ETL Project – Group 2 – Trending YouTube Videos

**Project Tasks**

1. **Create a `country\_db` database in pgAdmin 4 then create the following two tables within:**

**(2 people: \_\_\_\_\_ , \_\_\_\_\_)**

* A `US\_Videos` table that contains the columns ‘video\_rank’, ‘video\_id’, ‘country\_id’, ‘trending\_date’, ‘title’, ‘channel\_title’, ‘category\_id’, ‘publish\_time’, ‘views’, ‘likes’, ‘dislikes’, ‘comment\_count’
* A `GB\_Videos` table that contains the columns ‘video\_rank’, ‘video\_id’, ‘country\_id’, ‘trending\_date’, ‘title’, ‘channel\_title’, ‘category\_id’, ‘publish\_time’, ‘views’, ‘likes’, ‘dislikes’, ‘comment\_count’
* A `FR\_Videos` table that contains the columns ‘video\_rank’, ‘video\_id’, ‘country\_id’, ‘trending\_date’, ‘title’, ‘channel\_title’, ‘category\_id’, ‘publish\_time’, ‘views’, ‘likes’, ‘dislikes’, ‘comment\_count’
* A `CA\_Videos` table that contains the columns ‘video\_rank’, ‘video\_id’, ‘country\_id’, ‘trending\_date’, ‘title’, ‘channel\_title’, ‘category\_id’, ‘publish\_time’, ‘views’, ‘likes’, ‘dislikes’, ‘comment\_count’
* A `MX\_Videos` table that contains the columns ‘video\_rank’, ‘video\_id’, ‘country\_id’, ‘trending\_date’, ‘title’, ‘channel\_title’, ‘category\_id’, ‘publish\_time’, ‘views’, ‘likes’, ‘dislikes’, ‘comment\_count’
* Be sure to assign a primary key, as Pandas will not be able to do so. The primary key across all tables/CSV files will be ‘video\_id’.

SAVE CODE AS AN SQL FILE, UPLOAD TO GitHub

**In Jupyter Notebook perform all ETL:**

**(2 people: \_\_\_\_\_, \_\_\_\_\_)**

1. **Extraction**

* Put each CSV into a pandas DataFrame (download cleaned CSV files from GitHub to your local device, do a pd.read\_csv function in Jupyter Notebook to read

1. **Transform**

* Copy only the columns needed into a new DataFrame:

Do this in Jupyter:

df[[‘column 1’, column 2’, ‘column 3’, ‘etc’]].copy()

* Rename columns to fit the tables created in the database (we already dropped the columns we didn’t need manually, but we can do it Jupyter if needed).
* Set index to the previously created primary key (video\_id).

1. **Load**

* Create a connection to database.

Do this in Jupyter:

rds\_connection\_string = "<insert user name>:<insert password>@localhost:5432/nameofdatabase"

engine = create\_engine(f'postgresql://{rds\_connection\_string}')

* Check for a successful connection to the database and confirm that the tables have been created.

Do this in Jupyter: engine.table\_names()

* Append DataFrames to tables. Be sure to use the index set earlier.

Do this in Jupyter:

df.to\_sql(name=<database name>, con=engine, if\_exists='append', index=False)

* Confirm successful “Load” by querying database.

Do this in Jupyter:

pd.read\_sql\_query('select \* from <database name>’, con=engine).head()

* Join the tables (Postgres) and select the `video\_id` and other columns from the first table, and along with those same columns from the other tables.

1. **Analysis**

* Create 5-10 questions to ask/answer about the data sets.
* Analyze similarities/differences between countries