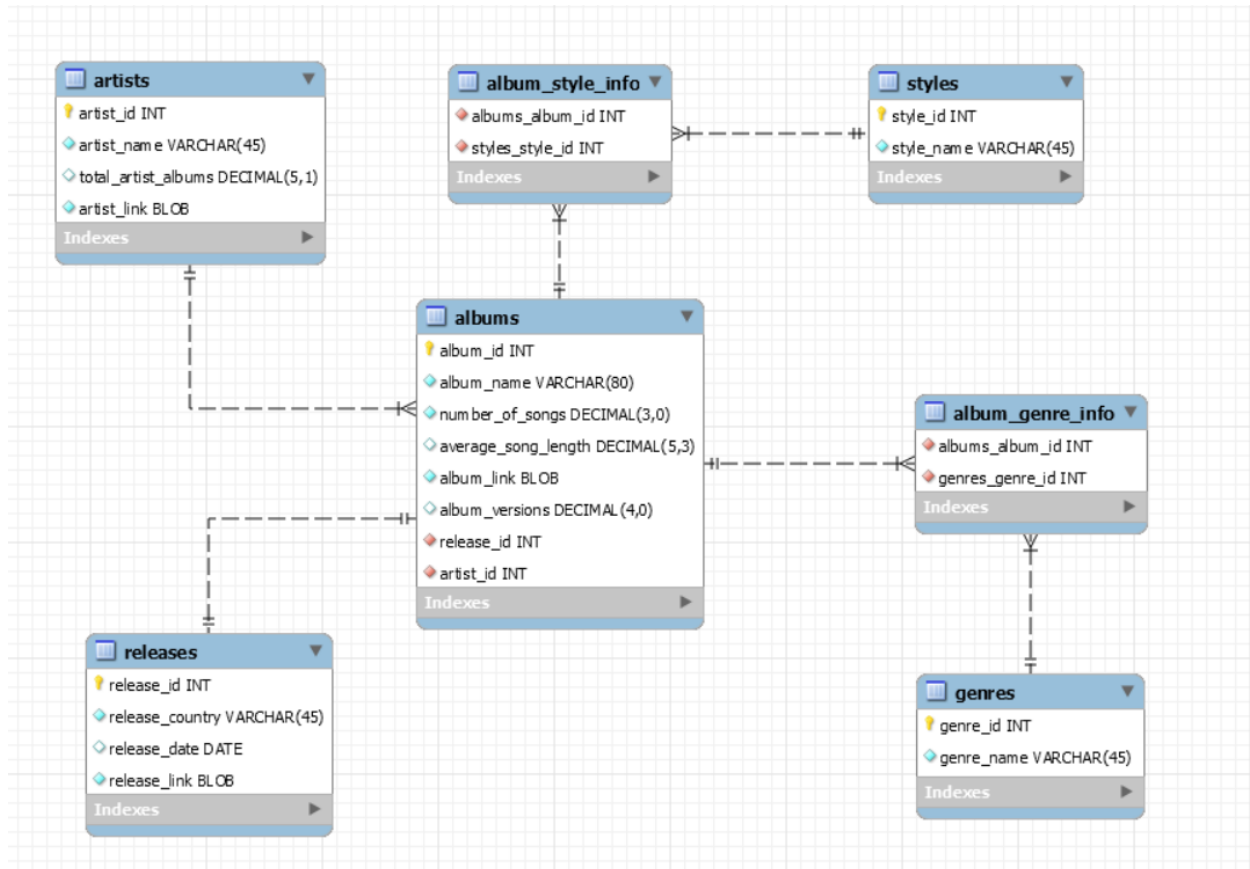Jason C. Bae

Project Proposal Report

Due: April 18th, 2021

ERD Diagram

Database Description

Discogs.com website provides a database of various albums for users to browse through and to make purchases for the ones they want. To improve user's experience, my database's primary objective is to display website information by separating the information under various entities to assist with organizations and to improve interactions. My secondary objective is to allow data manipulations for users to further explore their music taste, using the database.

Having flexibility in the dataset, the database allows the users to discover possible interests and/or explore other options relating to their music interests.

The database will be based on information retrieved from Kaggle.com that originally contained information about 1000 different albums from the Discogs.com website with 36 different attributes per album. From the source, my database will be focused on information from 20 different albums with 7 different entities and 2 of the entities serving as a linking table between the entities. These 20 albums will be pulled from the first 30 albums in the source that are in alphabetical order and average song length of the album is not a null value.

Sample Data Plan

As mentioned, my database will be drawing information about the database through Kaggle.com that contains a database with 1000 different albums with 36 different attributes. To gather more understanding about each album, the dataset has 3 different website links as attributes that are linked to either artist information, album information, or album information when it was first released for any need for manual data entry.

Compared to the original database, my database will be restructured to provide an easier understanding of the dataset, compared to the original database that only consisted of one entity with multiple attributes. My database will consist of various entities to allow an easier data manipulation to retract information, rather than limiting the capability to a data organization. Additionally, my database will introduce 4 new attributes to discover more about the information. The attribute is "release_date" and "style_name". The addition of these attributes will allow the users to interact with a full release information and music style information that may be necessary to explore their music tastes. These attributes will be inserted manually by me, using the website links provided for each album. Lastly, there will be some attributes excluded from the database to reduce ambiguity, confusion, and complexity. All attributes that divert away from meeting the secondary objective will be ignored. For example, other than release date, most of the date information will be unnecessary for exploring the user's music interests. Also, any sales and format information will not be useful in understanding the user's music preference. Even though finance and delivery of music are necessary to make a further judgment, the database's secondary goal does not explore the options that lead to any purchasing decisions.

Additionally, any other user's rating or user visits and statistics will also be excluded, due to ambiguous details that divert the users away from the secondary objective. From learning about popular music types, the users may develop biases and/or inclined to favor one song over another, which is contrary to the database intentions. Similarly, any recording company information or labels will be excluded, because they only include information that is arbitrary and adds complexity to the database. Also, similar to the user statistic information, the recording company can influence the users' preferences as big name companies can cause predetermined prejudices. Lastly, all genre type attributes become unnecessary to include as we plan to allocate an entire entity geared towards genre information about each album. By reducing complexity and reducing irrelevance, my database plans to contain attributes that are geared towards understanding the user's music taste with minimal distractions that influence user's preferences.

As shown by the ERD Diagram, my database will contain various relationships to fit my needs correctly. Table artists and albums will be connected as a one-to-many relationship. Despite having some albums sharing the same name, the database will be selective in which album it picks to establish a relational database that holds a relationship with each artist to publish multiple albums. Tables styles and genres will be connected to table albums as many-to-many relationships with linking tables album_style_info and album_genre_info, because each album can have multiple styles and genres while each genre or style can include multiple albums. Lastly, table releases will be connected to the table albums in an identifying, one-to-one relationship, because release information cannot exist without an album to have already existed. Furthermore, users will have access to various website links that are directly related to Discogs.com website that can provide more specific details about each album.

Progress Report

My team has accomplished various tasks since the completion of the proposal and had to make various revisions to meet primary/secondary objectives. Using my TA and AMP connections, we had to make various changes to how we distribute my entities. Initially, my original distribution of the database was exceeding the limit of my entity list. Thus, we had to reorganize my dataset to meet the limitations. During the separation, there were many issues to correctly normalize the given dataset. My original distribution of the dataset was insufficient

with dealing with multiple genres and multiple styles per album. Since each album can contain multiple genres and multiple styles, we struggled to correct and normalize to reduce duplicates of data. The original table had used truth and false methods to answer for multiple genres and styles, but the method made use of too many attributes and spaces. Thus, we had created new entities for genres and styles to allow better database organization.

Additionally, my database struggled with how to differentiate between foreign key, primary key, and composite keys. Thus, my AMP introduced us with feedback on how to correctly define each key and to display such an understanding onto the ERD diagram correctly since they had to be defined correctly per key separately. Outside connecting relationally, the database required separate definitions per key to uniquely identify each key per entity, so that each entity displayed the correct number of keys without establishing other relationships.

Changes from the initial Proposal

Compared to my original proposal, my database had undergone a major transformation, due to the misunderstanding of which attribute is available to develop my database. For example, my source did not contain much information about orders like card information, credit card type, and order number information and user account information like user interests and account number information, which resulted in deletion of such attributes. Thus, my database required a different goal than it had originally been planned for. Since the information we could work with varied, my goal had to be restated to meet the changes.

Furthermore, the restatement of my goal resulted in different numbers and types of entities. As my goal was more towards better user experience and to allow users to explore their interests, my database had to contain correct entities that addressed such aspects of the information to meet the goals. From understanding which information that we acquire and which information that we do not, we had to manually input for some aspects of the database, and we also had to exclude some information to reduce complexity and irrelevance.

Lastly, my database had to experience changes during the ERD Diagram development. As we explored more about the relationship understanding, we realized some of the attributes were not appropriate for some entities and had to be correctly allocated. As mentioned before,

we reduced some attribute information to reduce duplicates and make necessary adjustments to model correct normalization.

Plans for Remaining Work

  After establishing relational understanding, my database is left with the tasks of creating the actual database, manually inputting necessary information per album, and to accomplish view table understanding to display specific analysis understanding to meet the secondary goal. Since my secondary goal is geared towards exploring user's interests, we must establish view tables that display correct elements that may answer any questions that the user might have about their music interests. The view table must include any information that the user might require to better explore their music interests like which type of albums are in a certain genre, style, and format.