Final Project Proposal – Part IV

Assigned Date: 12/5/2023

Please add the following part to your final project proposal:

- You should generate a minimum of 100 data records for each table in a commonly used data format, such as .csv, .txt, etc. If there are tables that cannot provide up to 100 data records, please provide a reason.
- The data can be either real or synthetic. Please submit a brief report stating how you created these initial datasets. If it is a real dataset, please provide the sources of the data. If it is a synthetic dataset, please provide the source code used to generate that dataset.

- Customer table: The reason for having fewer than 100 customer records is due to the potential significant increase in the number of invoice items when adding more customers. This would exceed the desired scope for showcasing the consequences of our dataset.
- Employee table: Because the playlist comprises only 18 items, it is neither practical nor essential for the employee creating the playlist to include a total of 100 individuals.
- Genre table: Given the current existence of only approximately 25 music genres, the data volume will not reach 100 entries.
- MediaType table: The data volume will not reach 100 entries as there are only about a dozen formats for music files.
- Playlist table: Due to the difficulty of creating playlists, the data volume will not reach 100 entries.
- Publisher table: Today, the majority of music rights are held by these three music publishers, which is why these three companies have been selected to be used as publishers.

- The database we utilize was initially developed by Luis Rocha, a Canadian software developer. It was designed to represent a media store, with tables for artists, albums, and media tracks based on iTunes store libraries. Additionally, sales-related information, including customers, invoices, and employees, was randomly generated over a 4-year period. However, the dataset became outdated, prompting our decision to transition to an online streaming format. One of the key reasons for this change is the significant transformation in the music landscape, particularly with the emergence of streaming platforms like Spotify in 2008. To enhance the database, we introduced a new "PUBLISHER" table to specify the company that owns the reproduction rights to each song. We

leveraged Python to generate random outcomes for these companies and integrated them with the existing "TRACK" table. As a final improvement, we removed the "composer" attribute to streamline relations, as this information can now be accessed through the "ARTIST" table.