**ETL Project**

**Team Kappa**

*Main Insights*

1.- Background

The global **video streaming** market size was estimated at USD 42.6 billion in 2019 and is expected to reach USD 50.1 billion in 2020 and is expected to grow at a compound annual **growth** rate of 20.4% from 2020 to 2027, in order to reach USD 184.2 billion by 2027.

The video streaming services have experienced a rise of around 30% in viewership during the present lockdown.

Our project consists of extracting and loading **a Movie and Series Databases of the 4 major streaming platforms: Netflix, Prime Video, Disney+ and Hulu (**obtained from the site **Kaggle).**

* [**Tv shows data set**](https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney)
* [**Movies data set**](https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney)

In this sense, we loaded both databases to PgAdmin to join them in order to have one database with both movies and series by each platform. It is important to mention that the database only takes into account movies and series that are exclusive to their corresponding platforms.

2.- ETL process

Both databases have more than 20,000 records, with the Movies database being the largest one (16,000 records); however, the data extracted from source (Kaggle) is raw and with some problems. Therefore, it needs to be cleansed, mapped and transformed.

Some of the main validations that we did during this stage:

* Filtering – Select only certain columns to load
* Using rules and lookup tables for Data standardization
* Required fields should not be left blank
* Cleaning (for example, mapping NULL to 0, elimination of weird characters, etc.)
  + We eliminated 300 records where 3 important columns had NULLs such as Duration, Country and Language.
  + Replaced 31 special characters in some of the Movies columns.
* Use lookups to merge data

Loading

In this phase, we extracted and transformed the original data and loaded the info into PgAdmin to create two specific tables, Movies and Series, that were unified using the “Union” SQL function to consolidate in one table the data records of Movies and Series with the platform (in this case, the variable StreamPL) as the primary key. With this final step, we created a new database that has all the exclusive movies and series per each platform that can be used to perform some business intelligence analysis in the near future (not part of this ETL project).

3-. Our insights

The ETL process has to be well thought before starting it. How to clean the data & how to manage it has a direct relation to the purpose of use you want to give to the dataset. Also, you have to define if it needs a relational or not relational key for its tables and how you may want the final results look like.

Overthinking is not a good advice but to plan & design before you code is the best tip we learn in this project.