**Movies ETL**

Movie data from Wikipedia, Kaggle, and aggregated ratings were utilized to assemble a movie database from a clean dataset.

The ETL process was used to extract the Wikipedia and Kaggle data from their respective files, transform the datasets by cleaning rows and formatting datatypes, preforming joins, and loading the cleaned dataset into a SQL database.

**Overview**

The goal of this analysis is to create an automated pipeline that extracts, transforms and loads data. This analysis consists of four parts, where each step is building up from: the beginning of extracting data and function testing, transformation, cleaning process, to its final step of connect and load to the database. The entire process of ETL can be executed with a single function extract,\_transform, and\_load in the final step ETL\_create\_database.ipynb. The ETL process is broken down into four jupyter notebook files:

**ETL\_function\_test.ipynb**

* Data is extracted from the website in JSON and CSV formats.
* Data is transformed into Pandas data frames.
* JSON file requires extra step – loading file first and then transforming into data frame.

**ETL\_clean\_wiki\_movies.ipynb**

* Function clean\_movie combines scattered data of alternative languages into one column alt\_titles.
* Its subfunction change\_column\_name organizes column names into consistent pattern.
* In the function extract\_transform\_load the transformation process of wiki movies data begins and includes:  
  - Python list comprehensions.  
  - apply() and map() methods in combination with lambda functions.  
  - regular expressions or RegEx.

**ETL\_clean\_kaggle\_data.ipynb**

* Function extract\_transform\_load gets new tasks for cleaning Kaggle data and includes:  
  - Changing datatypes, using methods pd.to\_numeric, astype() and python comparison operators for Boolean types.  
  - Filling missing values and filtering unwanted columns.  
  - Merging data frames using pd\_merge method.

**ETL\_create\_database.ipynb**

* The function in its final step connects to the database by sqlalchemy library and to\_sql method.
* Complete ETL process can be executed with a single function extract\_transform\_load call.