**Extract Transform Load Project**

**Team Members:** Jana Simashkevich, Jenny Armour, Depa Ramcharran

**Project Summary:**

Perform Extract Transform Load on movie data sets in scope to prepare the data for further analysis on movie ratings, popularity, genre across 3 different sources of data such as IMDB, Netflix data and Rental information. The final data set could be used to analyze trends. Data was collected using different methods such as API calls, web scraping using BeautifulSoup library and CSV data loads using PostgreSQL.

**Extract**

The following 3 data sources were used for data extraction

1. **IMDB** <https://www.imdb.com/>

We used IMDB website to scrape the data for top 1000 movies for 3 years 2017-2019.

The data was scraped by calling advanced search link which provides 50 movies per page taking offset, start date and end date as arguments.

To collect 1000 movies for one year we had to scrape 20 links per year (60 links in total for 3 years). To avoid being blocked by IMDB we timed the requests by using random wait times.

Sample links

Processsing: <https://www.imdb.com/search/title/?release_date=2017-01-01,2017-12-31&start=1&ref_=adv_nxt> , 2017

.....Sleeping for 2

Processsing: <https://www.imdb.com/search/title/?release_date=2017-01-01,2017-12-31&start=51&ref_=adv_nxt> , 2017

.....Sleeping for 5

Processsing: <https://www.imdb.com/search/title/?release_date=2017-01-01,2017-12-31&start=101&ref_=adv_nxt> , 2017

.....Sleeping for 6

The following fields were captured from the IMDB website

movie\_list\_info.append({"move\_name":movie\_name,

"movie\_year":movie\_year,

"movie\_rating\_imdb":movie\_rating\_imdb,

"movie\_duration":movie\_duration,

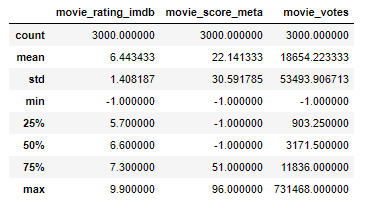
"movie\_genre":movie\_genre,

"movie\_description":movie\_description,

"movie\_score\_meta":movie\_score\_meta,

"movie\_votes":movie\_votes})

Data was temporarily saved in a pandas DataFrame structure

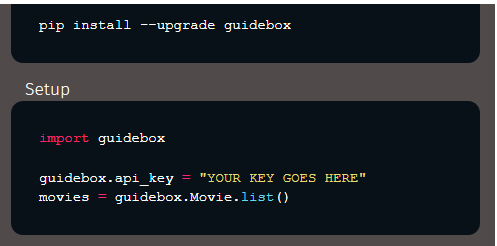




1. **GuideBox** <https://www.guidebox.com/docs>

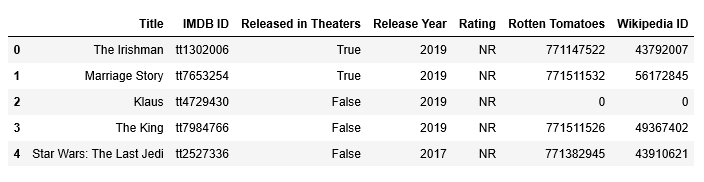
In 2014 Netflix discontinued its API. We found another API, Guidebox that gave access to over 240 streaming platforms including Netflix, Hulu, Prime Video, Showtime, HBO.

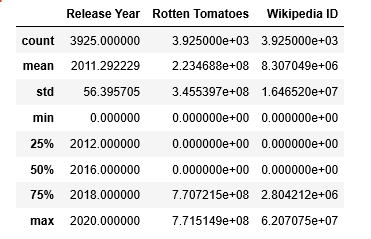
It was a free trial for a week but did limit 250 calls a minute.



A list of almost 4000 movies currently on Netflix was pulled using the offset feature (set at 249) and a loop. We included IMBD ID, Rotten Tomatoes reference number and wikipedia ID when available.

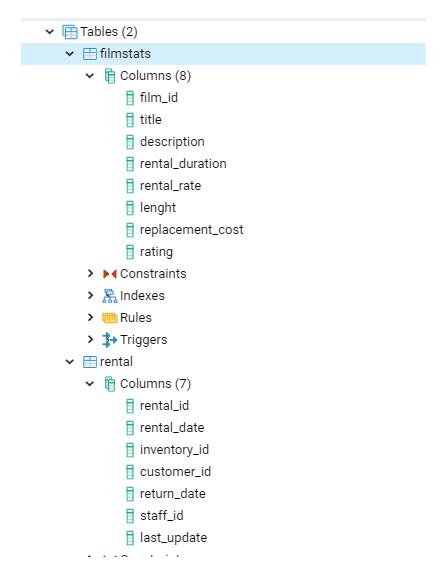
Data was temporarily stored in pandas DataFrame structure for further

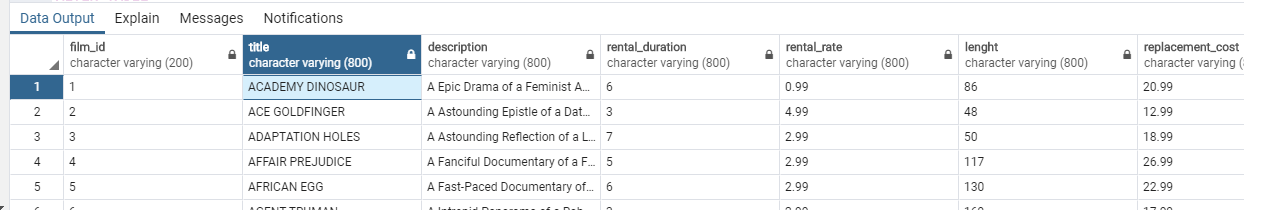




1. **Rental** Info (CSV based)

The Data was Extracted from 2 separate CSV files: Flim\_Actor and Rental CSV file.   
PostgreSQL platform was used to create the table





**Transform**

IMDB data transformations :

* Movie year was cleaned up to remove the brackets and extra characters
* Ratings were converted to floats, Number of voters was converted to integer
* Where ratings were not available data was updated to -1

GuideBox data transformations :

* NULL data was removed where data wasn’t available
* Data containing numbers was converted to integers/floats

Rental Info transformations:

To transform and clean the data the columns that had values that were not relevant to the topic of discussion where dropped. Any tables with the value Null was also dropped.  
ALTER TABLE Filmstats  
DROP COLUMN release\_year;

**Load**  
  
The final data set was loaded into MongoDB database so that further analysis can be performed  
