ASSIGNMENT 1 REPORT

1. ABSTRACT

The purpose of the assignment is to build a conceptual schema for a movie database gathering real world data. The data obtained from the three sources are cleaned, reformatted and then combined to fit the movie database schema.

2. DATA SOURCES AND DATA FIELDS:

To create a movie database (of movies released in 2017) using real world data, three sources of data are used:

1) A web scraper (Data scrapped from the IMDB site)
2) A web API (TMDB API used)
3) Raw csv (containing the info from IMDB site)

Web Scraper:

The web scraper is written in python using the beautiful soup package and using the pandas framework. The web scraper accesses the specified webpage by using the urllib and requests package in python. Multiple pages are scraped to gather around 500 records of data. The HTML tags are closely inspected for the information that needs to be scrapped. It then scrapes the page for specified sources of data and stores the scraped information in a pandas data frame. The scraped data is then exported to a CSV file.

Data fields scraped using the web scrapper include:

- Movie (Name of the movie)
- Genre (Genres to which the movie belongs)
- Runtime (Runtime of the movie)
- IMDB Rating (Rating for the film on IMDB)
- Metascore Rating (Rating for the film on Metascore)

Web API:

To extract movies data for films released in 2017 from a web API, the TMDB API is used. An api_key is used to gain access and get information from the TMDB API. Information is obtained from multiple pages (about 400 records of movies) and stored in a data frame and exported to a CSV file.

Data fields obtained from the web API include:

- Movie (Name of the movie)
- Release_date (The Release date of the movie)
- Popularity (Popularity of the movie)

Raw CSV:

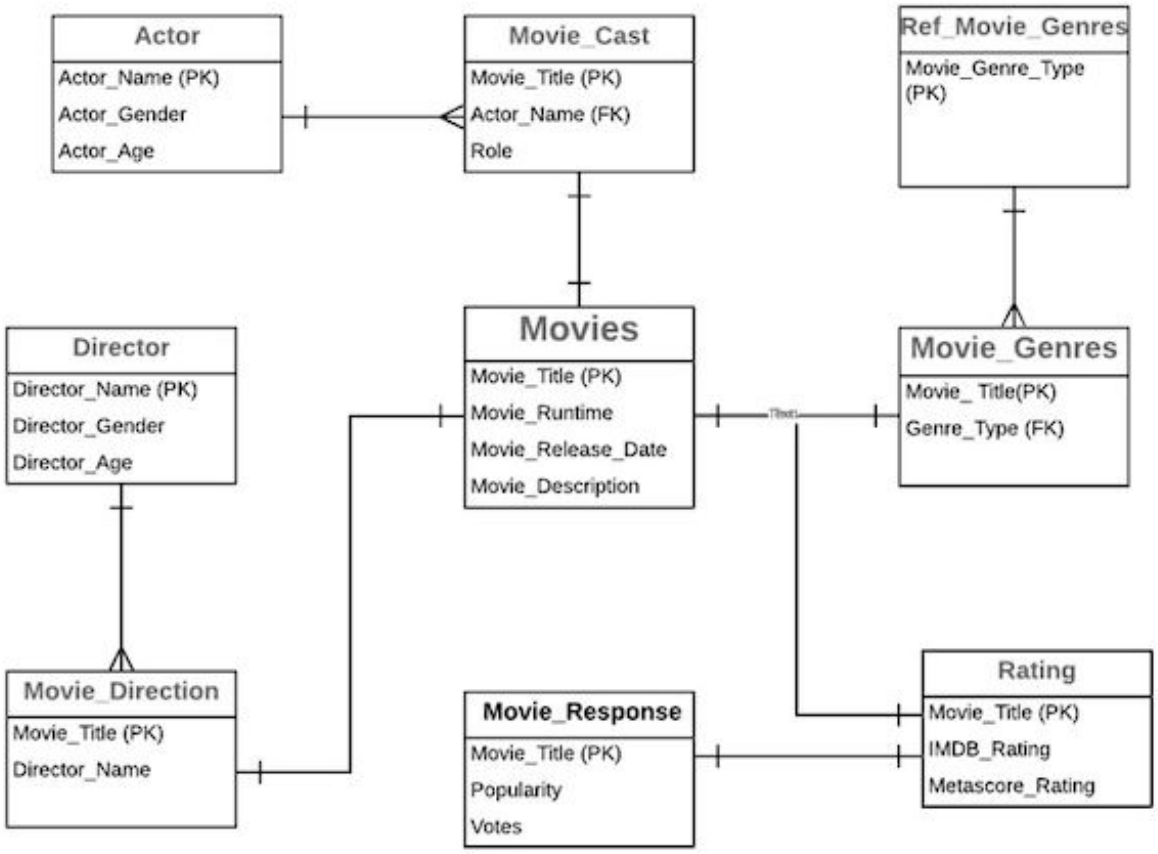
Raw CSV file obtained from the IMDB site is used as input and it has about 500 records.

Data fields obtained from the CSV include:

- Movie (Name of the movie)
- Votes (The number of votes for the movie)
- Director (Director information for the movie)
- Stars (List of stars present in the movie)
- Description (Movie description)

3. CONCEPTUAL SCHEMA EXPLANATION:

The conceptual schema for the movie database is as shown below:



The conceptual schema shows the entities and the relationship between the entities. The various entities are as follows:
Movies, movie cast, movie genres, rating, movie response, actors and director.

- Movies entity (Primary Key: Movie_Title) has attributes such as movie title, runtime, release_date and description.
- Movies_Genres (Primary Key: Movie_Title) entity has attributes such as movie title, genre type.
- Rating entity has attributes such as movie title, IMDB rating, Metascore Rating.
- Movie_response (Primary Key: Movie_Title) entity has attributes such as movie title, popularity, votes.
- Movie_description (Primary Key: Movie_Title) entity has attributes such as movie title, director name.
- Director entity (Primary Key: Director_Name) has attributes such as director name, gender and age.
- Actor entity (Primary Key: Actor_Name) has attributes such as actor name, gender and age.
- Movie_cast (Primary Key: Movie_Title) has attributes such as movie title, actor name and role.

Relationships between the entities are as follows:

- Movies and movie_genres has a one to one relationship as one entry in movies could have one matching entry in movie_genres table.
- Movies and rating has a one to one relationship as one entry in movies could have one matching entry in rating table.
- Movies and movie_cast has a one to one relationship as one entry in movies could have one matching entry in movie_cast table.
- Movies and director has a one to one relationship as one entry in movies could have one matching entry in director table.
- Actor and movie_cast has a one to many relationship as one actor could have cast in multiple movies.
- Director and movie_direction has a one to many relationship as one director could have directed many movies
- Movie_response and rating has a one to one relationship as one entry in movie_response could have one matching entry in rating table.
- Movie_genres and Ref_movie_genre has a one to many relationship as one entry in ref_movie_genres could have many matching entry in movie_genre table.

4. AUDIT VALIDITY:

Data is reformatted to fit the database schema. Using python code as seen below, the null values present in the dataset are eliminated.

⇨ Movies_final_list = Movie_final_join.dropna(how='any',axis=0)

The resulting dataset is clean and does not contain null values or any duplicates.

5. AUDIT COMPLETENESS

The Movies dataset obtained after cleaning is same as the real world data.

6. AUDIT CONSISTENCY/UNIFORMITY

The resulting movies dataset covers the entire possible range for the dataset. The data does not have any limitations or does not contain any invalid or null or negative values.

7. FILES INFORMATION:

The assignment includes the following files:

1) DMDD_Assignment_1_program.ipynb -> Jupyter file of the code.
2) Films_list_1.csv -> Output csv file containing the web scrapping output of the IMDB website (for movies released in 2017)
3) Films_list_2.csv -> Output csv file containing the movies gathered using the TMDB web API (for movies released in 2017)
4) Films_list_3.csv -> CSV file used as input (contains data of movies released in 2017)
5) Movies_Database.csv -> Final output csv file containing the movie dataset after merging data obtained from the three sources and cleaning the data.
6) Conceptual_Database_Schema.pdf -> Conceptual Database Schema for the Movies Database
7) Assignment 1 Report -> Report for the Assignment

Web Scraping:

IMDB website is scrapped for information which includes:
Movie title, genre, runtime, IMDB rating and Metascore Rating (Code present in the Jupyter notebook file DMDD_Assignment_1_program.ipynb)

Subset of the sample output is as follows:

Movie	Genre	Runtime	Imdb_Rating	Metascore_Rating
The Upside	Comedy, Drama	126 min	6.2	45
The Greatest Showman	Biography, Drama, Musical	105 min	7.6	48
The Wife	Drama	99 min	7.3	77
Get Out	Horror, Mystery, Thriller	104 min	7.7	84
Spider-Man: Homecoming	Action, Adventure, Sci-Fi	133 min	7.5	73

Web API:

TMDB API is used to get information of movies released in 2017. (Code present in the Jupyter notebook file DMDD_Assignment_1_program.ipynb)

Subset of the sample output is as follows:

Movie	Release_date	Popularity
The Ash Lad: In the Hall of the Mountain King	2017-09-29	65.911
Thor: Ragnarok	2017-10-25	46.242
Justice League	2017-11-15	40.615
Star Wars: The Last Jedi	2017-12-13	37.97
Spider-Man: Homecoming	2017-07-05	35.315

Raw CSV:

Raw CSV file contains data of movies (released in 2017). Sample format of the data is as follows:

Movie	Votes	Director	Stars	Description
The Upside	8424	Neil Burger	Kevin Hart,Bryan Cranston,Nicole Kidman,Aja Na...	A comedic look at the relationship between a w...
The Greatest Showman	186173	Michael Gracey	Hugh Jackman,Michelle Williams,Zac Efron,Zendaya	Celebrates the birth of show business and tell...
The Wife	10244	Björn Runge	Glenn Close,Jonathan Pryce,Max Irons,Christian...	A wife questions her life choices as she trave...
Get Out	363414	Jordan Peele	Daniel Kaluuya,Allison Williams,Bradley Whitfo...	A young African-American visits his white girl...
Spider-Man: Homecoming	387924	Jon Watts	Tom Holland,Michael Keaton,Robert Downey Jr.,M...	Peter Parker balances his life as an ordinary ...

The data from the three sources is joined and cleaned to remove null values.
Sample output of the final dataset is as follows:

Movie	Genre	Runtime	Imdb_Rating	Metascore_Rating	Release_date	Popularity	Votes	Director	Stars	Description
The Greatest Showman	Biography, Drama, Musical	105 min	7.6	48	2017-12-20	28.796	186173	Michael Gracey	Hugh Jackman,Michelle Williams,Zac Efron,Zendaya	Celebrates the birth of show business and tell...
Get Out	Horror, Mystery, Thriller	104 min	7.7	84	2017-02-24	25.046	363414	Jordan Peele	Daniel Kaluuya,Allison Williams,Bradley Whitfo...	A young African-American visits his white girl...
Spider-Man: Homecoming	Action, Adventure, Sci-Fi	133 min	7.5	73	2017-07-05	35.315	387924	Jon Watts	Tom Holland,Michael Keaton,Robert Downey Jr.,M...	Peter Parker balances his life as an ordinary ...
Thor: Ragnarok	Action, Adventure, Comedy	130 min	7.9	74	2017-10-25	46.242	427652	Taika Waititi	Chris Hemsworth,Tom Hiddleston,Cate Blanchett,...	Thor is imprisoned on the planet Sakaar, and m...

CONCLUSION AND INFERENCE:

In the assignment, all of the database tables are populated with real-world data collected from three sources: web scraper, web API and a raw CSV file. A conceptual database schema is created for the movies dataset. Data collected is reformatted and cleaned to fit the conceptual database schema. After auditing the data for quality and accuracy, the final dataset has about 250 rows which is same as the real-world data.

CITATIONS AND REFERENCE:

http://www.databaseanswers.org/data_models/imdb/index.htm
https://www.dataquest.io/blog/web-scraping-beautifulsoup/