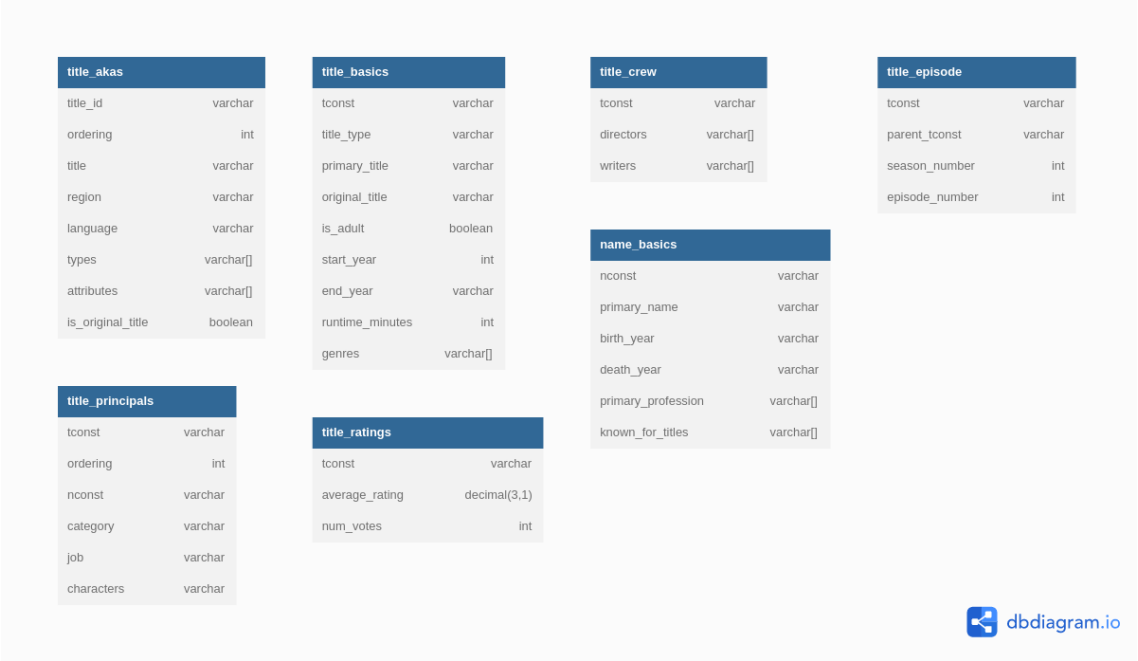
DBMS Project

Team : 3 Amigos

March 20, 2023

1 Raw Data Set

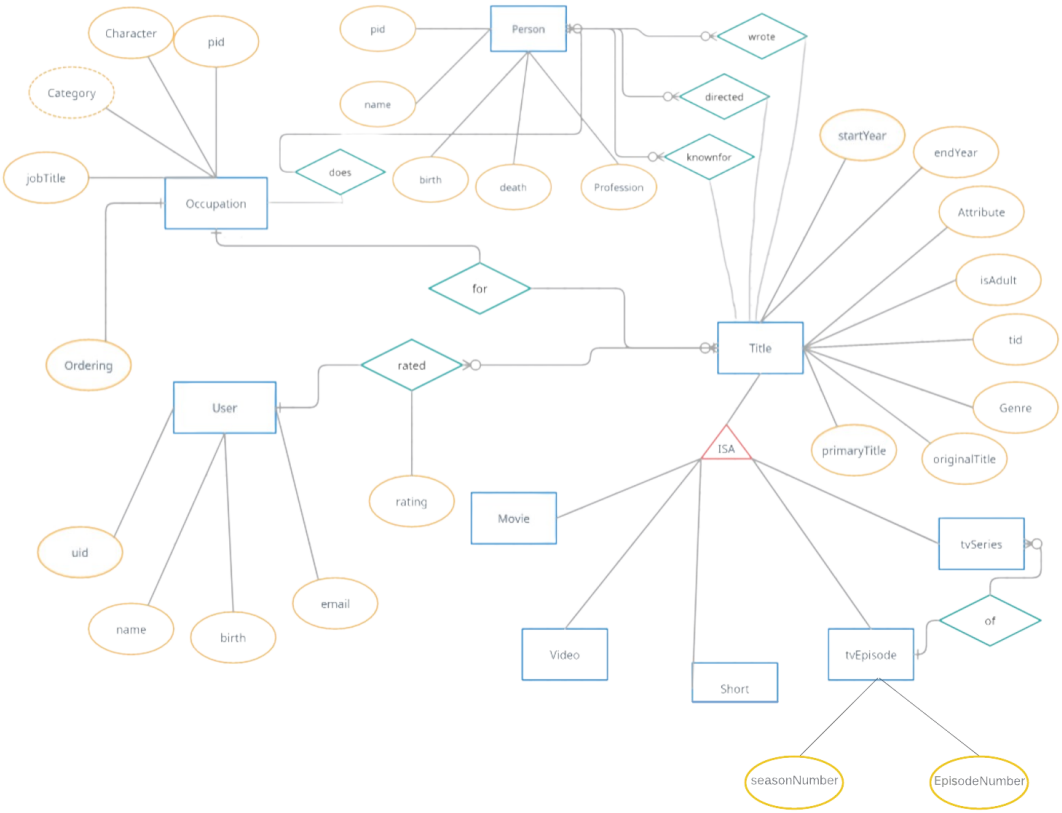
We initially began with the raw data set that we obtained from IMDB. The size of this data set is approximately 7 GB. We obtained 7 tables in .tsv files with the following schema description.



It is clear that the raw data that we have here has lots of redundancies, unwanted tables, non-atomic data types etc. We first drew an ER diagram for the database that we plan to build for the application. We went on to remove some unwanted data, i.e. tables that we do not require for the application we are meaning to build. We added some new tables for our implementation purpose, like tables containing users and rating information. Finally, we performed FD preserving normalization which has been described in section 3 below.

2 Entity-Relationship Diagram

Here is the entity relationship diagram that we intend to have for our database.



3 FD Preserving Normalisation Steps

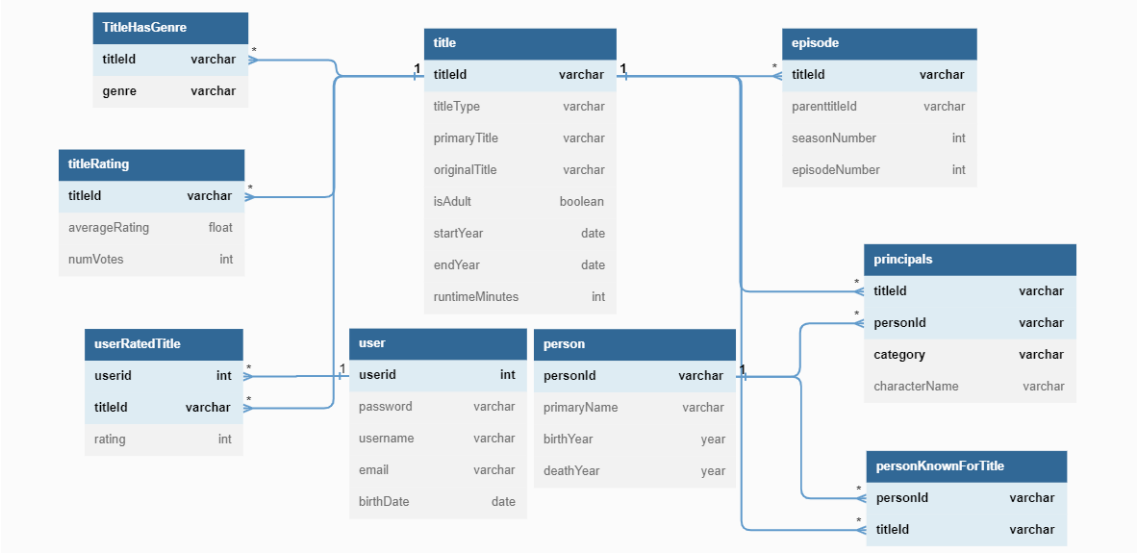
We performed the following normalization steps:

The raw database had multiple data redundancies and non-atomic attributes. To ensure data integrity and coherence in our database design, we used the following normalizations to ensure our database is in Boyce-Codd Normal Form.

- Originally the genre attribute in the 'title' table was a non-atomic varchar array. We normalized 'title' table and separated out the genre attribute to form another relation 'titleHasGenre' which has tuples of the form (title, genre).
- 'knownForTitles' was an array attribute in the original data, which we separated out into 'personKnownForTitle' relation. This relation contains (personId, titleId) tuples.
- We intend to store the ratings given by users to titles in the 'userRatedTitle' relation.

4 Relational Schema

Here is the relational schema for the final database we obtained after doing the normalisation steps. The attributes highlighted in bold are primary keys.



This diagram was made using an online tool DBDiagram.io.

5 Functional Dependencies

Following are the functional dependencies in the tables that we obtained after doing Functional Dependency Preserving Normalization.

5.1 title

titleid $\longrightarrow$ titletype primaryTitle originalTitle isAdult startYear endYear runtimeMinutes

5.2 user

userid $\longrightarrow$ password username email birthDate

5.3 person

personid $\longrightarrow$ primaryName birthYear deathYear

5.4 episode

titleid $\longrightarrow$ parenttitleid seasonNumber episodeNumber

5.5 principals

titleid personid category $\longrightarrow$ characterName

5.6 titleRating

titleid $\longrightarrow$ averageRating numVotes

5.7 userRatedTitle

userid titleid $\longrightarrow$ rating