UNIVERSITÉ DE NEUCHÂTEL

EXERCISE 10
DATA MANAGEMENT

Group 11

Alexandre Ducommun
Rabeb Ben Ramdhane
Abdallah Shukor

Professor

Iulian Ciorascu

09 December 2019

Question 1 - Analysis

The given dataset contains data on daily trending YouTube videos for 10 countries. Each countries data is in a separate file and split in two formats. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. A line in this file represent a trending video that has a unique "video_id" and "trending_date" as unique Identifier. It also includes a "category_id" field, which varies between countries, to retrieve the categories for a specific video in the associated JSON file. One such file contains a list of items where the categories are stored and is included for each of the countries in the dataset.

FRvideos.csv

Attribute Name	Type	Description
video_id	String	PK1, Unique attribute that identifies the video
trending_date	String	PK2, Attribute that defines the trending date of the video
title	String	Attribute that defines the title of the video
channel_title	String	Attribute that defines the title of the channel
category_id	ID Key	Unique attribute that identifies the category of the video from JSON
publish_time	Date	Attribute that defines the date of the publication
tags	String	Attribute that defines the tags used for the video
views	Integer	Attribute that defines the number of views
likes	Integer	Attribute that defines the number of likes
dislikes	Integer	Attribute that defines the number of dislikes
comment_count	Integer	Attribute that defines the number of comments
thumbnail_link	URL	Attribute that defines the URL of the thumbnail of the video
comments_disabled	Boolean	Attribute that defines if comments are disabled
ratings_disabled	Boolean	Attribute that defines if ratings are disabled
video_error_or_removed	Boolean	Attribute that defines if the video contains an error or was removed
description	String	Attribute that describes the video

FR_category_id.json


```
root: {} 3 items
  kind: youtube#videoCategoryListResponse
  etag: "ld9biNPKjAjgjV7EZ4EKeEGrhao/1v2mrzYSYG6onNLt2qTj13hkQZk"
  items: [] 31 items
    0: {} 4 items
      kind: youtube#videoCategory
      etag: "ld9biNPKjAjgjV7EZ4EKeEGrhao/Xy1mB4_yLrHy_BmKmPBggty2mZQ"
      id: 1
      snippet: {} 3 items
        channelId: UCBR8-60-B28hp2BmDPdntcQ
        title: Film & Animation
        assignable: true
```


The dataset includes data for the following regions: United States of America, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India. The data of these regions concerns the same time period. These data sources can be found on https://www.kaggle.com/datasnaek/youtube-new or be generated by using the Kaggle script on https://github.com/mitchelljy/Trending-YouTube-Scraper/blob/master/scraper.py .

Question 2 – Atomics Values

Data columns that haven't atomic value:

- **Publish_time:** It's considered to be a non-atomic attribute since it contains several values that of different data types. In our case, we want to split the publish time into **Date (Y/M/D) and Time (H/M/S).**
- **Tags:** It's considered to be a non-atomic attribute since it contains text that is separated by the following character "|". Practically speaking, this attribute should be split because it is a multivalued attribute that can be split and place in a new table with the key of the trend video.

Question 3 – Data Integration

The data sources for the different regions have the same structure and the same meaning. So we decided to put all data coming from csv files into one table called "TRENDING_VIDEOS". To manage the non-atomic attributes, we will split publish_time in three new columns as publish_time, publish_date and publish_tod. For tags we will create a new table called "VIDEO_TAGS" where we will add the key of the trending video (trending_video_id) and the tag name in each line. For categories, we will create another table called "CATEGORY" where we put the country code and the category_id as the key. In order to make the relationships more convenient we decided to add a surrogate key (trending_video_id) which will be the primary key of the "TRENDING_VIDEOS" table. In that way, we can directly pass this key as foreign in the other tables without passing 3 attributes as the entire primary key.

Question 4 – Data Traceability

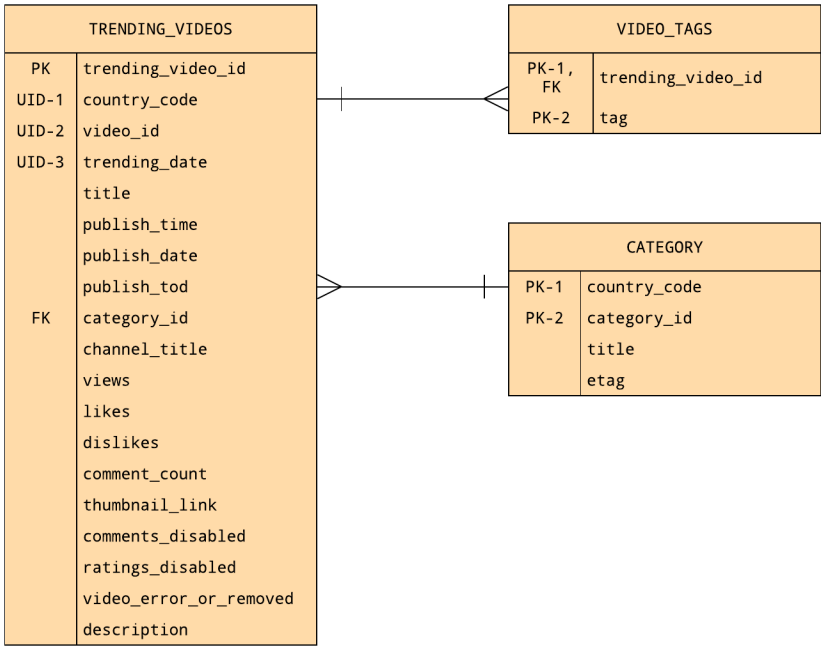
Because we integrated the data sources in one destination, we must manage the traceability of the data. We will add a new column that will contain the country code of the data. This column is populated with the country code of the file during the transformation step. We also add the "country_code" column to the table "CATEGORY".

Question 5 – Data sources interconnection

Data source contains two types of file, CSV and JSON. The CSV contains the information of the videos and the JSON a list of categories. The reason is that categories may differ between countries, so each country has its list of categories in a JSON file. Each video in the csv file contains a key that matches a category ID in the appropriate JSON file.

In the future database we will create a new table where to store all the categories id, title, etags and add a column "country_code" to keep traceability of the data. A line will represent a category with country code and id as key.

Question 6 – Design database



Question 7 – Steps for an automated system

Data Extraction

In order to create an automated system we should first extract the data. In order to do that we need to scrape the following url: https://www.kaggle.com/datasnaek/youtube-new/data in order to retrieve all the files found in the data sources list. This list contains several files (10 csv files and 10json files). The process behind the automated system is to be able to download all these files automatically and then merge them in a specific table. The csv files contain information about youtube videos of several countries while the json files represent the category of each of these video based on the country. So in order to retrieve the data from the json file, we need to store them in a dictionary and then we could save them in a csv file in order to be able to join several tables together. The system should scrape all the data using the YouTube API. It will save the data by country locally in csv files and json files.

Data Transformation

After the data is retrieved and before storing in a local database we need to do some further manipulation and transformation to the dataset. Since we want to merge all countries together in a single csv file we need to differentiate them using a new attribute called country_code which will basically identify the name of each country. This will also keep traceability of the data to the system. Then, in order to have a consistent and non-redundant database we need to handle non-atomic attributes. In the csv files, we notice that there are two non-atomic attributes which are tags and publish_time. In order to make the publish_time atomic we split it into publish_date, publish_time, publish_tod. As for the tags we split into a column called tag and we separate it into a new table "TAGS" using the surrogate key of the table "TRENDING_VIDEOS" as a foreign key in order to join the two tables. A line in TAGS will represents a tag name with the id of the trending video that it belongs. As a final step, in order to search for a specific category of videos we need to split the categories which contains title (example: Animation & Drama…). So putting the category in a separate table having the surrogate key "trending_video_id" as a foreign key in order to manipulate data within the two tables.

Data Loading

Once the transformation is finished, a new database will be created locally using Sqlite3. This database will include all the tables mentioned above and a new schema will be generated based on the tables we designed having the primary keys and all necessary attributes.The final database will be formed of 3 tables. The first table which is "TRENDING_VIDEOS" containing all the countries in the data source as well as the categories found in the json files related to each country. The second table which is VIDEOS_TAGS which will represent all the tags of a specific video The third table "CATEGORY" which represents the category of each video. This automated system will create a consistent, convenient and non-redundant database. It will be for visualization and analysis.