CSCI 620//Brown, Introduction to Big Data, Spring 2225

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 an ER diagram to represent the IMDb data. This does *not* mean your ER diagram should match the data files. You should aim for entity sets to represent things which will be stored in a database, which may not correspond to any particular data file. Restrict yourself to information on titles (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)**

2. Create a relational model to store IMDB information based on your ER diagram. Write SQL scripts to create all tables in the database including primary and foreign keys. 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. (Note that you can simply remove the two character prefix from all primary and foreign keys.) Provide a brief description on how you solved these issues. **(10 points)**

3. Provide a description of the contents of the files in the IMDB dataset. 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)**

4. Provide a program to load the IMDB data (excluding adult titles) into the database you created in Q2. Note that some foreign keys which you wish to define may be invalid. You can skip inserting rows with invalid foreign keys. Your program needs to load the whole database in approximately four hours using commodity hardware (actual runtime will likely be much less). This program may simply be a text file of SQL statements. Provide a description of how you solved these issues and report your timings. **(45 points)**

   Hint: You may find the Postgres COPY command useful. However, if you use this command, you will still need to manipulate the data after loading. If you use the copy command, you may need the option QUOTE E'\b' to properly handle the quotes present in the file.

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 so that the transaction aborts. Your program should check to ensure that the database is in

the same state as before. (That is, row #1 should not have been inserted.)
(15 points)