Edward Rees

Computer Science

CS 333 Introduction to Database Systems

Spring 2022

Design, Load, and Explore a Movies database

Chapter 1: Project Description

The Goal of this Project

The goal of this project is to better understand the process of creating and working with a database. This will be done by understanding various components of the database design, such as: loading data from a file, designing and building a database based on information provided, testing the database with sample queries, exploring the database, querying the database created, and optimizing queries for the database. In doing these, a stronger foundation and understanding of database design will be gained.

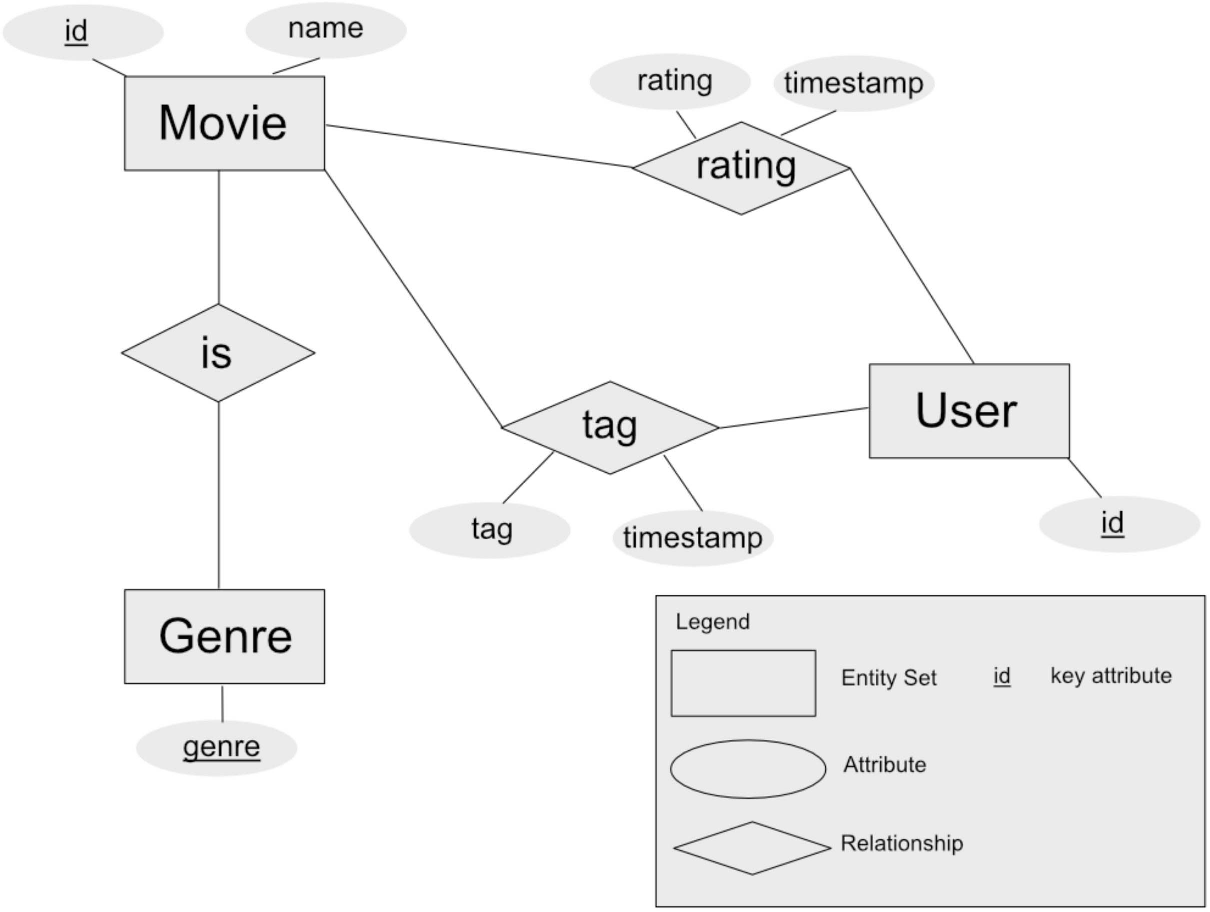
Data Exploration

The dataset contains three txt files. The first is the movies.txt file, which contains a list of movies with each row containing the movie id, the movie name with the year released included, and a list of genres. The movie id is an integer, while the movie name and the genres are all strings. The second is the ratings.txt file, which contains a list of ratings from a given user id, movie id, rating, and the timestamp for the rating. All of the attributes from the ratings.txt are integers, as they are all represented as numbers. The third file is the tags.txt file, which contains the user id, the movie id, the tag, and the timestamp. The user id, movie id, and timestamp are all integers, while the tag itself is a string.

Chapter 2: Database Design

E/R Diagram

When looking at the data, I came up with the E/R diagram below. I saw that Movie would have to be its own entity set, with the attributes of id and name. I initially thought of Rating and Tag as another entity set, until noticing that userId is common in both, which made me think that User can be its own entity set, with an ID attribute. This leads to Rating and Tag both being relationships between Movie and User, with Rating have rating and timestamp as additional attributes, and Tag having tag and timestamp as additional attributes. I then considered Genre as its own entity set connecting with Movie, with genre being a primary key, with Movie and Genre having a many-to-many relationship.



Logical Schema

Movie (id: text, name: text)
User (id: text)

Rating (movieId: text, userId: text, rating: text, timestamp: timestamp)

Tag (movieId: text, userId: text, tag: text, timestamp: timestamp)

Genre (genre: text)