislikes and comment count:

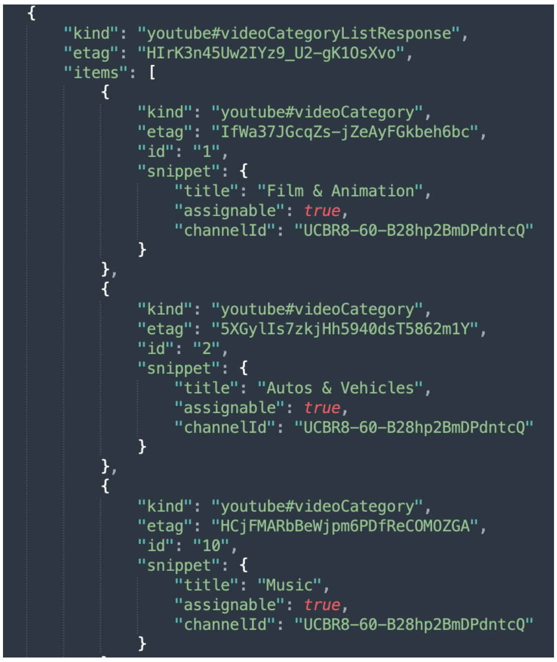


video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date	view_count	likes	dislikes	comment_count	comments_disabled
3C66w5Z0ixs	I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11T19:20:14Z	UCvtRTOMP2TqYqu51xNrqAzg	Brawadis	22	2020-08-12T00:00:00Z	1514614	156908	5855	35313	FALSE
M9Pmf9AB4Mo	Apex Legends | Stories from the Outlands ,Äì ,ÄúThe Endorsement,Äù	2020-08-11T17:00:10Z	UC0ZV6M2THA81QT9hrVWJG3A	Apex Legends	20	2020-08-12T00:00:00Z	2381688	146739	2794	16549	FALSE
J78aPJ3VyNs	I left youtube for a month and THIS is what happened.	2020-08-11T16:34:06Z	UCYzPXprvl5Y-Sf0g4vX-m6g	jacksepticeye	24	2020-08-12T00:00:00Z	2038853	353787	2628	40221	FALSE
kXLn3HkpjaA	XXL 2020 Freshman Class Revealed - Official Announcement	2020-08-11T16:38:55Z	UCbg_UMjlHJg_19SZckaKajg	XXL	10	2020-08-12T00:00:00Z	496771	23251	1856	7647	FALSE
VIUo6yapDbc	Ultimate DIY Home Movie Theater for The LaBrant Family!	2020-08-11T15:10:05Z	UCDVPcEbVLQgLZX0Rt6jo34A	Mr. Kate	26	2020-08-12T00:00:00Z	1123889	45802	964	2196	FALSE
w-aidBdvZo8	I Haven't Been Honest About My Injury.. Here's THE TRUTH	2020-08-11T20:00:04Z	UC5zJwsFtEs9WYe3A76p7xlA	Professor Live	24	2020-08-12T00:00:00Z	949491	77487	746	7506	FALSE
uet14uf9NsE	OUR FIRST FAMILY INTRO!!	2020-08-12T00:17:41Z	UCDSJCBYqL7VQrlXfhr1RtwA	Les Do Makeup	26	2020-08-12T00:00:00Z	470446	47990	440	4558	FALSE
ua4QMFQATco	CGP Grey was WRONG	2020-08-11T17:15:11Z	UC2C_jShtL725hvbm1arSV9w	CGP Grey	27	2020-08-12T00:00:00Z	1050143	89190	854	6455	FALSE
SnsPZj91R7E	SURPRISING MY DAD WITH HIS DREAM TRUCK!! | Louie's Life	2020-08-10T22:26:59Z	UCZDdF_p-L88NWVpzF0vjvMQ	Louie's Life	24	2020-08-12T00:00:00Z	1402687	95694	2158	6613	FALSE

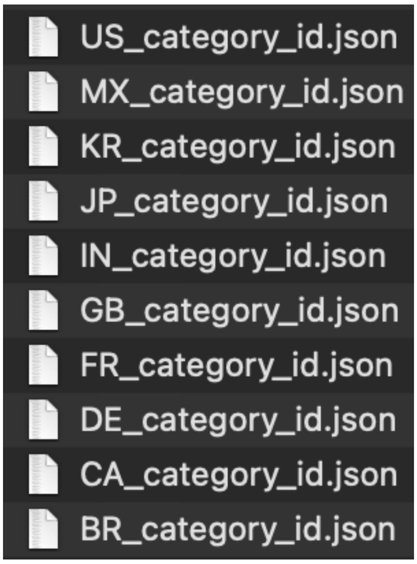


GB_youtube_trending_data.csv
JP_youtube_trending_data.csv
KR_youtube_trending_data.csv
IN_youtube_trending_data.csv
MX_youtube_trending_data.csv
FR_youtube_trending_data.csv
DE_youtube_trending_data.csv
CA_youtube_trending_data.csv
BR_youtube_trending_data.csv
US_youtube_trending_data.csv

The data also includes a category_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the 10 regions in the dataset.



```
{
    "kind": "youtube#videoCategoryListResponse",
    "etag": "HIrK3n45Uw2IYz9_U2-gK1OsXvo",
    "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"
            }
        },
        {
            "kind": "youtube#videoCategory",
            "etag": "HCjFMARbBeWjpm6PDfReCOMOZGA",
            "id": "10",
            "snippet": {
                "title": "Music",
                "assignable": true,
                "channelId": "UCBR8-60-B28hp2BmDPdntcQ"
            }
```



US_category_id.json
MX_category_id.json
KR_category_id.json
JP_category_id.json
IN_category_id.json
GB_category_id.json
FR_category_id.json
DE_category_id.json
CA_category_id.json
BR_category_id.json

Tasks:

You will need your cloud storage account on Microsoft Azure and your Snowflake account which were set up for the lab 2.

Your tasks will be:

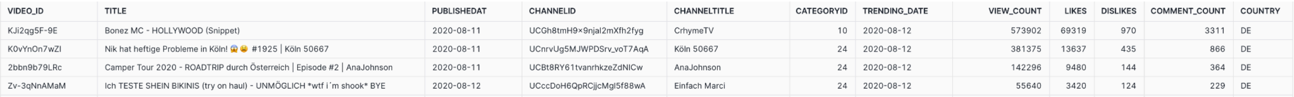
PART 1: Data Ingestion

Provide a sql file containing all the sql code used in Snowflake for part 1 and called it "**part_1.sql**":

1. Download the (compressed) dataset on:
   a. Trending data: https://drive.google.com/file/d/14xKzN4MEtCr1IZ_8w0JKwBTCjo-CBLIL/view?usp=sharing

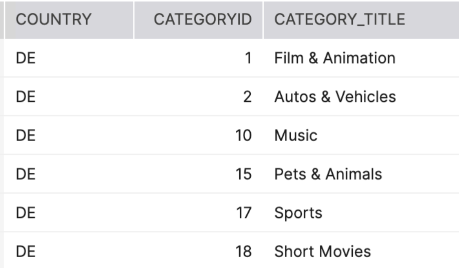
    b. Category data: https://drive.google.com/file/d/1uhkOwCCQK7LoER6tXZpsVbIfAr-CJomJ/view?usp=sharing

2. Upload the dataset in your storage account on Azure
3. On Snowflake:
    a. Create a database called: "***assignment_1"***
    b. Create a stage called "***stage_assignment"***, pointing to your azure storage
4. Ingest the data as external tables on Snowflake
    a. Create two external tables "***ex_table_youtube_trending"*** and ***"ex_table_youtube_category"*** *with the correct data type.*
5. Transfer the data from external tables into tables with the following columns:
    a. For trending data create a table called "***table_youtube_trending***" with:



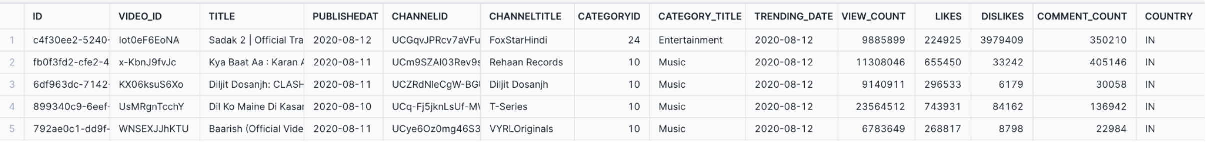
VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE	CATEGORYID	TRENDING_DATE	VIEW_COUNT	LIKES	DISLIKES	COMMENT_COUNT	COUNTRY
KJi2qg5F-9E	Bonez MC - HOLLYWOOD (Snippet)	2020-08-11	UCGh8tmH9×9njal2mXfh2fyg	CrhymeTV	10	2020-08-12	573902	69319	970	3311	DE
K0vYnOn7wZI	Nik hat heftige Probleme in Köln! 😱😩 #1925 | Köln 50667	2020-08-11	UCnrvUg5MJWPDSrv_voT7AqA	Köln 50667	24	2020-08-12	381375	13637	435	866	DE
2bbn9b79LRc	Camper Tour 2020 - ROADTRIP durch Österreich | Episode #2 | AnaJohnson	2020-08-11	UCBt8RY61tvanrhkzeZdNICw	AnaJohnson	24	2020-08-12	142296	9480	144	364	DE
Zv-3qNnAMaM	Ich TESTE SHEIN BIKINIS (try on haul) - UNMÖGLICH *wtf i´m shook* BYE	2020-08-12	UCccDoH6QpRCjjcMgl5f88wA	Einfach Marci	24	2020-08-12	55640	3420	124	229	DE

    b. For category data create a table called "***table_youtube_category***" with:



COUNTRY	CATEGORYID	CATEGORY_TITLE
DE	1	Film & Animation
DE	2	Autos & Vehicles
DE	10	Music
DE	15	Pets & Animals
DE	17	Sports
DE	18	Short Movies

6. Create a final table called "***table_youtube_final***" by combining "***table_youtube_trending***" and "***table_youtube_category***" on *country* and *categoryid* (**be careful to not lose any records**), while adding a new field called ***id*** by using the "UUID_STRING()" function :



	ID	VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE	CATEGORYID	CATEGORY_TITLE	TRENDING_DATE	VIEW_COUNT	LIKES	DISLIKES	COMMENT_COUNT	COUNTRY
1	c4f30ee2-5240-	lot0eF6EoNA	Sadak 2 | Official Tra	2020-08-12	UCGqvJPRcv7aVFu	FoxStarHindi	24	Entertainment	2020-08-12	9885899	224925	3979409	350210	IN
2	fb0f3fd2-cfe2-4	x-KbnJ9fvJc	Kya Baat Aa : Karan A	2020-08-11	UCm9SZAI03Rev9s	Rehaan Records	10	Music	2020-08-12	11308046	655450	33242	405146	IN
3	6df963dc-7142-	KX06ksuS6Xo	Diljit Dosanjh: CLASH	2020-08-11	UCZRdNleCgW-BGI	Diljit Dosanjh	10	Music	2020-08-12	9140911	296533	6179	30058	IN
4	899340c9-6eef-	UsMRgnTcchY	Dil Ko Maine Di Kasar	2020-08-10	UCq-Fj5jknLsUf-MV	T-Series	10	Music	2020-08-12	23564512	743931	84162	136942	IN
5	792ae0c1-dd9f-	WNSEXJJhKTU	Baarish (Official Vide	2020-08-11	UCye6Oz0mg46S3	VYRLOriginals	10	Music	2020-08-12	6783649	268817	8798	22984	IN

You should end up with **2,667,041** rows in ***table_youtube_final***

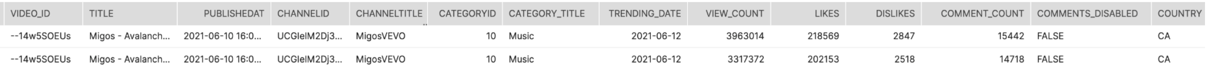
PART 2: Data Cleaning

Provide a sql file containing all the sql code used in Snowflake for part 2 and called it "**part_2.sql**" (add comments to separate each questions):

1. In "*table_youtube_category"* which *category_title* has duplicates if we don't take into account the *categoryid (return only a single row)*?
2. In "*table_youtube_category"* which *category_title* only appears in one country?
3. In "*table_youtube_final*", what is the *categoryid* of the missing *category_title*?

4. Update the *table_youtube_final* to replace the NULL values in *category_title* with the answer from the previous question.
5. In “*table_youtube_final*”, which video doesn’t have a *channeltitle (return only the title)*?
6. Delete from “*table_youtube_final*“, any record with *video_id* = “#NAME?”

The “*table_youtube_final*“ contains duplicates with the same *video_id*, *country* and *trending_date* however their metrics (likes, dislikes, etc..) can be different. E.g:



VIDEO_ID	TITLE	PUBLISHEDAT	CHANNELID	CHANNELTITLE	CATEGORYID	CATEGORY_TITLE	TRENDING_DATE	VIEW_COUNT	LIKES	DISLIKES	COMMENT_COUNT	COMMENTS_DISABLED	COUNTRY
--14w5SOEUs	Migos - Avalanch...	2021-06-10 16:0...	UCGIeIM2Dj3...	MigosVEVO	10	Music	2021-06-12	3963014	218569	2847	15442	FALSE	CA
--14w5SOEUs	Migos - Avalanch...	2021-06-10 16:0...	UCGIeIM2Dj3...	MigosVEVO	10	Music	2021-06-12	3317372	202153	2518	14718	FALSE	CA

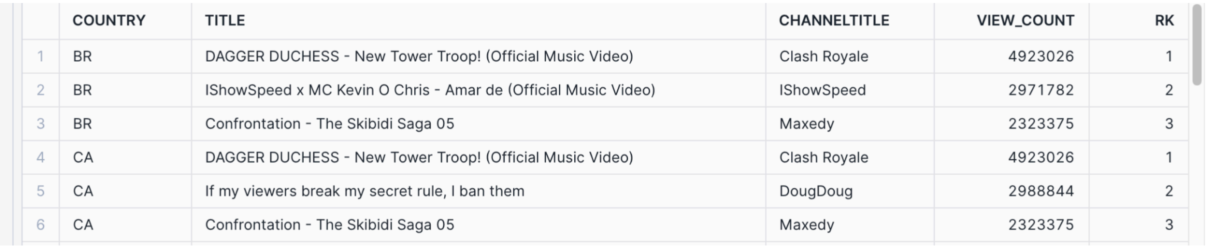
We can assume that the highest number of *view_count* will be the record to keep when we have duplicates.

7. Create a new table called “*table_youtube_duplicates”* containing only the “bad” duplicates by using the *row_number()* function.
8. Delete the duplicates in “*table_youtube_final*“ by using “*table_youtube_duplicates”*.
9. Count the number of rows in “*table_youtube_final*“ and check that it is equal to **2,597,494 rows.**

PART 3: Data Analysis

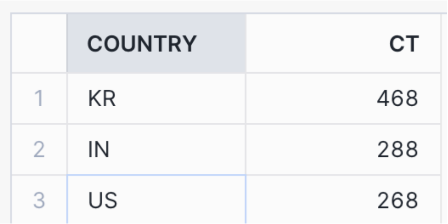
Provide a sql file containing the sql code used:

1. What are the 3 most viewed videos for each country in the Gaming category for the *trending_date* = ‘’2024-04-01 . Order the result by *country* and the *rank*, e.g:



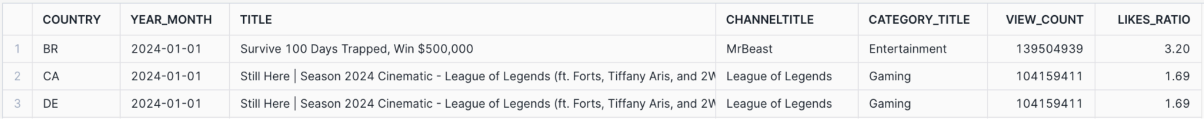
	COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
1	BR	DAGGER DUCHESS - New Tower Troop! (Official Music Video)	Clash Royale	4923026	1
2	BR	IShowSpeed x MC Kevin O Chris - Amar de (Official Music Video)	IShowSpeed	2971782	2
3	BR	Confrontation - The Skibidi Saga 05	Maxedy	2323375	3
4	CA	DAGGER DUCHESS - New Tower 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

2. For each country, count the number of **distinct** video with a title containing the word “BTS” (case insensitive) and order the result by count in a descending order, e.g:



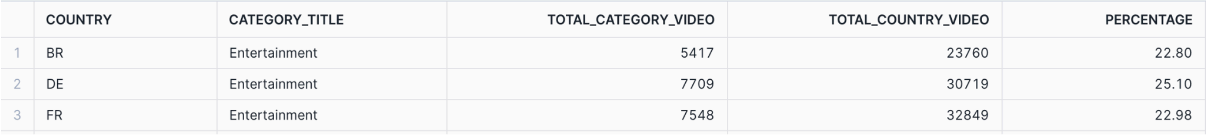
	COUNTRY	CT
1	KR	468
2	IN	288
3	US	268

3. For each *country*, *year* and *month* (in a single column) and only for the year 2024, which video is the most viewed and what is its likes_ratio (defined as the percentage of likes against view_count) truncated to 2 decimals. Order the result by *year_month* and *country*. The output should like this:



	COUNTRY	YEAR_MONTH	TITLE	CHANNELTITLE	CATEGORY_TITLE	VIEW_COUNT	LIKES_RATIO
1	BR	2024-01-01	Survive 100 Days Trapped, Win $500,000	MrBeast	Entertainment	139504939	3.20
2	CA	2024-01-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2W	League of Legends	Gaming	104159411	1.69
3	DE	2024-01-01	Still Here | Season 2024 Cinematic - League of Legends (ft. Forts, Tiffany Aris, and 2W	League of Legends	Gaming	104159411	1.69

4. For each *country*, which *category_title* has the most **distinct** videos and what is its percentage (2 decimals) out of the total **distinct** number of videos of that *country*? Only look at the data from 2022. Order the result by category_title and *country*. The output should like this:



	COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
1	BR	Entertainment	5417	23760	22.80
2	DE	Entertainment	7709	30719	25.10
3	FR	Entertainment	7548	32849	22.98

5. Which *channeltitle* has produced the most **distinct** videos and what is this number ?

PART 4: Business Question

Provide a single sql file containing all the queries used:

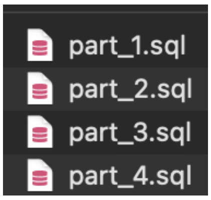
If you were to launch a new Youtube channel tomorrow, which category (excluding “Music” and “Entertainment”) of video will you be trying to create to have them appear in the top trend of Youtube ? Will this strategy work in every country?

This is an individual assignment but each student will be marked individually.

Deliverables:

Each student will have to submit

- SQL queries (.sql files) used for parts:



part_1.sql
part_2.sql
part_3.sql
part_4.sql

- A “handover” written report
- Any other relevant documents

The report should not exceed 2000 words (figures and tables are not counted).

Compress all deliverables into a single zip file and use the following file naming format for the submission:
Assignment_1_FirstName_LastName.zip

A good “handover” report should contained:

1. High-level view of your project.
2. Explanation for the different steps of your project.
3. Any issues/bugs you faced and how you solved them.

4. Answers to the different questions.
5. Relevant screenshots/images/diagrams/flows if necessary.

You can assume that the reader of your report will have a similar understanding and knowledge of any technical skills.

A good way to know if you have a good “handover” report is to ask one of your classmates/groupmates to read through it and see if he/she will be confident to “take over” your work.

Example 1
Example 2

Assessment Criteria:

- Quality of code.
- Justification of data transformation, data formats, data storage and accuracy of results with evidence supporting claims.
- Quality of findings and recommendations for business questions.
- Clarity and quality of written report.

Criteria Details and weights:

Criteria	Further Details
Quality of code	1. Code can be executed without raising an error. 2. Code achieved the goal of the brief 3. Code is well commented.
Justification of any data processing (transformation, formats, storage, etc.)	1. High level explanation of each major step and decision. 2. Follows the good “handover” report guidelines
Accuracy of results with evidence supporting claims	1. Correct answers to the different questions (Part 2 and 3). 2. Answers output are in the same shape as the example (column name, column format).
Quality of findings and recommendations for business questions.	1. Correct answer to the business questions. 2. Relevant queries are provided to support the answer.
Clarity and quality of written report.	1. Complete and professionally formatted report (spelling, grammar, punctuation, layout). 2. Report is not exceeding the maximum length

This assignment will count **30%** of your final mark.

Due Date:

All assignments need to be submitted before the **due date (2nd September 2024)** on Canvas.

Late submission will be penalised 10 pts per day after the due date.