**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.

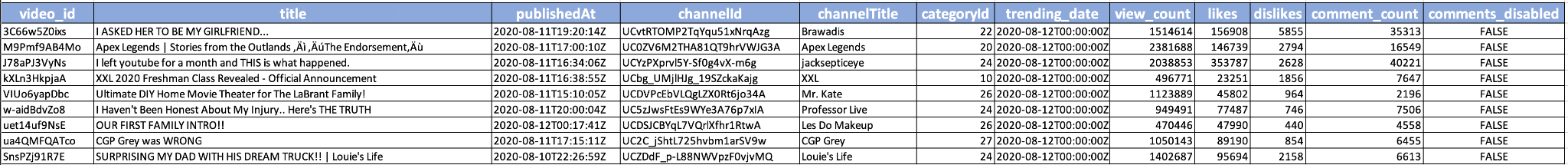
**Introduction to the dataset**

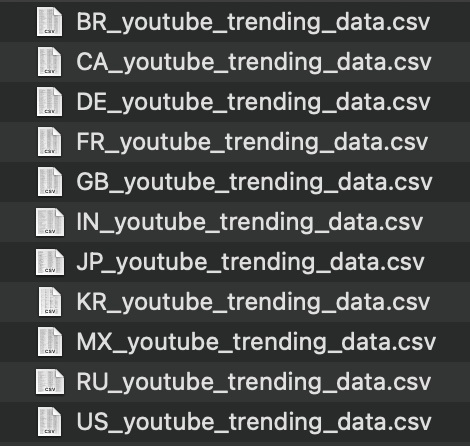
YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users' interactions (e.g. number of views, shares, comments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

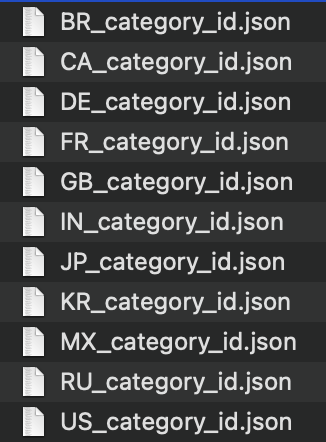
Each region’s data is in a separate file. Data includes the video title, channel title, published time, views, likes and dislikes and comment count:





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.





**Tasks:**

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

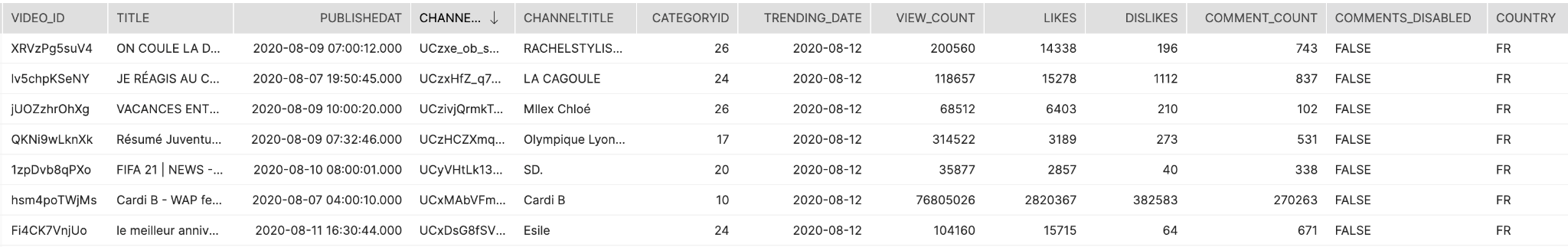
Your tasks will be:

**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”:

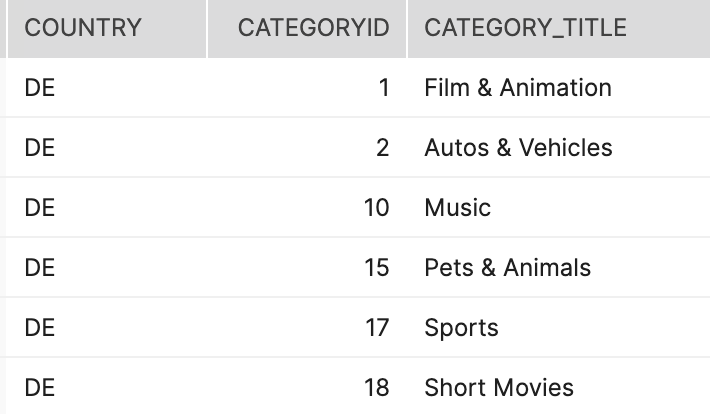


1. Download the (compressed) dataset on:
   1. Trending data: <https://drive.google.com/file/d/1IORoX-kbleMhEvwi0-_VrySvjieFj7Dt/view?usp=sharing>
   2. Category data: <https://drive.google.com/file/d/15n9LKvg6PbTN3nXXE1M7qrk30piWjLOs/view?usp=sharing>
2. Upload the dataset in your storage account on Azure
3. Ingest the data as external tables on Snowflake
4. Transfer the data from external tables into tables with the following columns:
   1. For trending data create a table called “*table\_youtube\_trending*” with:

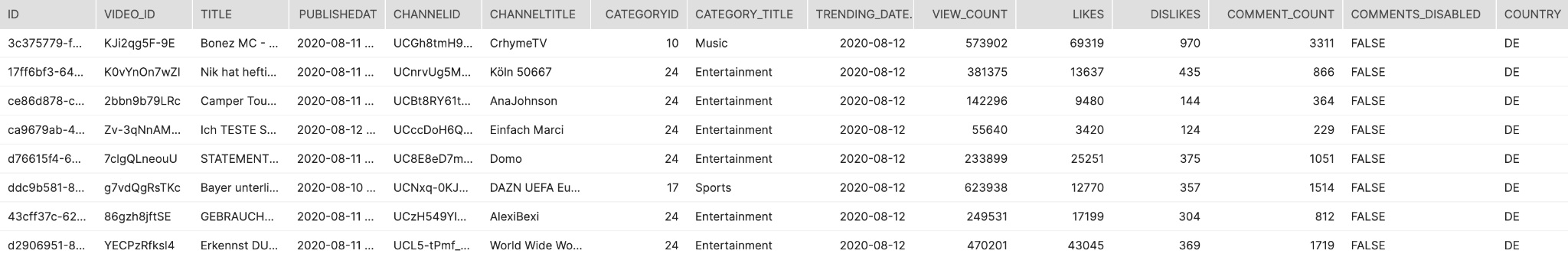


\

* 1. For category data create a table called “*table\_youtube\_category*” with:

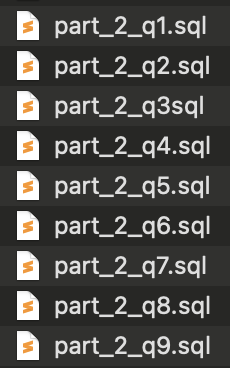


1. Create a final table called “*table\_youtube\_final*” by combining “*table\_youtube\_trending*” and “*table\_youtube\_category*” on *country* and *categoryid* (**be careful to not lose any records**), while adding a new field called “id” by using the “UUID\_STRING()” function :



**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:

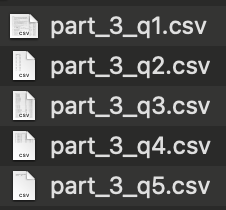
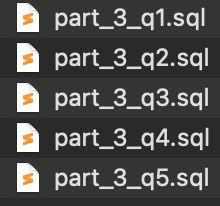


We can assume that the highest number of *view\_count* will be the record to keep when we have duplicates.

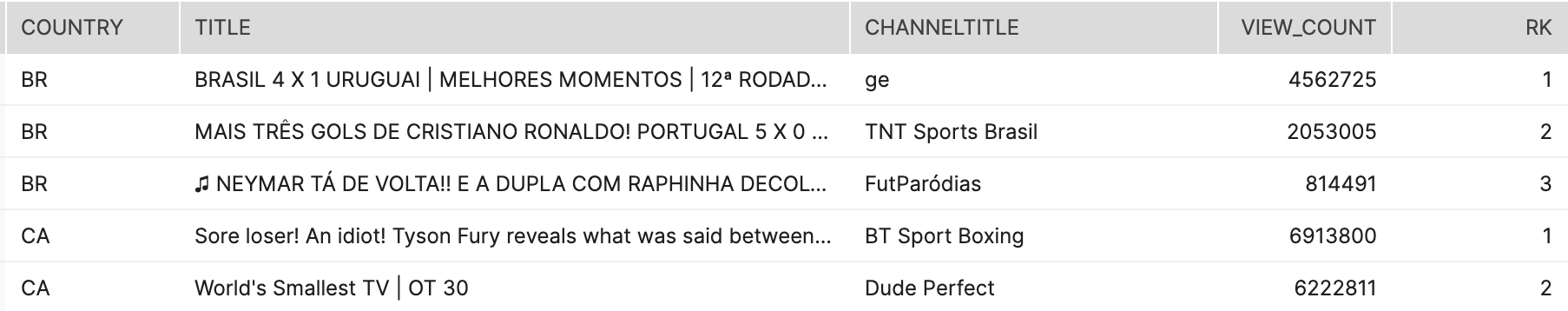
1. Create a new table called “*table\_youtube\_duplicates”*  containing only the “bad” duplicates by using the *row\_number()* function.
2. Delete the duplicates in “*table\_youtube\_final*“ by using “*table\_youtube\_duplicates”*.
3. Count the number of rows in “*table\_youtube\_final*“ and check that it is equal to **1,123,017 rows.**

**PART 3: Data Analysis**

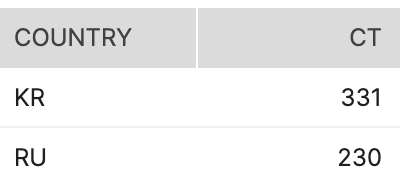
For each question provide a sql file containing the sql code used **AND** a csv containing the output (you can use the snowflake export feature):

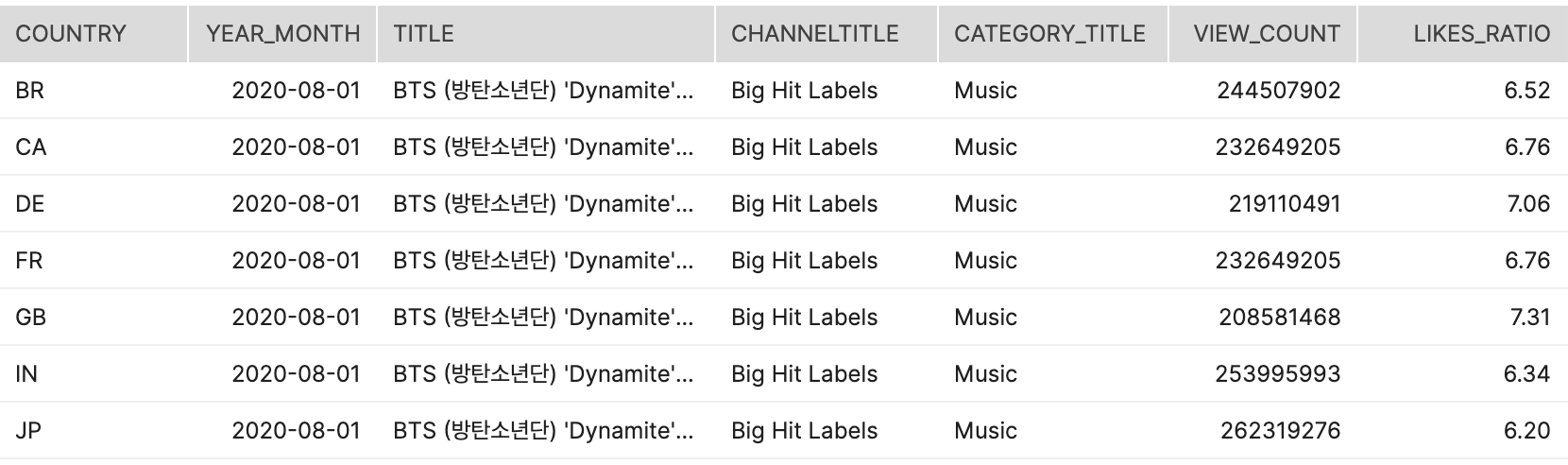
1. What are the 3 most viewed videos for each country in the “Sports” category for the *trending\_date* = ‘'2021-10-17'’. Order the result by *country* and the *rank*, e.g:



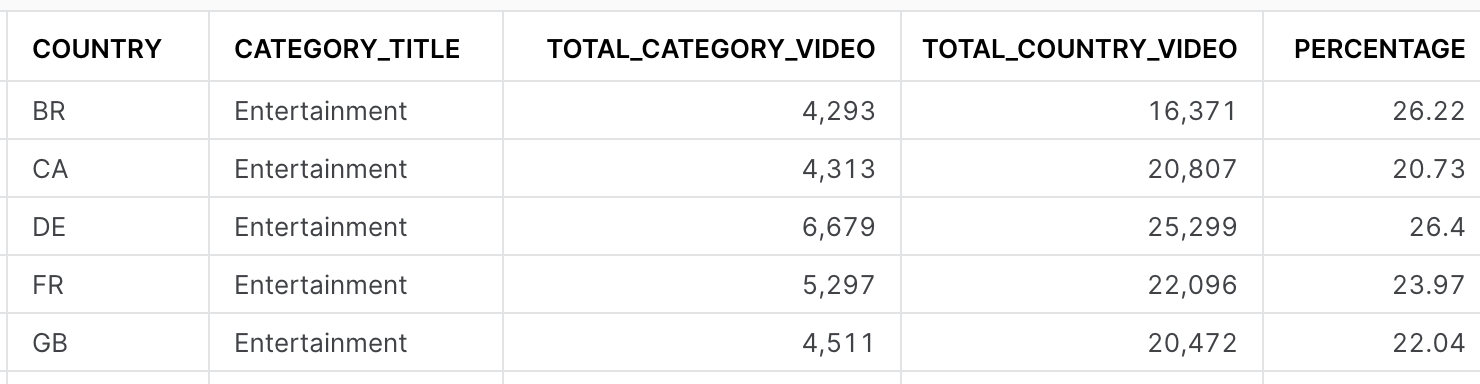
1. For each country, count the number of **distinct** video with a title containing the word “BTS” and order the result by count in a descending order, e.g:



1. 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:



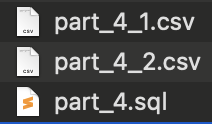
1. For each *country*, which *category\_title* has the most **distinct** videos and what is its percentage (2 decimals) out of the total **distinct** number of videos of that *country*? Order the result by category\_title and *country*. The output should like this:



1. 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](https://drive.google.com/file/d/1_oJlSMUlvefQBLikFbMCb8sP9s0FmOym/view?usp=sharing)

[Example 2](https://drive.google.com/file/d/1LXi98PBpD-T9FoDeTJZChlAqU8vU-0JR/view?usp=sharing)

**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 (5th September 2023)** on Canvas. Penalties will be applied for late submission