Name: Fuad Choudhury
Course: CIS4400, **Semester:** Spring 2025
Professor: Jefferson Bien-Aime

Homework: 1

Movie Rating Analysis based on IMDB

Business Requirements:

Goal: Based on IMDb ratings, analyze and compare three to four movies to determine which one has the highest rating.

- This will help users quickly identify highly rated movies.
- Additionally, the analysis will consider the release year of each movie.

Functional Requirements:

A simple program that allows users to compare movie ratings.

- The interface should be user-friendly and provide relevant details.
- Display the release year along with the rating score for better comparison.
- Users can filter movies based on year to find top-rated films.
- Provide the option to compare multiple movies.
- Include graphical analysis for visual comparison.

Data Requirements:

- IMDb Movie Ratings
- Release Year
- Number of Votes

Data Sourcing:

- Web Scraping using Python for scraping movie details.
- Using two CSV files from IMDb websites.

Information Architecture:

The information architecture for the Movie Rating Analysis Based on IMDb project illustrates the end-to-end workflow of processing IMDb data to prepare it for analysis and reporting. The project utilizes two main datasets: Title Basic, which includes metadata about movies, and Title Rating, which contains rating information. These datasets are collected from IMDb in CSV format and temporarily stored for initial processing. The data then undergoes a cleaning stage to eliminate missing values and correct formatting issues, followed by a reformatting step where data types are standardized, and column names are updated for consistency. In the transformation phase, the cleaned datasets are joined, and new attributes are derived, such as computed metrics or normalized formats. Once transformed, the data is consolidated into structured tables and loaded into a centralized data warehouse hosted on Microsoft Azure. This warehouse serves as the foundation for efficient querying and downstream analysis.

Data Architecture:

The data architecture represents the overall data pipeline at a high level. It begins with the ingestion of raw IMDb data, which is first stored temporarily before undergoing an integration process. This process includes all steps such as cleaning, transforming, and merging the datasets, as detailed in the information architecture. The integrated data is then loaded into a data warehouse where it is organized and optimized for performance. Finally, the data is made available to visualization tools for building reports, dashboards, and conducting deeper analytical exploration.

Dimensional Modeling:

- The data warehouse is composed of a fact table that includes ratings, votes, and the year, along with dimension tables that classify movies based on their release year.

- Also, a surrogate key is employed for both the fact and dimension tables.

Deliverables:

- Links to data sources (https://datasets.imdbws.com/)

- Azure Storage Link:

Blob-SAS-Token:

sp=r&st=2025-05-03T20:09:13Z&se=2025-05-31T04:09:13Z&spr=https&sv=2024-11-04&sr=c&sig=yS7zZiabUpvHN5lEa0zoCvtYR4PVXdtqU21tv6vVvQE%3D

Blob-SAS-URL:
https://cis4400datastorage2025.blob.core.windows.net/cis4400homework?sp=r&st=2025-05-03T20:09:13Z&se=2025-05-31T04:09:13Z&spr=https&sv=2024-11-04&sr=c&sig=yS7zZiabUpvHN5lEa0zoCvtYR4PVXdtqU21tv6vVvQE%3D

- Data origin: IMDb Non-Commercial Datasets; Customers can access subsets of IMDb data for personal and non-commercial purposes.

  (https://developer.imdb.com/non-commercial-datasets/)

- Data dictionary (Using MS Excel)
- Git repository Link (https://github.com/fuadchoudhury/CIS4400_Homework)
- Data Model (Using Draw.io)
- Accessible IMDb Data Warehouse:

  (https://developer.imdb.com/non-commercial-datasets/)