



ITESM & Trilogy

Anacondas team:

Daniela Villarreal

Raúl Flores Palacios

Michelle García

Juan Ramón Félix

Professors:

Mr. Alejandro Estrella Gabilondo

Mr. Jorge Luis Ramos Zavaleta

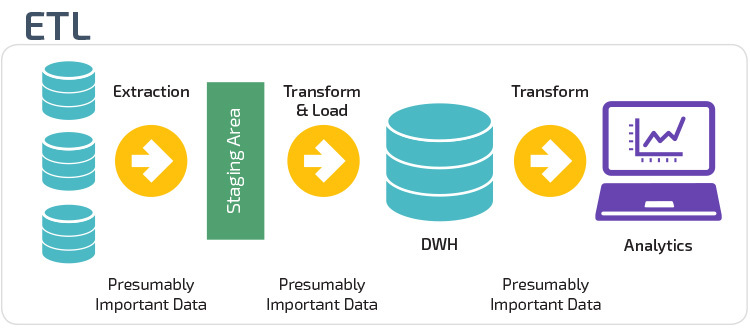
Project ETL

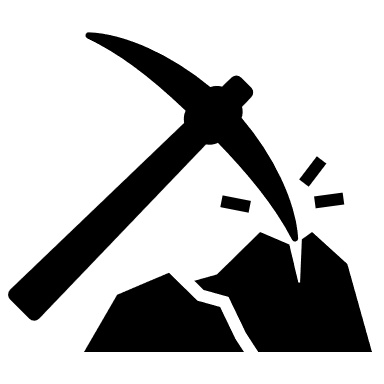
Data Analytics Bootcamp

18 de Julio de 2020

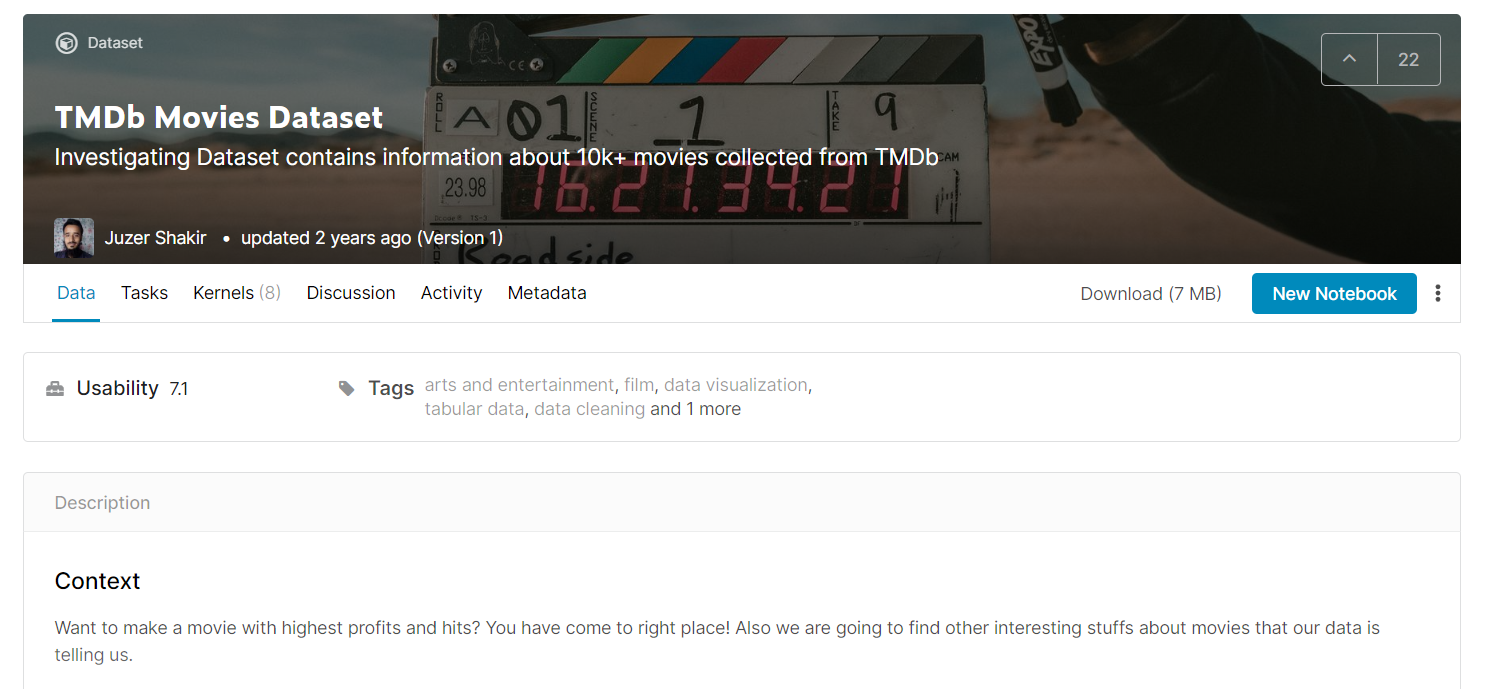
Introduction

For this project we have decided to work with information from movies as is something that we like and have interest in, and also because currently we have allot available spare time to watch a movie, so to make a better selection we have created this database. We were interested to have the ratings of the movies, that is why we use the database of reviews and qualifications of IMBD and Rotten Tomatoes, also we wanted to have a section where we can show in which streaming platform these movies were available. We follow the ETL process to acquire, transform and display our data, starting from the data extraction from databases available, then transforming our data with cleansing and adjustments and then saving this data in more readable and accessible format.

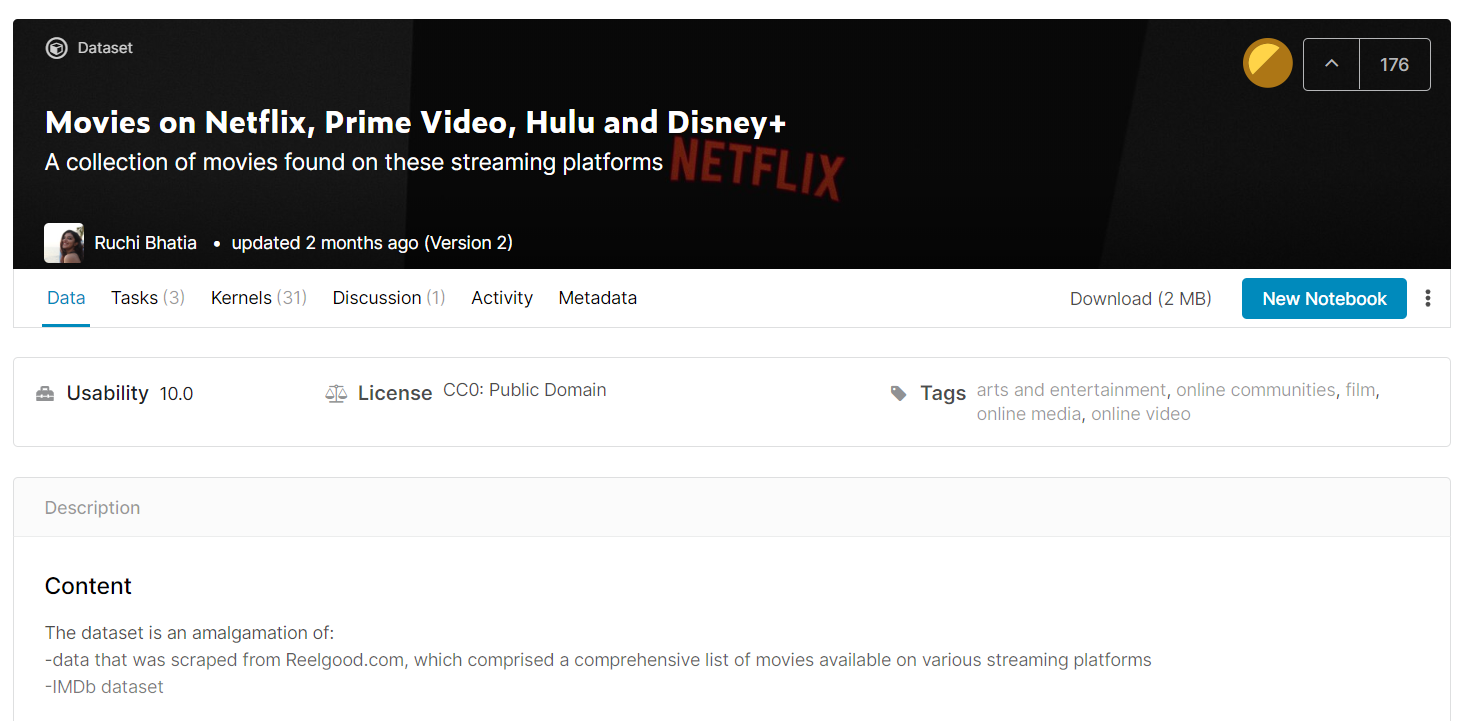


Extraction:

We have decided to use Kaggle as our starting point as is a coarse and useful web pages where we can find multiple Data bases of various topics. We search from its database collection on movies related, and we have found this to databases, that match the best to what we were looking for.

<https://www.kaggle.com/juzershakir/tmdb-movies-dataset>

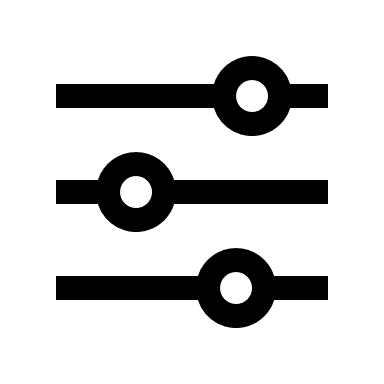
<https://www.themoviedb.org/?language=es>



<https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney>

<https://reelgood.com/>

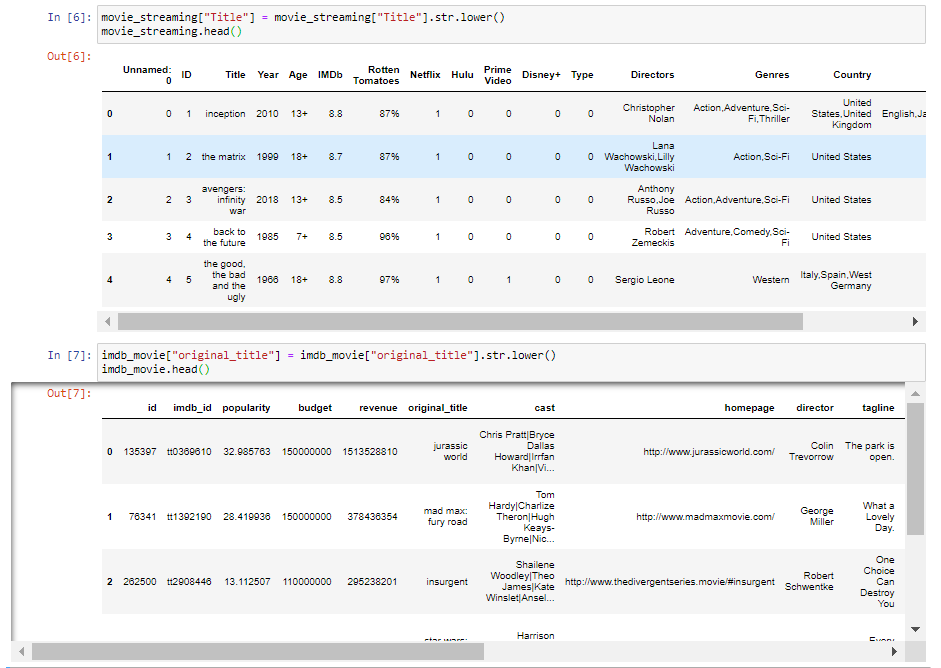
This tow data bases where available in a CSV documents. That is handily as is easier to use for extracting the data.

Transformation:

This was the most complex part of the whole project as we required to do many modifications to the databases in order to adjust them to our preferences and make it more easy for surf through database and let us see only what we wanted. Some of the modification that we made are the following.

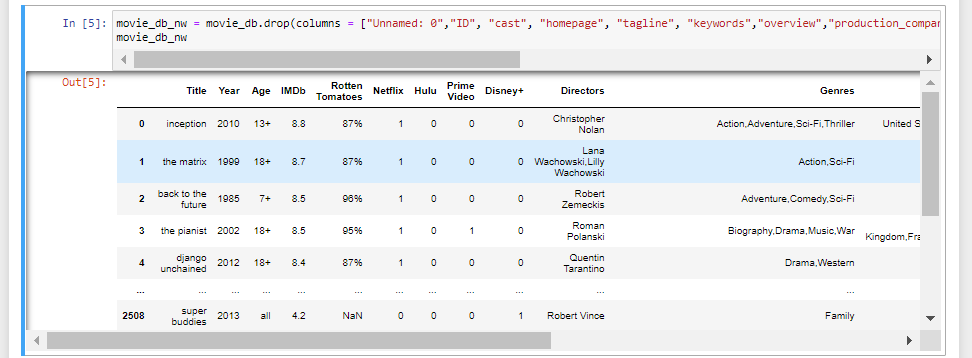
Modification and cleansing 1:

The first modification that we encounter that we required was to adjust the titles or movies names in our CSV files, so they match in format and in name. This way is will be easier and without problems in the merge of the datasets, as is will find more titles that match. One of this modification was to change all the titles to lower case in both datasets.



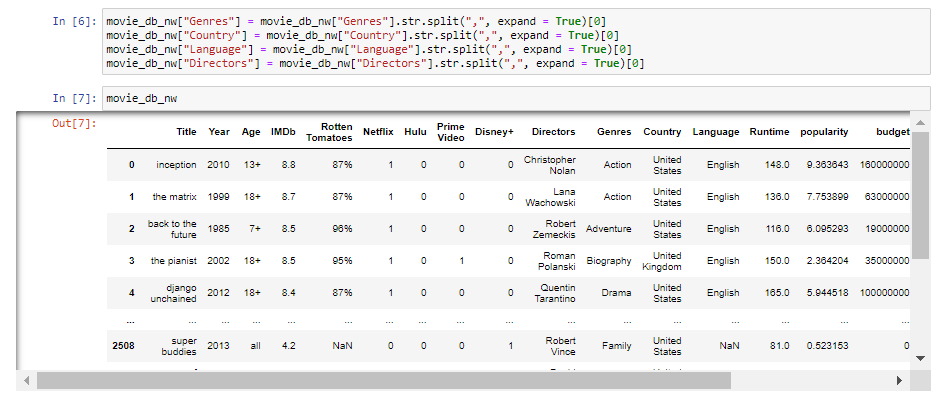
Modification and cleansing 2:

Once we have merged our datasets the we have deleted all the columns that were not interesting for our research Cast, home page, tagline, key words, over view, production company, etc.…, also we have delated the columns that were repetitive. We have as well save some interesting columns that we wanted to have in our data base as are director name, run time, categories, genres, country of production, etc.…



Modification and cleansing 3:

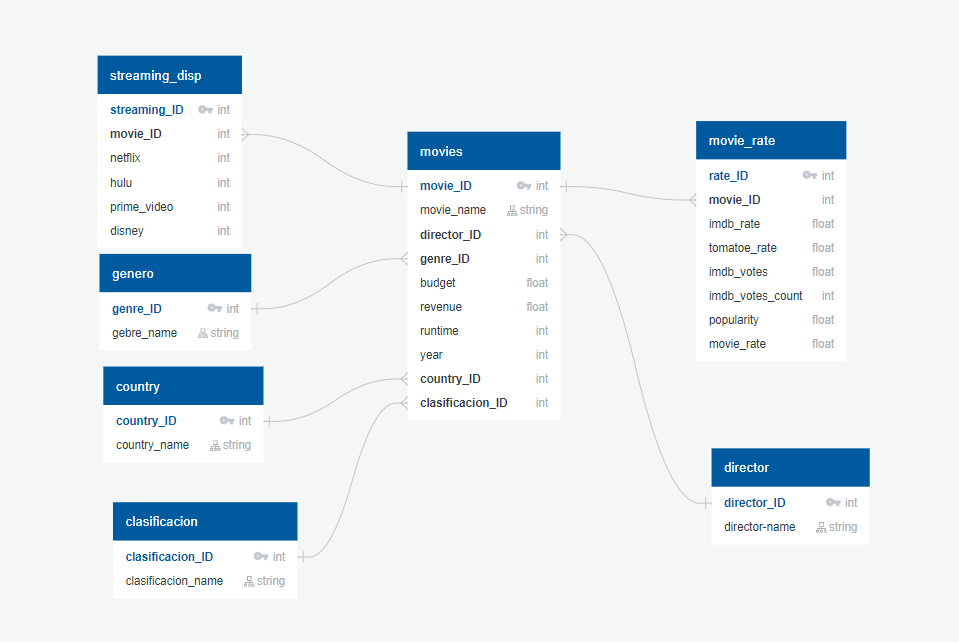
We found that some values in the columns have a list per line, there for we have taken the decision that we were only to use the first value of the list in order to facilitate the searching of the directories. There for if the movie have the genre as action, adventure, sci-fi and thriller, we have only taken the value "action" or if the film was produced in the United States and United Kingdom and Germany have only saved the United States as the country of origin.

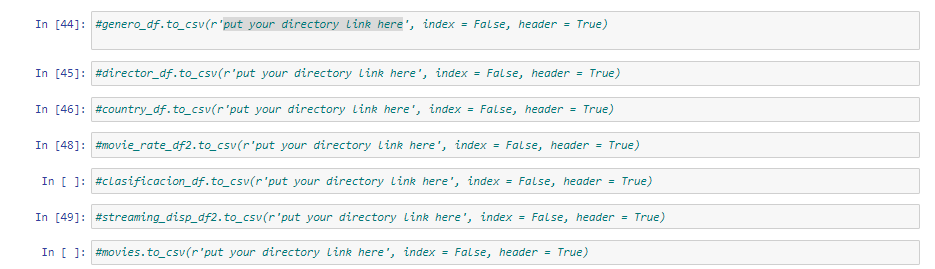


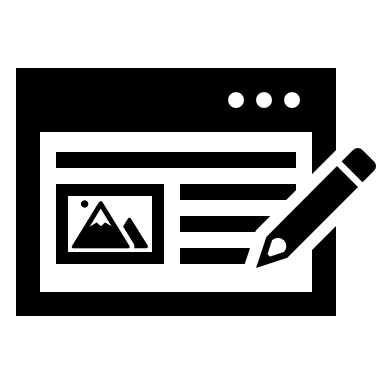
Modification and cleansing 4:

Once we have our main data set clean, we have processed then to divide it in directories to make it in smaller tables that are easier to relate and to review. And in other to use them in our SQL we have make CSV´s of this tables.

DBD diagram





Loading:

We have use PostgreSQL to make a new database with our new tables and their relation references this way you can check the tables of our data base more easy and you can read only the information that you want using queries.

Making the shell scheme:

DROP TABLE IF EXISTS movies CASCADE;

DROP TABLE IF EXISTS movie\_rate CASCADE;

DROP TABLE IF EXISTS streaming\_disp CASCADE;

DROP TABLE IF EXISTS director CASCADE;

DROP TABLE IF EXISTS country CASCADE;

DROP TABLE IF EXISTS genero CASCADE;

DROP TABLE IF EXISTS clasificacion CASCADE;

CREATE TABLE director (

    director\_ID SERIAL PRIMARY KEY,

    director\_name character varying(250) NOT NULL

);

CREATE TABLE genero (

    genre\_ID SERIAL PRIMARY KEY,

    genre\_name character varying(250) NOT NULL

);

CREATE TABLE country (

    country\_ID SERIAL PRIMARY KEY,

    country\_name character varying(250) NOT NULL

);

CREATE TABLE clasificacion (

    clasificacion\_ID SERIAL PRIMARY KEY,

    clasificacion\_name character varying(250) NOT NULL

);

CREATE TABLE movies (

    movie\_ID SERIAL PRIMARY KEY,

    movie\_name character varying(250) NOT NULL,

    director\_ID integer NOT NULL REFERENCES director(director\_ID),

    genre\_ID integer NOT NULL REFERENCES genero(genre\_ID),

    budget FLOAT NOT NULL,

    revenue FLOAT NOT NULL,

    runtime integer NOT NULL,

    year integer NOT NULL,

    country\_ID integer NOT NULL REFERENCES country(country\_ID),

    clasificacion\_ID integer NOT NULL REFERENCES clasificacion(clasificacion\_ID)

);

CREATE TABLE movie\_rate (

    rate\_ID SERIAL PRIMARY KEY,

    movie\_ID INT NOT NULL REFERENCES movies(movie\_ID),

    imdb\_rate FLOAT NOT NULL,

    tomatoe\_rate FLOAT NOT NULL,

    imdb\_votes FLOAT NOT NULL,

    imdb\_votes\_count integer NOT NULL,

    popularity FLOAT NOT NULL,

    movie\_rate FLOAT NOT NULL

);

CREATE TABLE streaming\_disp (

    streaming\_ID SERIAL PRIMARY KEY,

    movie\_ID INT NOT NULL REFERENCES movies(movie\_ID),

    netflix integer NOT NULL,

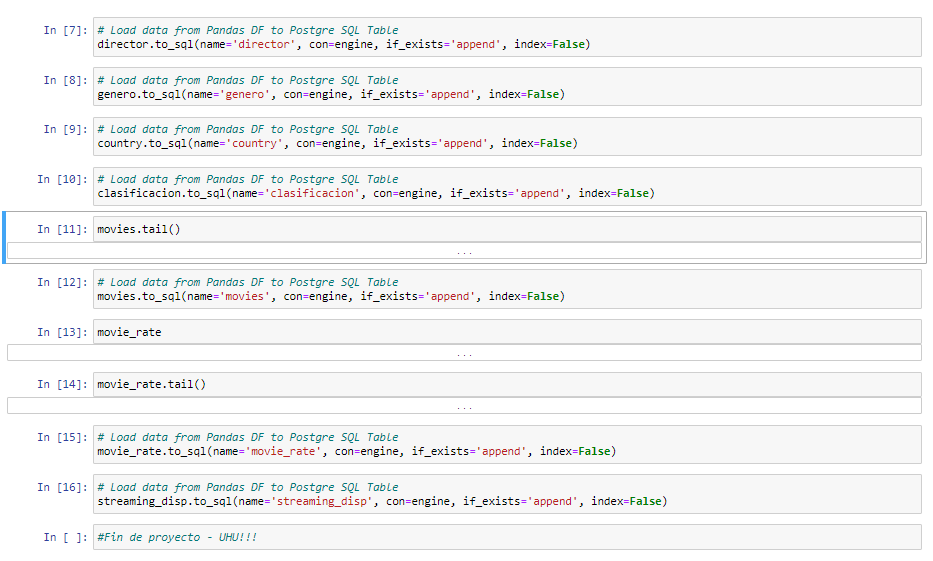
    hulu integer NOT NULL,

    prime\_video integer NOT NULL,

    disney integer NOT NULL

);

Loading the data:



Challenges and conclusion

This was a fascinating project as it was our first encounter that we can do analysis and extraction from data in the real world and not just from academic content that is previously checked and reviewed for us the students. We think that things same really simple at the beginning but then you realize that you have make a mistake in the cleansing and transformation, that is a really common mistake that happens in the task of reviewing data. We found that we have need to make many adjustments to have the data according to our original design and that is not simple task as you need to clearly understand what you have, what you want and how to reference correctly the data to achieve that design that you have started with. Also, we found that there is much more extend to our analysis and that we can still contribute more to our database perhaps information from the directors, have more ratings from other sites etc... In conclusion ETL is a really powerful tool that we can use to analyze any data that we want from the web, but it only can be achieved if you put a lot of effort in understanding and cleaning your data.

References

Bhatia, R. (2020, May 22). Movies on Netflix, Prime Video, Hulu and Disney+. Retrieved July 19, 2020, from https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

Regodee. (n.d.). Regodee: Where to Stream Movies &amp; TV Shows on Every Service. Retrieved July 19, 2020, from https://reelgood.com/

Shakir, J. (2018, March 24). TMDb Movies Dataset. Retrieved July 19, 2020, from https://www.kaggle.com/juzershakir/tmdb-movies-dataset

TMBD. (n.d.). The movie data base. Retrieved July 19, 2020, from https://www.themoviedb.org/?language=es