CSCI 620/Section 02/Mior, Introduction to Big Data, Spring 2195

Assignment 1 – Relational model

Let's model IMDB (the Internet Movie Database). You can find a description and the datasets to download here: https://www.imdb.com/interfaces. The database system we are going to use in this assignment is PostgreSQL (any 9.X or later version will work).

Your tasks

1. Provide a description of the contents of the files. This should attempt to explain the purpose of each file and its contents. You should not simply copy the text from the web page.
   (15 points)

2. Provide an ER diagram which represents the IMDb data. This does *not* mean your ER diagram should exactly match the data files. You should aim for your entity sets to represent things stored in the database, which may not always correspond to a data file. Restrict yourself to non-adult movies (including genres, ratings and votes), actors (not including character information), directors, writers and producers (we are not interested in jobs, professions or "known for titles"). **(15 points)**

3. Provide a relational model to store IMDB information. Write SQL scripts to create the whole schema including users and privileges. You need to explore the dataset to decide the sizes of the attributes in advance and be careful when using reserved words like order or character. Integers should be used instead of strings for primary keys. Provide a brief description on how you solved these issues. **(10 points)**

4. Provide a program to load the IMDB data from the text files into the database. You should take into account the charset of the text files. Your program needs to load the whole database in approximately four hours using commodity hardware. Provide a brief description on how you solved these issues and report your timings.
   (45 points)

5. Provide a program that connects to the previous database and creates a transaction to insert three rows of data. Force an error in row #2 in such transaction and ensure that the database is in the same state as before.
   (15 points)