Católica
Universidade Católica de Brasília

Pró-Reitoria Acadêmica
Curso de Ciência da Computação
Atividade de Modelagem de Banco de Dados

ATIVIDADE EM GRUPO MBD

Grupo: Carlos Fernandes, Vitor Bittencourt, Wesley Moreira, Vinicius Lacerda
Professor: João Robson

1. Introdução

Os dados utilizados neste trabalho foram extraídos de uma base real de músicas mais ouvidas no Spotify, disponível no site Kaggle, intitulada "Spotify Most Streamed Songs". Essa base de dados contém informações detalhadas sobre as músicas que obtiveram mais streams na plataforma Spotify, incluindo dados como nome da música, nome dos artistas, número de streams, ano de lançamento, popularidade em diferentes playlists e plataformas, além de características musicais, como danceabilidade, energia e acústica.

O objetivo principal do trabalho foi utilizar essa base de dados para construir um banco de dados relacional utilizando o Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL. A partir desse banco de dados, foram criadas consultas SQL relevantes, que permitem explorar questões como quais são as músicas mais populares em determinados anos, os artistas com mais músicas de sucesso e identificar os anos com maior quantidade de músicas que ultrapassaram 1 milhão de streams.

Apresentação e Explicação do Processo de Extração e Importação dos Dados

O processo começou com a escolha dos dados, seguidos pela manipulação e formatação para adequação ao ambiente de banco de dados relacional.

1. Extração dos Dados

Inicialmente, os dados foram tratados no próprio site do kaggle, que disponibiliza uma aba de notebooks, onde é possível rodar scripts em python para visualização de dados. Utilizando a linguagem de programação Python e a biblioteca pandas, foi feita uma amostra aleatória de 70 linhas do arquivo CSV, para reduzir o volume de dados e focar no essencial para o desenvolvimento da modelagem e consultas SQL. Esse subset de 70 linhas foi salvo em um novo arquivo CSV.

```python
import pandas as pd

arquivo_csv = '/kaggle/input/spotify-most-streamed-songs/Spotify Most Streamed Songs.csv'
dados = pd.read_csv(arquivo_csv)
dados_selecionados = dados.sample(n=70, random_state=42)
print(dados_selecionados)
dados_selecionados.to_csv('linhas_selecionadas.csv', index=False)
```

Com esse novo arquivo CSV, utilizamos o ChatGPT para tratamentos dos dados, já que inserir dado por dado nas tabelas seria muito trabalhoso. Pedimos para que ele retornasse os dados já no padrão do SQL para inserção nas tabelas.

2. Preparação do Banco de Dados

O próximo passo foi a criação do banco de dados utilizando o SGBD MySQL. Foram definidas cinco tabelas principais, organizadas da seguinte maneira:

- Tabela Musicas: Contendo informações sobre o nome da música e a quantidade de streams.
- Tabela Artistas: Contendo os nomes dos artistas correspondentes às músicas.
- Tabela Lançamentos: Registrando os anos, meses e dias de lançamento das músicas.
- Tabela DesempenhoSpotify: Registrando o desempenho das músicas em playlists e charts do Spotify.
- Tabela DesempenhoApple: Registrando o desempenho das músicas em playlists e charts da Apple Music.

3. Inserção e Importação dos Dados

Os dados foram processados e ajustados para inserção no banco de dados MySQL. Uma vez que esses dados estavam limpos e preparados, foram gerados scripts SQL para a criação das tabelas e a inserção de dados. Esses scripts foram executados diretamente no MySQL, utilizando o MySQL Workbench

4. Consultas Realizadas

As consultas SQL foram projetadas para responder perguntas específicas relacionadas aos dados. As três consultas principais realizadas foram:

- Músicas com mais streams em determinado ano: Uma query para identificar as músicas mais populares em um ano específico, classificadas pelo número de streams.
- Artistas com mais músicas populares: Uma query que conta quantas músicas populares cada artista tem, com base nos dados de streams.
- Anos com mais músicas que ultrapassaram 1 milhão de streams: Uma query para identificar quais anos tiveram mais músicas com streams superiores a 1 milhão.

Consultas Realizadas

1. Quais são as músicas com mais streams em determinado ano?

```
SELECT M.track_name, M.streams, L.released_year
FROM Musicas M
JOIN Lancamentos L ON M.id_musica = L.id_lancamento
WHERE L.released_year = 2022  -- Substitua pelo ano que desejar
ORDER BY M.streams DESC;
```

Esta consulta retorna as músicas que tiveram o maior número de streams em um ano específico. A query faz uma junção (JOIN) entre a tabela Musicas e a tabela Lancamentos, associando cada música ao seu ano de lançamento. O filtro WHERE L.released_year = 2022 restringe os resultados para o ano escolhido (aqui, 2022). A ordenação dos resultados (ORDER BY M.streams DESC) garante que as músicas com mais streams apareçam primeiro.

Essa consulta é relevante para responder perguntas sobre tendências de consumo musical em determinado período. Por exemplo, identificar as músicas mais ouvidas de um ano pode ajudar a entender o impacto de lançamentos de grandes artistas ou identificar novos fenômenos musicais que dominaram as plataformas de streaming em um determinado ano.

2. Quais artistas têm mais músicas populares?

```
SELECT A.artist_name, COUNT(M.id_musica) AS num_musicas
FROM Artistas A
JOIN Musicas M ON A.id_artista = M.id_musica
GROUP BY A.artist_name
ORDER BY num_musicas DESC;
```

Essa consulta calcula o número total de músicas populares que cada artista tem na base de dados. Para isso, a query faz uma junção entre as tabelas Artistas e Musicas, associando as músicas aos respectivos artistas. A função de agregação COUNT() é utilizada para contar quantas músicas cada artista tem. O agrupamento é feito com GROUP BY A.artist_name, que agrupa as músicas por artista, e a ordenação ORDER BY num_musicas DESC garante que os artistas com mais músicas populares apareçam no topo.

Esta consulta é útil para determinar quais artistas têm um maior número de músicas populares no Spotify. Isso pode ser útil tanto para as gravadoras, que podem usar esses dados para medir a popularidade de seus artistas, quanto para profissionais de marketing e influenciadores do setor musical, que desejam entender quais artistas estão dominando as plataformas de streaming.

3. Em qual ano houve mais músicas com mais de 1 milhão de streams?

```sql
SELECT L.released_year, COUNT(M.id_musica) AS num_musicas
FROM Musicas M
JOIN Lancamentos L ON M.id_musica = L.id_lancamento
WHERE M.streams > 1000000
GROUP BY L.released_year
ORDER BY num_musicas DESC;
```

Essa consulta identifica os anos em que mais músicas tiveram mais de 1 milhão de streams. A query faz uma junção entre as tabelas Musicas e Lancamentos, e filtra as músicas que têm mais de 1 milhão de streams com WHERE M.streams > 1000000. Em seguida, as músicas são agrupadas por ano de lançamento (GROUP BY L.released_year), e a contagem de músicas por ano é calculada com COUNT(M.id_musica). A ordenação final ORDER BY num_musicas DESC garante que os anos com o maior número de músicas populares sejam listados primeiro.

Essa consulta é interessante para detectar os anos com maior produção e consumo de sucessos musicais no Spotify. Isso pode ajudar a identificar anos de grande relevância na indústria musical, possivelmente correlacionados com grandes lançamentos, mudanças nas tendências de consumo, ou até mesmo eventos externos que influenciaram o comportamento dos ouvintes.

3. Modelo Conceitual