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

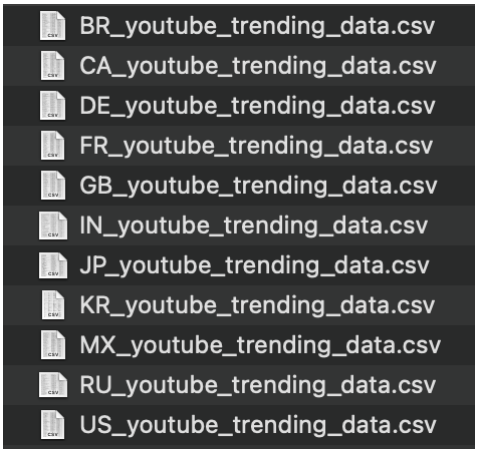
YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users' interactions (e.g. number of views, shares, comments 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 today) of data of daily trending YouTube videos. Data is included for the IN, US, GB, DE, CA, FR, RU, BR, MX, KR, and JP regions (India, USA, Great Britain, Germany, Canada, France, Russia, Brazil, Mexico, South Korea, and, Japan respectively), with up to 200 listed trending videos per day.

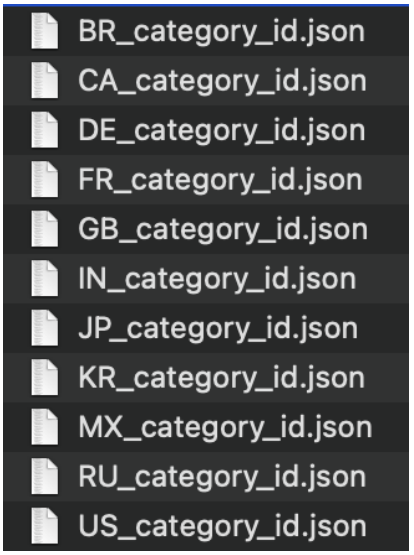
Each region’s data is in a separate file. Data includes the video title, channel title, published time, views, likes and dislikes and comment count:

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	UC0ZV6M2THA81QT9hrVWiG3A	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-aid8dvZo8	I Haven't Been Honest About My Injury... Here's THE TRUTH	2020-08-11T20:00:04Z	UCSzJwsFtEs9WYe3A76p7xIA	Professor Live	24	2020-08-12T00:00:00Z	949491	77487	746	7506	FALSE
uet14uf9NsE	OUR FIRST FAMILY INTRO!!	2020-08-12T00:17:41Z	UCDSJCBYqL7VQrXfhr1RtwA	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



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 11 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"
            }
```

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”:

part_1.sql

1. Download the (compressed) dataset on:
    a. Trending data: https://drive.google.com/file/d/1bsRxgSTXenOhKCjN3nSqmisy9aMokdeW/view?usp=sharing

    b. Category data: https://drive.google.com/file/d/13818ZbLMSpCNHR9CO3Ecty7iv_-HEHhx/view?usp=sharing

2. Upload the dataset in your storage account on Azure
3. Ingest the data as external tables on Snowflake
4. 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	CHANNE... ↓	CHANNELTITLE	CATEGORYID	TRENDING_DATE	VIEW_COUNT	LIKES	DISLIKES	COMMENT_COUNT	COMMENTS_DISABLED	COUNTRY
XRVzPg5suV4	ON COULE LA D...	2020-08-09 07:00:12.000	UCzxe_ob_s...	RACHELSTYLIS...	26	2020-08-12	200560	14338	196	743	FALSE	FR
Iv5chpKSeNY	JE RÉAGIS AU C...	2020-08-07 19:50:45.000	UCzxHfZ_q7...	LA CAGOULE	24	2020-08-12	118657	15278	1112	837	FALSE	FR
jUOZzhrOhXg	VACANCES ENT...	2020-08-09 10:00:20.000	UCzivjQrmkT...	Mlex Chloé	26	2020-08-12	68512	6403	210	102	FALSE	FR
QKNi9wLknXk	Résumé Juventu...	2020-08-09 07:32:46.000	UCzHCZXmq...	Olympique Lyon...	17	2020-08-12	314522	3189	273	531	FALSE	FR
1zpDvb8qPXo	FIFA 21 | NEWS -...	2020-08-10 08:00:01.000	UCyVHtLk13...	SD.	20	2020-08-12	35877	2857	40	338	FALSE	FR
hsm4poTWjMs	Cardi B - WAP fe...	2020-08-07 04:00:10.000	UCxMAbVFm...	Cardi B	10	2020-08-12	76805026	2820367	382583	270263	FALSE	FR
Fi4CK7VnjUo	le meilleur anniv...	2020-08-11 16:30:44.000	UCxDsG8fSV...	Esile	24	2020-08-12	104160	15715	64	671	FALSE	FR

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

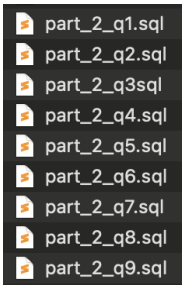
COUNTRY	CATEGORYID	CATEGORY_TITLE
DE	1	Film & Animation
DE	2	Autos & Vehicles
DE	10	Music
DE	15	Pets & Animals
DE	17	Sports
DE	18	Short Movies

5. 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	COMMENTS_DISABLED	COUNTRY
3c375779-f...	KJl2qg5F-9E	Bonez MC - ...	2020-08-11 ...	UCGh8tmH9...	CrhymeTV	10	Music	2020-08-12	573902	69319	970	3311	FALSE	DE
17ff6bf3-64...	K0vYnOn7wZI	Nik hat hefti...	2020-08-11 ...	UCnrvUg5M...	Köln 50667	24	Entertainment	2020-08-12	381375	13637	435	866	FALSE	DE
ce86d878-c...	2bbn9b79LRc	Camper Tou...	2020-08-11 ...	UCBt8RY61t...	AnaJohnson	24	Entertainment	2020-08-12	142296	9480	144	364	FALSE	DE
ca9679ab-4...	Zv-3qNnAM...	Ich TESTE S...	2020-08-12 ...	UCccDoH6Q...	Einfach Marci	24	Entertainment	2020-08-12	55640	3420	124	229	FALSE	DE
d76615f4-6...	7clgQLneouU	STATEMENT...	2020-08-11 ...	UC8E8eD7m...	Domo	24	Entertainment	2020-08-12	233899	25251	375	1051	FALSE	DE
ddc9b581-8...	g7vdQgRsTKc	Bayer unterli...	2020-08-10 ...	UCNxq-0KJ...	DAZN UEFA Eu...	17	Sports	2020-08-12	623938	12770	357	1514	FALSE	DE
43cff37c-62...	86gzh8jftSE	GEBRAUCH...	2020-08-11 ...	UCzH549Yl...	AlexiBexi	24	Entertainment	2020-08-12	249531	17199	304	812	FALSE	DE
d2906951-8...	YECPzRfksI4	Erkennst DU...	2020-08-11 ...	UCL5-tPmf_...	World Wide Wo...	24	Entertainment	2020-08-12	470201	43045	369	1719	FALSE	DE

PART 2: Data Cleaning

For each question provide a sql file containing the sql code used:



1. In "*table_youtube_category"* which *category_title* has duplicates if we don't take into account the *categoryid*?
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*?
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...	UCGIelM2Dj3...	MigosVEVO	10	Music	2021-06-12	3963014	218569	2847	15442	FALSE	CA
--14w5SOEUs	Migos - Avalanch...	2021-06-10 16:0...	UCGIelM2Dj3...	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 is it equal to **1,123,017 rows.**

PART 3: Data Analysis

For each question provide a sql file containing the sql code used **AND** a csv containing the output (you can use the snowflake export feature):

1. What are the 3 most viewed videos for each country in the "Sports" category for the *trending_date* = ''2021-10-17''. Order the result by *country* and the *rank*, e.g:

COUNTRY	TITLE	CHANNELTITLE	VIEW_COUNT	RK
BR	BRASIL 4 X 1 URUGUAI | MELHORES MOMENTOS | 12ª RODAD...	ge	4562725	1
BR	MAIS TRÊS GOLS DE CRISTIANO RONALDO! PORTUGAL 5 X 0 ...	TNT Sports Brasil	2053005	2
BR	♫ NEYMAR TÁ DE VOLTA!! E A DUPLA COM RAPHINHA DECOL...	FutParódias	814491	3
CA	Sore loser! An idiot! Tyson Fury reveals what was said between...	BT Sport Boxing	6913800	1
CA	World's Smallest TV | OT 30	Dude Perfect	6222811	2

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

COUNTRY	CT
KR	331
RU	230

3. For each *country*, *year* and *month* (in a single column), 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
BR	2020-08-01	BTS (방탄소년단) 'Dynamite'...	Big Hit Labels	Music	244507902	6.52
CA	2020-08-01	BTS (방탄소년단) 'Dynamite'...	Big Hit Labels	Music	232649205	6.76
DE	2020-08-01	BTS (방탄소년단) 'Dynamite'...	Big Hit Labels	Music	219110491	7.06
FR	2020-08-01	BTS (방탄소년단) 'Dynamite'...	Big Hit Labels	Music	232649205	6.76
GB	2020-08-01	BTS (방탄소년단) 'Dynamite'...	Big Hit Labels	Music	208581468	7.31
IN	2020-08-01	BTS (방탄소년단) 'Dynamite'...	Big Hit Labels	Music	253995993	6.34
JP	2020-08-01	BTS (방탄소년단) 'Dynamite'...	Big Hit Labels	Music	262319276	6.20

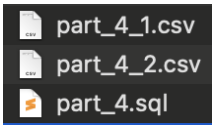
4. For each *country*, which *category_title* has the most **distinct** videos and what is its percentage (2 decimals) out of the total **distinct** number of videos of that *country*? Order the result by *country*. The output should like this:

COUNTRY	CATEGORY_TITLE	TOTAL_CATEGORY_VIDEO	TOTAL_COUNTRY_VIDEO	PERCENTAGE
BR	Entertainment	4,293	16,371	26.22
CA	Entertainment	4,313	20,807	20.73
DE	Entertainment	6,679	25,299	26.4
FR	Entertainment	5,297	22,096	23.97
GB	Entertainment	4,511	20,472	22.04

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 and one csv file per output, e.g:



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:
    - 1 file for part 1
    - 9 files for part 2
    - 5 files for part 3
    - 1 file for part 4
- CSV files which are the SQL queries output for parts:
    - 5 files for part 3
    - At least 1 file for part 4
- 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	Weight
Quality of code	1. Code can be executed without raising an error. 2. Code is well commented.	15%
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	20%
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).	40%
Quality of findings and recommendations for business questions.	1. Correct answers to the business questions. 2. Relevant outputs are provided to support answers.	15%
Clarity and quality of written report.	1. Complete and professionally formatted report (spelling, grammar, punctuation, layout). 2. Report is not exceeding the maximum length	10%

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

Due Date:

All assignments need to be submitted before the **due date (4th September 2022)** on Canvas. Penalties will be applied for late submission