**Project-2 Proposal:**

**Project team:**

1. Hamid Zarringhalam
2. Hazim Hamadneh
3. Loic Tiemani
4. Ahmad Abo Alafa

This Project is the continuation of Project-1 which was about Netflix Movies and TV-Series Data Analysis.

This Project is consisting of three parts:

* Extract the data
* Transform the data
* Load the data

# Data Extraction:

Data extraction will include the following tasks:

1. Extract relevant data from data sources:

* https://www.kaggle.com/shivamb/netflix-shows?select=netflix\_titles.csv
* Public movie APIs:
  + 1. OMDB - OMDB to obtain information and metadata about movies.
    2. TMDB - TMDB for movie discovery.

1. Reconcile records with the source data
2. Make sure that no spam/unwanted data loaded
3. Data type check
4. Remove all types of duplicate/fragmented data

# Data Transformation:

Data transformation will include following tasks:

1. Extracted data is coming in different formats, hence data will be standardized
2. Database tables will be created and normalized
3. Data Frames will be merged if required
4. Filtering – Select only certain columns to load
5. Character Set Conversion and encoding handling
6. Conversion of Units of Measurements like Date Time Conversion, currency conversions, numerical conversions, etc.
7. Data threshold validation check. For example, movie release date cannot be before the year 1900
8. Data flow validation from the staging area to the intermediate tables.
9. Required fields should not be left blank.
10. Cleaning (for example, mapping NULL to 0)
11. First normal form, split a column into multiples and merging multiple columns into a single column.
12. Transposing rows and columns
13. Use lookups to merge data

# Data Load:

Data loading will include following tasks:

1. Loading the Data Frames in to the Data Bases
2. Making sure that the process is completed successfully