ETL Challenge – The Movies Dataset

March 13, 2021

Team Members:

Mona Peteet

Banu Nathan

Dan Freda

Jenish Jariwala

TABLE OF CONTENTS

1. INTRODUCTION

Movies can entertain, educate, and inspire the viewer in many ways. Every movie is set and developed in a particular culture. Movies are an integral part of us because they mirror what we believe in and how we coexist as people in a society. Movies are not only influential, but it is a source of entertainment and escapism. In movies you can come across music, stories and pictures all in one. When we watch movies, we forget what is happening out in the real world, but at the same time gives audiences a reality check.

Movies more than entertains us, because they can create awareness about the importance in sports, art, politics and education and they can also warn us about the dangers of alcohol, drugs and other criminal activities. Movies helps us understand more about other cultures.

And movie genres help awaken our sense of responsibly and empathy towards such situations. Different movie genres represent different content that us viewers understand the story telling behind a movie. Movies consist of multiple genres and these genres do their part in telling the story from the beginning till the end.

2.1 EXTRACTION & TRANSFORMATION

Data source:

The dataset we used for this project is obtained from grouplens.org called MovieLens Latest Datsets. This dataset provies ratings from over 600 users of 9000 movies.

Each data file in the dataset is in csv file. The csv files include various columns like userId, movieId, title, genres, rating, etc.

The csv files we used to extract and transform data are movies.csv, ratings.csv, tags.csv, link.csv.

For data extraction and transformation activities we included relational database design, diagramming and creation based on the information we were able to extract from csv. Pandas was used for data transformation which includes cleaning and

filtering the data. And data was extracted and transformed by running queries and creating tables in pgAdmin.

Transformation:

A clean dataset was created containing data from the csv files listed under Data source in the previous paragraph. Jupyter Notebook was used to achieve this.

Steps:

- Read each of the three csv files to a pandas data frame.
- Cleaned the data frames to include only relevant columns that could be useful for analysis.
- Cleaned the data frame for null and empty columns.
- Cleaned movie data frame was created to delete columns that were empty and not relevant.
- Converted data dictionary into data frame to illustrated

Screenshot of Clean Data Frame:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

2.2 LOADING OF DATA TO DATABASE:

Quick database diagrams environment was used to create the diagram and load name and columns of the csv files to understand the relationship between different data files.

PostgreSQL was used to load the extracted and transformed data. Here tables were created from the diagram. Then queries were structed to show the analysis of the dataset for better representation.

Query Editor Query History

```
CREATE TABLE "movies" (
    "movieId" INT   NOT NULL,
    "title" varchar   NOT NULL,
    "genres" varchar   NOT NULL,
    CONSTRAINT "pk_movies" PRIMARY KEY (
        "movieId"
     )
);

CREATE TABLE "ratings" (
    "userId" INT   NOT NULL,
    "movieId" INT   NOT NULL,
    "rating" FLOAT  NOT NULL,
    "timestamp" INT   NOT NULL);
--     CONSTRAINT "pk_ratings" PRIMARY KEY (
--         "userId"
--      )
-- );

CREATE TABLE "tags" (
    "userId" INT   NOT NULL,
    "movieId" INT   NOT NULL,
    "tag" VARCHAR   NOT NULL,
    "timestamp" INT   NOT NULL);
--     CONSTRAINT "pk_tags" PRIMARY KEY (
--         "userId"
--      )
-- );

```

DATA SOURCE

This dataset is available on Grouplens.org.

https://grouplens.org/datasets/movielens/latest/