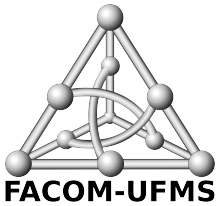
Big Data e Processamento de Dados: Uma Jornada para a Descoberta de Insights

Isadora S. da Silva

Mateus B. Cassiano

Trabalho de Conclusão de Curso

Orientação
Prof. Dra. Graziela Santos de Araújo



Faculdade de Computação
Universidade Federal de Mato Grosso do Sul

Campo Grande - MS
2024

Resumo

Este trabalho visa explorar como o uso de técnicas de *Big Data*, aliado ao processo de ETL (Extração, Transformação, Carga), pode transformar grandes volumes de dados em informações úteis para aprimorar a experiência do usuário. Com o aumento exponencial na geração de dados na era digital, impulsionado por dispositivos conectados, redes sociais e serviços de *streaming*, o tratamento adequado desses dados é essencial para gerar *insights* e auxiliar na tomada de decisões estratégicas. Os resultados demonstram a eficácia do processo na criação de um *dashboard* interativo, onde usuários podem explorar rankings de músicas e artistas com base em métricas, visualizar gráficos e personalizar filtros de acordo com suas preferências. Essa abordagem possibilitou destacar tendências entre diferentes gêneros musicais por meio da análise de dados. O trabalho conseguiu simular um cenário de *Big Data* e demonstrar como o uso estruturado de dados pode criar valor. O tratamento adequado de dados não apenas melhora a experiência do usuário, mas também tem o potencial de revolucionar o uso de informações em diversos setores.

Palavras-chave: *Big Data*, *ETL*, *Spotify*, *Visualização de Dados*.

Abstract

This work aims to explore how the use of Big Data techniques, combined with the ETL (Extract, Transform, Load) process, can transform large volumes of data into valuable insights to enhance the user experience. With the exponential growth of data generation in the digital age, driven by connected devices, social networks, and streaming services, the proper handling of this data is crucial for generating insights and supporting strategic decision-making. The results demonstrate the effectiveness of the process in creating an interactive dashboard, where users can explore music and artist rankings based on various metrics, view graphs, and customize filters according to their preferences. This approach made it possible to highlight trends between different musical genres through data analysis. The work successfully simulated a Big Data scenario and showed how the structured use of data can generate value. Proper data handling not only improves the user experience but also has the potential to revolutionize the use of information across various sectors.

Keywords: *Big Data*, *Data Visualization*, *ETL*, *Spotify*.

CAPÍTULO 1

Introdução

Com a popularização de *smartphones*, redes sociais, serviços de *streaming*, dispositivos integrados à Internet das Coisas (IoT) e o avanço da digitalização em atividades cotidianas, como transações financeiras e compras *online*, a geração de dados atingiu níveis sem precedentes em volume, velocidade e variedade. De acordo com estimativas [1], atualmente 402,74 milhões de *terabytes* de dados são gerados diariamente, o que equivale em até 147 *zettabytes* por ano. Esses dados, provenientes de diferentes fontes e formatos, como textos, imagens, vídeos e interações digitais, possuem o potencial de serem reciclados para agregarem benefícios.

A frase *"Data is the new oil. Like oil, data is valuable, but if unrefined it cannot really be used"* popularizada pela Forbes [2], destaca a importância estratégica dos dados, comparando-os a um recurso valioso como o petróleo que, quando devidamente refinado, pode gerar grande valor econômico. Nesse contexto, a exacerbada quantidade de dados da era digital pode ter uma influência econômica significativa, desde que seja processada adequadamente para obter *insights* valiosos.

Diante disso, originou-se o conceito de *Big Data*, que engloba conjuntos de dados que são muito complexos para serem gerenciados e analisados por ferramentas de dados tradicionais [3]. Sendo assim, *Big Data* é caracterizado pelo alto volume, velocidade e variabilidade dos dados gerados, como também pela veracidade e valor de suas informações.

Exemplificando a magnitude do *Big Data* à disposição, plataformas como *Google* lidam com cerca de 20 *petabytes* de dados todos os dias, resultantes das suas aproximadamente 3,5 bilhões de pesquisas diárias, enquanto em um dia, o *YouTube* transmite aproximadamente 720.000 horas de vídeos, o que leva a produção de 4,3 *petabytes* distribuídos entre diferentes resoluções e formatos, variando desde 144p a 4K, cada uma com seu respectivo consumo de dados [4].

Como mencionado, o avanço tecnológico vivenciado nas últimas décadas propiciou um cenário favorável para o desenvolvimento do *Big Data*, pois a democratização digital impulsiona a geração de dados por pessoas e dispositivos. No entanto, o montante de dados em si não gera informações úteis se não for analisado corretamente. Desta maneira, o processo de Extração, Transformação e Carga (ETL - do inglês *Extract,*

Transform and Load) surge como uma técnica para melhor preparar os dados gerados para análises futuras.

Considerando este cenário e pensando em explorar mais o assunto, buscou-se neste trabalho o desenvolvimento de um processo para simular um *Big Data* e conhecer todas as etapas para obtenção dos dados para composição do mesmo, bem como tratamento e análise dos dados para promover geração de informação de valor e, com isso, a melhor tomada de decisões por parte dos detentores destas informações.

Dessa forma, o objetivo geral deste trabalho é utilizar uma API do *Spotify* para obter dados, aplicar o processo ETL e realizar análises baseadas em características específicas, aprimorando a experiência do usuário.

Este desenvolvimento envolve o levantamento bibliográfico sobre os conceitos aplicados, como *Big Data* e ETL. Em seguida, a montagem de um *framework* ETL, escolhendo quais ferramentas serão utilizadas, para então realizar a implementação prática deste trabalho ao coletar os dados do *Spotify*, processá-los e a partir disso, montar um *dashboard* explicativo demonstrando um caso de uso para um usuário, obtendo sugestões musicais.

O texto deste trabalho está organizado como segue. O capítulo 2 descreve o conceito dos temas abordados de *Big Data*, ETL, visualização de dados, e os trabalhos relacionados que auxiliaram no entendimento da problemática. O capítulo 3 apresenta a metodologia utilizada para o desenvolvimento deste estudo. Os resultados obtidos são mostrados e analisados no capítulo 4. Por fim, o capítulo 5 apresenta as conclusões obtidas a partir deste estudo e propostas para trabalhos futuros.

CAPÍTULO 2

Contextualização

Com a integração crescente de dispositivos conectados à internet, a popularização da Internet das Coisas (IoT) e o avanço das plataformas digitais, a era da digitalização é marcada pela geração exponencial de dados. Este cenário cria uma escala e complexidade que caracteriza o fenômeno conhecido como *Big Data*. Nas próximas seções, será utilizado o temo *dataset* para se referir a uma base de dados.

2.1 Big Data

Big Data se refere a conjuntos de dados com características que não viabilizam serem extraídos, gerenciados ou processados em um tempo razoável por ferramentas tradicionais de gerenciamento de dados [3].

As primeiras definições para considerar uma coleção de dados como *Big Data* se baseiam no conceito de "3 V's"proposto por Laney em 2001 [5] (apud [6]).

O primeiro V, **volume**, refere-se à grande quantidade de dados gerados e armazenados. Esta característica pode variar consideravelmente dependendo de diversos fatores, pois, o que é considerado uma grande quantidade de dados para uma organização pode não ser para outra. Além disso, as tecnologias de armazenamento estão em constante evolução, de modo que o que era um desafio significativo para armazenamento no passado pode não ser mais um problema no futuro [7].

O segundo V, **velocidade**, diz respeito à rapidez em que esses novos dados são gerados e exigem processamento, criando a necessidade de sistemas capazes de lidar com fluxos de dados em alta velocidade.

Já o terceiro V, **variedade**, aborda a diversidade na formatação desses dados, uma vez que diferentes fontes podem gerar informações em formatos variados, como estruturados, semi-estruturados ou não estruturados.

Com o passar do tempo o conceito de *Big Data* passou a abranger "5 V's", de maneira que enquanto a definição de Laney foca em definir pontos estruturais e operacionais dos dados, os dois V's subsequentes enfatizam sua utilidade prática e na aplicação [8].

Sendo assim, o quarto V, **veracidade**, aborda a importância da confiabilidade em um banco de dados. Dados imprecisos ou incompletos podem levar a decisões erradas e resultados enganosos. Por conta disso é imperativo assegurar que os dados sejam precisos.

Por fim, o quinto V, foca no objetivo real de trabalhar com *Big Data*, obter **valor**. *Datasets* repletos de dados brutos são inúteis se não forem devidamente processados visando extrair informações que possam auxiliar análises e tomadas de decisões estratégicas trazendo benefícios às empresas e ao usuário final.

Nesse contexto, para auxiliar a compreensão do processo de transformação dos dados, usa-se o modelo DIKW (*Data, Information, Knowledge, Wisdom*), no qual é proposto uma hierarquia que ajuda a entender como os dados podem evoluir para se tornarem sabedoria. Cada nível representa um estágio progressivo de complexidade e valor, em que os dados brutos são transformados em algo significativo e útil [9].

Figura 2.1: Hierarquia do modelo DIKW. Fonte: Adaptada de [10]

O primeiro nível da hierarquia é composto pelos **dados**, entidades brutas e descontextualizadas, que não possuem significado por si só. Eles podem ser números, símbolos ou fatos simples, mas que não oferecem utilidade prática em seu estado original. Para que se tornem valiosos, os dados precisam ser organizados e analisados, então são o ponto de partida de todo o processo de transformação.

O segundo nível é a **informação**, que surge quando os dados são processados, organizados ou estruturados de modo a adquirirem significado em um contexto específico. Enquanto os dados são insumos brutos, a informação pode ser compreendida e interpretada. Ela transmite algo significativo e ajuda na tomada de decisões.

No terceiro nível está o **conhecimento**, que é a aplicação da informação. O conhecimento envolve interpretar a informação e usá-la de maneira prática, incorporando experiências, habilidades e compreensão mais profunda. Ele permite que as pessoas tomem decisões informadas e resolvam problemas com base na análise da informação disponível.

No topo da hierarquia encontra-se a **sabedoria**, que representa a aplicação crítica e ética do conhecimento. A sabedoria vai além de simplesmente utilizar o conhecimento para resolver problemas; ela envolve uma visão mais profunda e holística, levando em consideração as implicações a longo prazo das decisões e ações. Aplicar a sabedoria

requer discernimento e a capacidade de tomar decisões ponderadas com base em valores éticos.

Diante disso, aplicando técnicas de *Big Data* junto a modelos de negócio, as empresas podem alcançar diversos benefícios, como [7] [11]:

- **Eficiência operacional:** análises de *Big Data* podem identificar fragilidades nos sistemas operacionais, além de montar planos que resultam em redução de custos, aumento de produtividade e melhoria de performance;
- **Desenvolvimento inovador:** o *Big Data* pode ser alinhado a modelos preditivos, de modo a antecipar movimentos de mercado e permitindo que empresas estejam preparadas para atender demandas, obtendo vantagens competitivas;
- **Melhor atender expectativas do usuário:** com *Big Data*, os dados que os clientes fornecem às empresas permitem criar experiências de uso customizadas para cada indivíduo e ofertas personalizadas.

A *Netflix* é um dos maiores exemplos de empresa orientada por dados. Utilizando abordagens *Big Data* a partir dos dados coletados sobre comportamentos de seus usuários, como séries e filmes assistidos, horários de visualização, pausas e até abandonos, ela conseguiu revolucionar a indústria do entretenimento e se tornar a líder no setor de *streaming* digital. Em contrapartida, sua empresa rival, *Blockbuster*, protagonista no setor de locação de filmes, não adaptou-se às inovações tecnológicas e à cultura orientada por dados, de modo que não se adequou às mudanças de comportamentos de seus consumidores e acabou anunciando falência em 2013 [12].

O uso adequado de *Big Data* para o crescimento da *Netflix* é destacado pelo seu sistema de recomendação, que utiliza algoritmos de *Machine Learning* para realizar análises preditivas para cada perfil cadastrado. Para isso, utiliza-se um método híbrido de filtragem, baseado em: (i) filtragem baseada em conteúdo, que analisa os programas e filmes já consumidos por um usuário e identifica padrões em características como gênero, atores e temas, sugerindo opções semelhantes; e (ii) filtragem colaborativa, que utiliza o comportamento de outros usuários com perfis parecidos para recomendar títulos, aumentando ainda mais a precisão das sugestões [13] [14].

O sistema de recomendação também abrange estratégias visuais para engajar os consumidores, indo além da simples sugestão de títulos. O catálogo da plataforma conta com diversas artes que abrangem temas, e até mesmo atores, diferentes para cada obra, como pode-se observar na Figura 2.2 para a série *Stranger Things*, de modo a aumentar a personalização da experiência de cada assinante [13] [14].

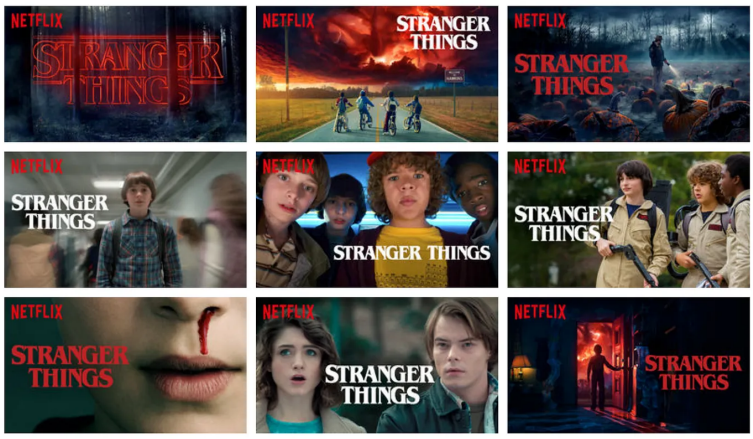


Figura 2.2: Diferentes artes para a série *Stranger Things*. Fonte: [15]

Note o exemplo da Figura 2.3, no caso de um telespectador que tenha um histórico de preferências para filmes da Uma Thurman, então o filme *Pulp Fiction* será recomendado com a capa personalizada destacando a atriz, diferente de um cliente que tenha preferências em filmes do John Travolta, o que aumenta a probabilidade de escolha em assistir este título. Essa customização visual, somada ao algoritmo preditivo, ajuda a manter os assinantes engajados e a aumentar o tempo que passam na plataforma [15].

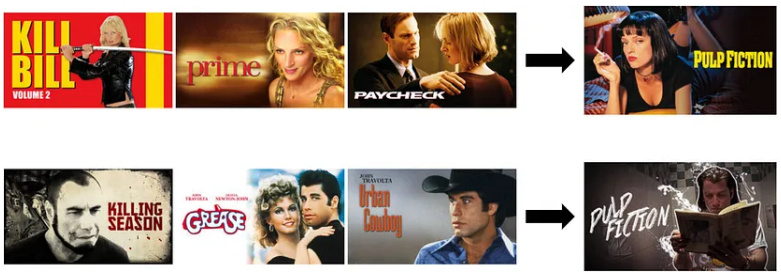


Figura 2.3: Diferentes artes para o filme *Pulp Fiction*. Fonte: [15]

Logo, dada a grande quantidade de oportunidades que a análise de dados proporciona, é imprescindível lidar com o *Big Data* corretamente para explorar seus benefícios. Na próxima subseção, será explorada uma dessas estratégias, conhecida como *Extract, Transform and Load* (ETL).

2.2 ETL

O processo de ETL é uma abordagem eficiente para garantir integração e gerenciamento de dados, uma vez que sua estrutura permite preparar um grande volume de dados de diversas fontes para análises posteriores. Ele organiza, limpa e padroniza dados brutos de acordo com regras de negócio para melhor atender seu contexto corporativo, garantindo que estes sejam consistentes e adequados para futuras consultas e aplicações [16]. Dessa forma, o ETL é fundamental em ambientes que lidam com o *Big Data*. Nas próximas subseções, serão apresentadas as fases do processo ETL.

2.2.1 Extração

A etapa de **extração** é a primeira fase do processo ETL, em que os dados brutos são extraídos de suas fontes que podem ser APIs, redes sociais, bancos de dados relacionais, e até dados gerados por dispositivos IoT (Internet das Coisas) [17]. Dada a natureza heterogênea das fontes, o formato dos dados pode variar entre: (i) estruturados, como tabelas relacionais; (ii) semi-estruturados, como arquivos JSON e XML; (iii) não estruturados, como imagens, vídeos e textos.

É crucial nesse processo que os dados extraídos se mantenham íntegros, evitando perdas ou corrupções. Além disso, é importante que a extração não impacte o desempenho dos sistemas fontes, o que poderia acarretar em perdas significativas à companhia.

Note que atualizações continuam a acontecer nos sistemas fonte após a extração no processo, de modo que dados já extraídos podem ficar desatualizados. Sendo assim, existem três métodos de manter os registros em dia, segundo Singh [18]:

- **Extração parcial com notificação de atualização:** o sistema de origem informa quando um dado é modificado, de modo que apenas registros atualizados são re-extraídos;
- **Extração parcial sem notificação de atualização:** sem o sistema de notificação, é então necessário que investigações por alterações sejam feitas regularmente em intervalos de tempo, como semanalmente. Caso um registro modificado seja encontrado, ele é re-extraído;
- **Extração completa:** em situações em que não é possível verificar por atualizações diretamente com o sistema de origem, todo o *dataset* é extraído novamente para se comparar com os dados extraídos previamente em busca de modificações. Contudo, por ser uma operação muito custosa, não recomenda-se para conjuntos grandes de dados.

Após a extração, os dados podem ser temporariamente salvos em uma *Data Staging Area* (DSA). Este armazenamento intermediário permite que os dados obtidos possam ser validados, mas também garante que caso ocorra uma falha nas etapas seguintes, todo o processamento já realizado não precise ser refeito [18].

2.2.2 Transformação

Em seguida, os dados extraídos em sua forma bruta são enriquecidos durante a etapa de **transformação**, por meio de técnicas de refinamento de ajuste, visando agregar valor para utilizá-los em análises futuras.

Existem diversas transformações que podem ser aplicadas nesta etapa, as principais, segundo [16] e [19], são:

- **Limpeza de dados:** consiste na remoção de dados duplicados, correção de erros e preenchimento de valores ausentes;
- **Agregação e enriquecimento:** dados podem ser agregados ou enriquecidos com informações adicionais, como a combinação de dados de diferentes fontes para criar um conjunto mais completo e informativo;

- **Aplicação de regras de negócio:** durante a transformação, regras específicas da organização podem ser aplicadas aos dados, como cálculos de métricas ou a categorização de informações;
- **Conversão de dados:** dados podem ser convertidos de um formato para outro, como a conversão de dados de XML para JSON ou de tabelas relacionais para formatos de dados NoSQL;
- **Transformação e normalização:** dados são convertidos para um formato consistente. Isso pode envolver a transformação de formatos de data, a padronização de unidades de medida ou a normalização dos dados.

Ainda sobre a Normalização, refere-se a um processo estruturado em um conjunto de regras conhecidas como Formas Normais. A Primeira Forma Normal (1FN) garante que todos os dados sejam atômicos. A Segunda Forma Normal (2FN) elimina dependências parciais, o que significa que todos os atributos não-chave devem depender totalmente da chave primária como um todo, e não apenas de partes dela. Já a Terceira Forma Normal (3FN) remove dependências transitivas, assegurando que todos os atributos dependam diretamente e exclusivamente da chave primária [20] [21].

2.2.3 Carga

Por fim, na etapa de **carga**, os dados são carregados para seu destino final. Esta etapa é essencial para garantir que os dados estejam disponíveis para serem consultados, analisados e utilizados pelas aplicações empresariais e ferramentas analíticas.

Há diferentes formas de realizar o carregamento de registros para o destino final, conforme [18] e [19]:

- **Carregamento inicial:** é o carregamento completo de todos os dados ao destino final;
- **Carregamento incremental:** são cargas periódicas que trazem registros modificados entre os sistemas origem e destino, mantendo-o atualizado;
- **Atualização completa:** consiste em excluir todo o conteúdo de ao menos uma tabela, substituindo-o com uma carga de dados novos.

Há diversas maneiras de desenvolver uma estrutura de ETL, e a escolha da abordagem mais adequada depende de critérios como escalabilidade, velocidade de processamento, custo-benefício e eficiência operacional [22]. Com base nesses fatores, é possível optar por diferentes tipos de ferramentas que se alinhem às necessidades específicas do projeto, sendo algumas dessas [23]:

- **Ferramentas *Cloud***: são oferecidas por provedores de nuvem como AWS, *Azure* e *Google Cloud*. Elas são escaláveis e flexíveis, permitindo que empresas ajustem rapidamente a capacidade de processamento conforme necessário. Exemplos incluem *AWS Glue* e *Azure Data Factory*;

- **Ferramentas *Open Source***: oferecem uma alternativa robusta e econômica às ferramentas proprietárias. Com essas opções, as empresas podem construir, adaptar e escalar suas *pipelines* de dados com total controle sobre o código e a infraestrutura. Exemplos de ferramentas de código aberto incluem *Apache NiFi* e *Hadoop*;
- **Ferramentas *Code-Based***: proporcionam a liberdade de poder construir soluções ETL totalmente personalizadas, sendo possível criar fluxos de trabalho complexos e otimizados para tarefas específicas. Exemplos de ferramentas baseadas em código são a linguagem de programação *Python* junto a bibliotecas *Pandas* e *PySpark*.

2.3 Visualização de Dados

A visualização de dados tem um papel fundamental na comunicação dos resultados extraídos do processo de ETL, pois transforma dados complexos e estruturados em representações visuais que podem ser facilmente interpretadas por públicos diversos [24]. A escolha de uma ferramenta de visualização envolve diversos fatores, como facilidade de uso, integração com fontes de dados, customização, escalabilidade, custo e suporte [25].

O *Grafana* é uma ferramenta de código aberto usada para visualização e análise de dados em tempo real. Ele permite a criação de painéis interativos e gráficos dinâmicos a partir de diversas fontes de dados, como bancos de dados, serviços de monitoramento, e APIs [26]. Outras ferramentas com características semelhantes incluem *Tableau*, *Power BI* e *Qlik Sense*.

2.4 Trabalhos Relacionados

Alguns trabalhos relacionados serviram como fonte de inspiração para este estudo. De maneira geral, no que tange temas de ETL e *Big Data*, discute-se como um processamento eficiente de dados pode trazer benefícios a diferentes setores, como pode-se ver a seguir:

- ***Towards a Big Data Analytics Framework for IoT and Smart City Applications***: este estudo apresenta uma estrutura para a análise de *Big Data*, enfocando a integração de métodos de coleta, armazenamento e análise de dados. Ele combina processamento em lote e em fluxo, visando desenvolver soluções eficazes para cidades inteligentes [27];
- ***Building an Efficient ETL/ELT Process for Data Delivery***: discute as melhores práticas para construir processos ETL eficientes, com foco na otimização do pipeline de dados para garantir entregas rápidas e seguras, especialmente em ambientes de grandes volumes de dados [28].

Além disso, uma vez que este trabalho utiliza dados musicais, os artigos a seguir são relevantes por empregar a ferramenta de API do *Spotify* na coleta de dados sobre músicas:

- ***Exploring the world of music with the Spotify API and Postbot***: com o intuito de obter recomendações de músicas de hip hop acústicas, configurou-se a API para extrair dados de acordo com as métricas desejadas. Em seguida, foi utilizada a ferramenta *Postbot* para gerar uma interface visual dos resultados obtidos [29];
- ***Using the Spotify API for data-driven analysis of my playlists***: utilizou-se a API do Spotify para realizar uma análise orientada por dados de *playlists* pessoais, observando padrões de escuta e métricas musicais relevantes [30] [31];
- ***Using Spotify data to predict which* "Novidades da semana" *songs would become hits***: os dados coletados foram empregados junto a modelos preditivos de *Machine Learning* para tentar prever quais dos lançamentos da semana se tornariam *hits* [32].

Sendo assim, estes estudos nos auxiliaram a entender como a API do *Spotify* funcionava, além de compreender a vasta possibilidade de análises que podem ser feitas a partir dos dados que o *Spotify* disponibiliza via API [33].

CAPÍTULO 3

Metodologia

Este trabalho segue uma abordagem exploratória e descritiva. A natureza exploratória se justifica pelo conhecimento superficial que tínhamos sobre o tema antes da pesquisa, o que nos levou a buscar um entendimento mais aprofundado em diversos âmbitos do ramo de dados, como extração, manipulação, armazenamento e visualização. Já o caráter descritivo, buscamos nos aprofundar no assunto, e isto se dá a partir do objetivo geral deste trabalho de realizar análises musicais de acordo com determinadas características obtidas de uma amostra de dados.

A pesquisa foi inicializada a partir do estudo de trabalhos relacionados que se aprofundam em técnicas de tratamento de *Big Data* com foco central no processo ETL. Em seguida, com um entendimento mais fundamentado desta técnica, deu-se início ao planejamento da estruturação do *framework* de ETL.

Como este trabalho requer a extração de dados de uma fonte de origem, foi feita uma pesquisa acerca das possibilidades de se utilizar como fonte de extração. Dentre as opções foi escolhido o *Spotify*, por fornecer uma API pública. Essa escolha se deu devido à capacidade da API de fornecer informações detalhadas sobre faixas e álbuns, como título, duração, artistas, além de valores numéricos para diferentes métricas musicais. Essas propriedades são fundamentais para a análise e recomendação de músicas, permitindo um entendimento mais preciso das características musicais.

Diversas métricas são disponibilizadas pela API do *Spotify*, a análise de dados deste trabalho se concentrou nas seguintes [33]:

- **Dançabilidade:** descreve quão adequada é uma faixa para dançar, com base em uma combinação de elementos musicais, incluindo andamento, estabilidade de ritmo, força de batida e regularidade geral;
- **Energia:** representa uma métrica perceptiva de intensidade e atividade. Normalmente, as faixas energéticas parecem rápidas, altas e barulhentas. Por exemplo, o *Death Metal* tem alta energia, enquanto um prelúdio de Bach tem pontuação baixa na escala. As características perceptivas que contribuem para este atributo incluem faixa dinâmica, volume percebido, timbre, taxa de início e entropia geral;
- **Acusticidade:** uma métrica de confiança para saber se a faixa é acústica;

- **Vocalização:** detecta a presença de palavras faladas em uma faixa. Quanto mais exclusivamente falada for a gravação (por exemplo, *talk show*, audiolivro, poesia), maior será essa métrica;
- **Vivacidade:** detecta a presença de público na gravação. Valores mais altos de vivacidade representam uma probabilidade maior de que a faixa tenha sido tocada ao vivo;
- **Instrumentabilidade:** prevê se uma faixa não contém vocais. Os sons "ooh"e "aah"são tratados como instrumentais neste contexto. Faixas de rap ou palavras faladas são claramente vocais. Quanto maior o valor dessa métrica, maior será a probabilidade da faixa não conter conteúdo vocal;
- **Valência:** uma métrica que descreve a positividade musical transmitida por uma música. Faixas com alta valência soam mais positivas (por exemplo, feliz, alegre, eufórica), enquanto faixas com baixa valência soam mais negativas (por exemplo, triste, deprimida, irritada);
- **Volume geral:** a intensidade sonora geral de uma faixa em decibéis (dB), úteis para comparar a intensidade sonora relativa das faixas. A intensidade sonora é a qualidade de um som que é o principal correlato psicológico da força física (amplitude);
- **Tempo musical:** o tempo geral de uma faixa estimado em batidas por minuto (BPM). Na terminologia musical, o tempo é a velocidade ou o ritmo de uma determinada peça e deriva diretamente da duração média da batida.

Além disso, foi necessário garantir que o conjunto amostral de dados fosse diversificado. Isso se deu para evitar que, ao utilizar critérios específicos como "dançabilidade" durante as análises, algum gênero predominasse. Logo, com o intuito de assegurar que a média entre as métricas de gêneros distintos não fossem muito distantes, foi feito um script em *Python* no *Google Colab* utilizando bibliotecas como *Matplotlib*, *NumPy* e *Spotipy*, que recebe métricas de músicas contidas em *playlists* de diferentes gêneros, retiradas do *Spotify*. Após avaliar os resultados obtidos em cada gênero testado, foram escolhidos os seguintes: clássica, EDM, pagode, pop, rap e rock, pois foram os que apresentaram um bom índice de diversidade, como mostrado na Figura 3.1.

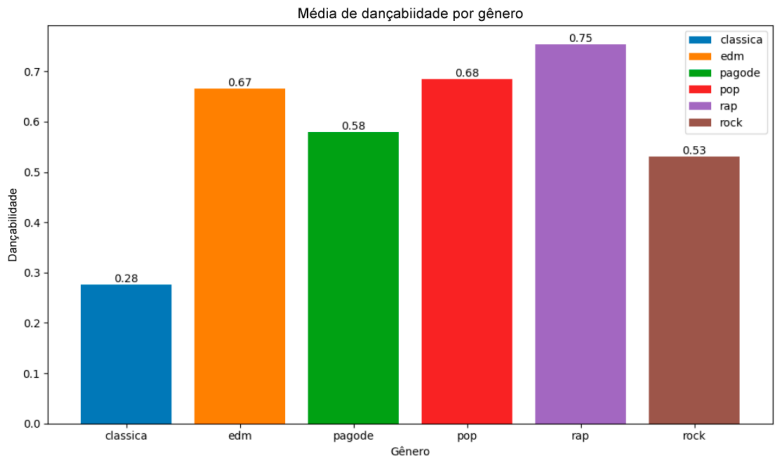


Figura 3.1: Média de dançabilidade para os gêneros escolhidos.

Dessa forma, uma vez definido que os dados seriam extraídos do *Spotify* e manipulados via código, o projeto passou a ser desenvolvido utilizando as seguintes tecnologias:

- **Ambiente de execução:** *notebook* com *Windows 11* (*Intel Core i7-11800H*, 32 GB de RAM, SSD NVMe);
- **Linguagem de programação:** *Python* 3.12;
- **Bibliotecas auxiliares:** *Spotipy*, Pandas;
- **Banco de dados relacional:** *PostgreSQL* 16;
- **Ferramenta de visualização de dados:** *Grafana* OSS 11.3.

A primeira etapa realizada foi a **extração** via um script em *Python* junto à biblioteca *Spotipy*, que recebe a ID do *Spotify* de 5 *playlists* musicais de gêneros distintos. Os dados são extraídos de duas maneiras: a primeira consiste em percorrer as *playlists* e extrair os dados para todas as músicas nelas contidas, já a segunda forma se baseia em encontrar o álbum em que cada canção contida na playlist foi lançada e extrair todos os dados das outras músicas presentes neste mesmo álbum.

Ao fim da extração, o conteúdo obtido foi armazenado em diversos arquivos JSON separados por gêneros, como `album_tracks_playlists_*`, `albums_playlist_*`, `features_album_tracks_playlists_*`, `features_playlist_*` e `playlist_*`, onde o símbolo "∗"representa cada um dos 5 gêneros utilizados como base. Cada arquivo possui níveis de complexidade diferentes, pois englobam múltiplos graus de elementos interconectados, como listas e dicionários aninhados.

A Figura 3.2 exemplifica a estrutura e a conexão entre os arquivos extraídos por meio do campo *id*, representadas pelas linhas completas no diagrama. No caso dos arquivos `playlist_*.json`, além de conterem dados sobre cada faixa, existem campos destacados em azul que correspondem a estruturas mais complexas que encapsulam informações mais detalhadas, como `album` e `artists`, que por sua vez podem conter subestruturas, como `images`. Esse mesmo padrão hierárquico e interconectado é observado em outros arquivos, como mostra a Figura 3.3, reforçando a modularidade e a complexidade dos dados extraídos.

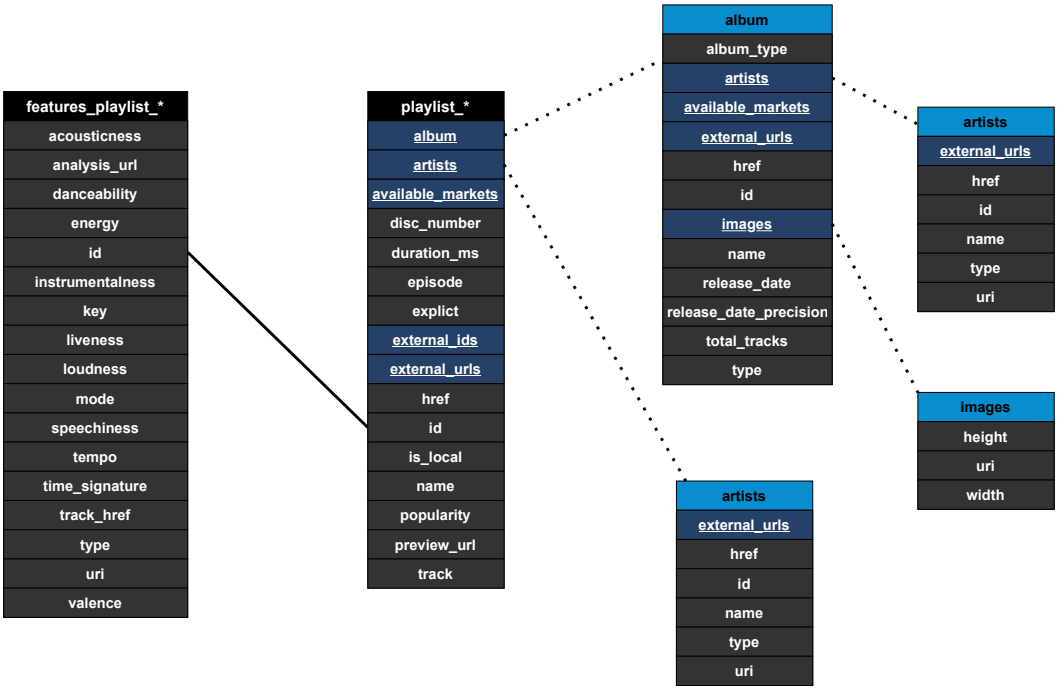


Figura 3.2: Organização dos dados extraídos da primeira forma

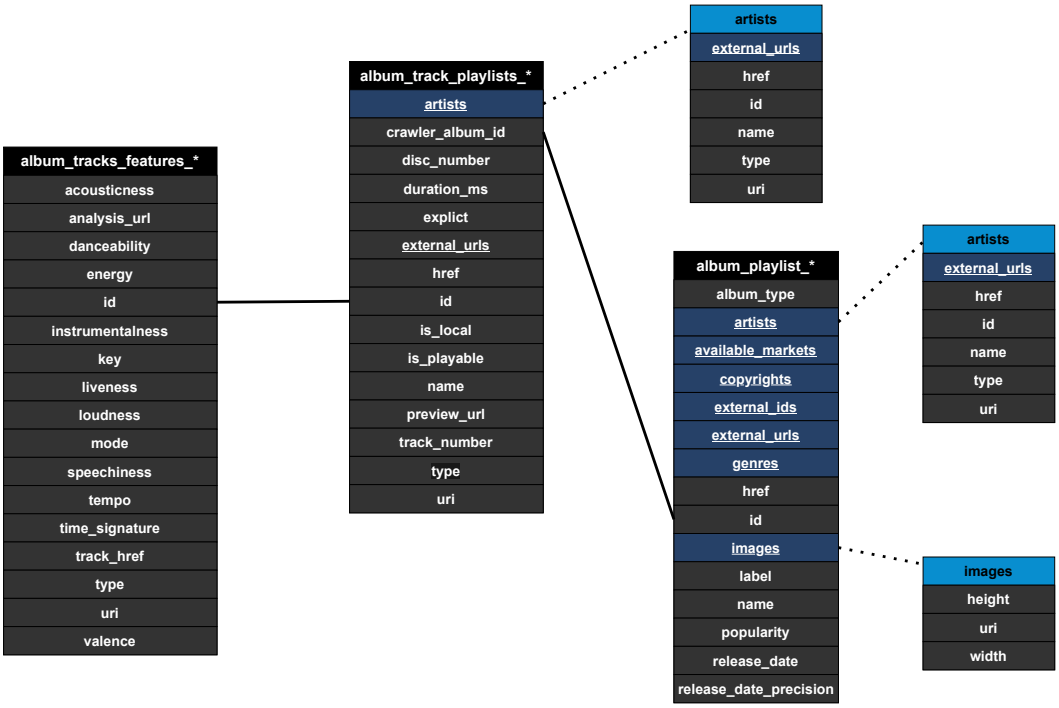


Figura 3.3: Organização dos dados extraídos da segunda forma

Diante disso, é importante utilizar técnicas de manipulação de dados durante a etapa de **transformação** para organizá-los e simplificá-los. A primeira etapa desse processo consiste em simplificar a estrutura dos dados, reorganizando os campos para que contenham apenas valores simples, eliminando listas e dicionários aninhados. Para

isso, foi desenvolvido um script em *Python* que percorre recursivamente esses campos, extraindo os dados aninhados e convertendo-os em tabelas independentes.

Na estrutura original, os campos foram substituídos por dicionários contendo apenas identificadores únicos (IDs), mantendo as informações compactadas pelo maior tempo possível. Essa abordagem reduz a redundância durante o processamento e facilita a manipulação dos dados ao longo das etapas subsequentes. Como resultado, novos arquivos são gerados para armazenar as tabelas independentes, enquanto os arquivos originais são atualizados com os campos reorganizados.

Com o grau de complexidade estrutural reduzido, o próximo passo de modelagem foi mesclar e concatenar os dados, uma vez que por conta das duas formas de extração, informações semelhantes foram distribuídas em arquivos diferentes, como o caso das métricas das músicas que estão espalhadas nos arquivos `features_playlist_*.json` e `features_album_tracks_playlists_*.json`, resultando na estrutura mostrada pela Figura 3.4. Além disso, é importante garantir a remoção de itens duplicados nesta fase, uma vez que itens podem ter sido extraídos mais de uma vez, e essas duplicidades não acrescentam valor à análise.

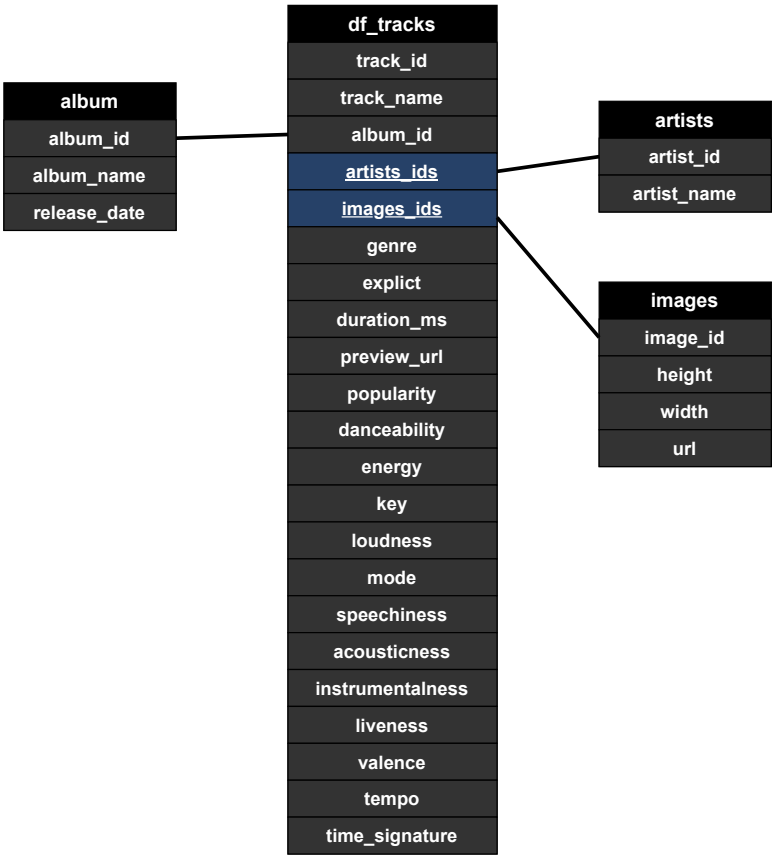


Figura 3.4: Organização dos dados mesclados

Em seguida, para que os dados fiquem prontos para análise, é importante garantir a normalização das tabelas. Conforme exibe a Figura 3.5, esse processo envolveu assegurar que as tabelas contivessem apenas valores atômicos, eliminando elementos como

listas nos campos *artists_ids* e *images_ids* e evitando a redundância de informações. Dado que as relações entre músicas e artistas, bem como entre músicas e imagens, são do tipo muitos-para-muitos, foram criadas duas tabelas intermediárias para gerenciar essas associações. Essas tabelas contêm combinações de *track_id* com *artist_id* e *track_id* com *image_id*, organizadas de forma que cada combinação seja representada em uma linha separada.

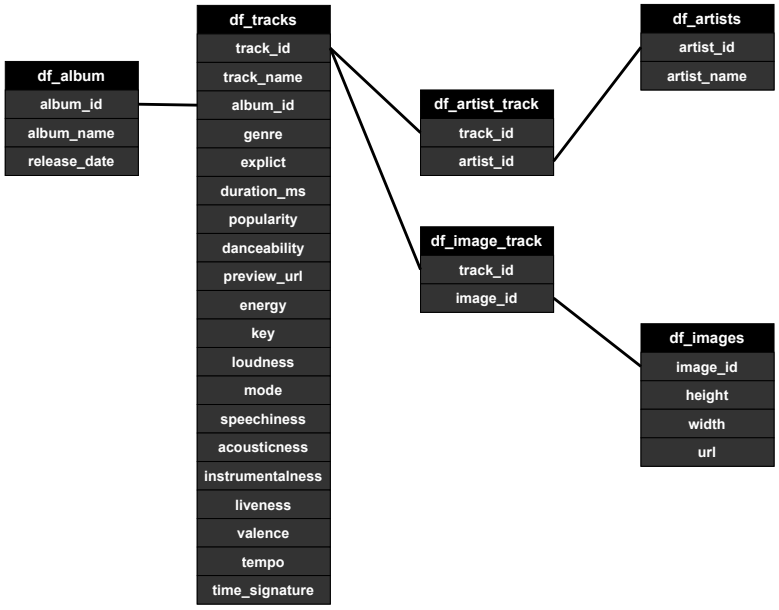


Figura 3.5: Organização dos dados normalizados

Até então, os dados extraídos foram armazenados em diretórios localmente entre cada fase de processamento. Agora, com os dados prontos para análise, inicia-se a etapa de **carga**, que consiste em transferi-los para um banco de dados relacional, como o *PostgreSQL*. Para isso, foram desenvolvidos dois scripts: o primeiro é um código em SQL que cria os esquemas das tabelas no banco de dados, já o segundo é feito em *Python* e converte cada elemento presente nos arquivos em formato JSON obtidos na etapa anterior em comandos SQL equivalentes, para que esses dados possam ser armazenados nas tabelas criadas pelo primeiro script.

Com os dados armazenados no *PostgreSQL*, foi gerado, diretamente pelo banco, o diagrama Entidade-Relacionamento apresentado na Figura 3.6. Esse diagrama ilustra a distribuição das informações entre as tabelas do banco relacional e facilita a visualização da cardinalidade entre os objetos, contribuindo para uma compreensão mais clara da estrutura dos dados.

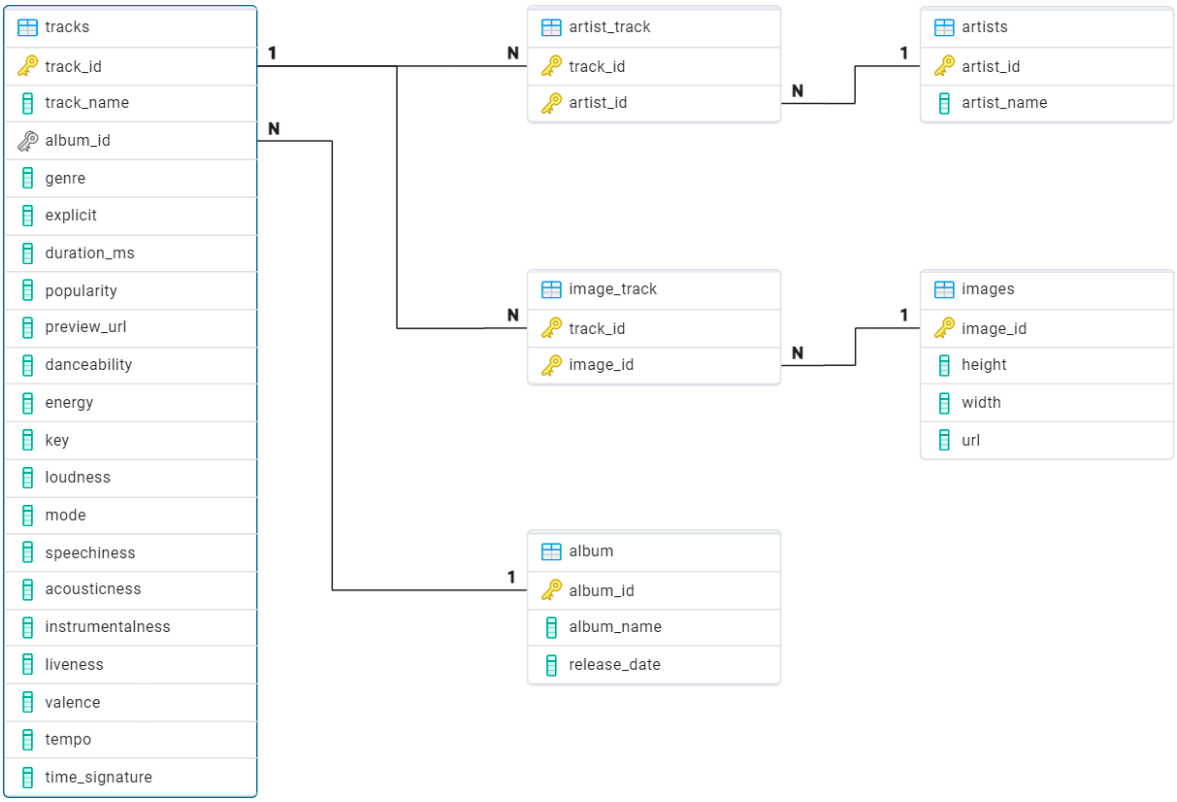


Figura 3.6: Diagrama Entidade-Relacionamento

Por fim, com os dados prontos e centralizados no banco de dados, é feita uma conexão entre o *PostgreSQL* e a ferramenta de análise e visualização escolhida, o *Grafana*. Dessa forma, é possível criar um *dashboard* com todas as visualizações de dados desejadas, além de que a conexão entre o *PostgreSQL* e o *Grafana* garante que dados recém adicionados ao banco de dados sejam enviados para análise, mantendo assim as visualizações no *dashboard* atualizadas sem intervenção externa.

O código dos scripts desenvolvidos para este trabalho e do *dashboard* criado para a visualização dos dados foram publicados em um repositório no *GitHub* [34].

CAPÍTULO 4

Análise de Resultados

Os dados extraídos e transformados permitiram a construção de um *dashboard* interativo, capaz de trazer informações desejadas de acordo com os gostos pessoais do usuário final. A seguir, detalham-se os principais resultados.

A utilização da API do Spotify foi essencial para o processo de coleta de dados deste estudo. Inicialmente, a estratégia de extração baseava-se em obter os dados a partir de *playlists* públicas e de tamanho na escala de centena de faixas, para cada gênero escolhido para análise. Dessa forma, o primeiro conjunto de dados reuniu informações para 1691 músicas. Embora relevante, consideramos esse volume de dados ainda pequeno para representar a diversidade de valor que é possível obter com *Big Data*.

Nesse contexto, o método de identificar os álbuns das faixas do conjunto inicial e coletar dados das outras músicas contidas neles surgiu como uma forma de expandir significativamente a quantidade de dados obtidos. Como resultado do uso das duas abordagens, estiveram disponíveis para análise cerca de 14810 músicas, de 2756 artistas, contidas em 1319 álbuns.

Esse aumento criou um cenário mais próximo ao *Big Data*, não somente pela maior possibilidade de resultados que podem ser retornados na análise final, mas também por permitir melhor simular as dificuldades de trabalhar com uma larga escala de dados com estruturas variadas.

Os dados fornecidos tem um potencial valioso por conterem métricas que não estão disponíveis aos usuários nos aplicativos oficiais disponibilizados pelo *Spotify*. Destacam-se os parâmetros (métricas) que descrevem características musicais em uma escala de 0.0 a 1.0 baseada em seus elementos sonoros, como dançabilidade, energia, volume geral, tempo musical, vocalização, acusticidade, instrumentalidade, vivacidade e valência.

Todavia, uma limitação para este estudo é o retorno de valores nulos para o campo de gênero musical. Como alternativa, foi necessário preencher esse atributo baseado no estilo principal da *playlist* em que a faixa foi retirada para obter uma aproximação dos gêneros associados às músicas. Essa abordagem não é totalmente precisa, uma vez que depende da curadoria dos terceiros que montaram as *playlists* para garantir que contenham apenas músicas do gênero esperado, assim como presume-se que todas as canções de um álbum pertençam ao mesmo estilo.

Quanto às adaptações estruturais, os dados vêm originalmente com conjuntos de campos compactados em dicionários e, devido aos métodos diferentes de extração, a hierarquia para acessar um registro não é padronizada. Por conta disso, é importante normalizar e unificar as informações para preparar um ambiente que propicie consultas otimizadas ao banco de dados. Isto é feito ao reunir todos os conteúdos semelhantes em arquivos organizados por temáticas, como faixas ou artistas, em vez de hierarquicamente. Essa estruturação, além de facilitar eliminar duplicidades, também prepara os arquivos para serem convertidos em tabelas de um banco de dados centralizado, otimizando as análises futuras.

É válido ressaltar que durante as alterações na estrutura, dentro dos arquivos destinados às faixas musicais, mantiveram-se temporariamente os campos *artists_ids* e *images_ids*. Essa escolha foi feita, pois, cada música é relacionada a pelo menos 3 imagens iguais, mas com resoluções diferentes, além de existirem músicas associadas a mais de um artista, sendo assim, compactar os identificadores únicos em dicionários evita ter que lidar com arquivos com o tamanho, no mínimo, três vezes maior. Por fim, a normalização é o passo transformador final dos dados, em que os campos mencionados anteriormente são eliminados e trocados por tabelas intermediárias reduzindo redundâncias, de acordo com a primeira forma normal (1FN).

Portanto, com base nos dados extraídos e estruturados, foi possível desenvolver um *dashboard* com experiências de usuário personalizadas e dinâmicas. Consistindo em um conjunto de painéis customizáveis, o *dashboard* criado permite visualizar rankings das principais músicas e artistas com base em diferentes métricas, além de aplicar filtros personalizados, como selecionar gêneros musicais específicos e definir a quantidade de músicas e artistas exibidos em cada painel.

A Figura 4.1 ilustra como ficam os painéis configurados para exibir as 50 principais músicas para cada métrica, exibindo os títulos das faixas, artistas que participam delas, o álbum de origem e sua imagem de capa, e ao clicar na capa do álbum ou no título da música, é possível ampliar a imagem ou escutar uma prévia sonora da canção, respectivamente. Já a Figura 4.2 mostra painéis com *rankings* dos artistas com a maior média em cada métrica, de acordo com os mesmos filtros utilizados nos painéis da figura anterior.

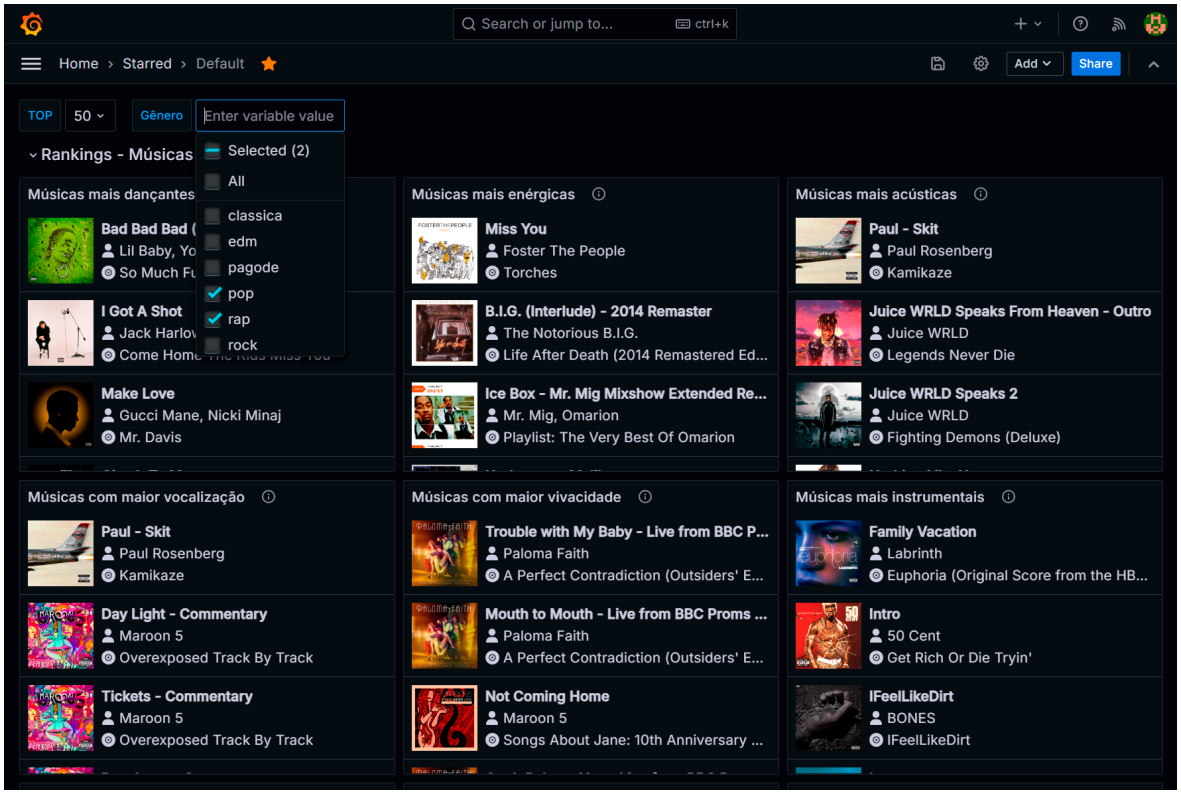


Figura 4.1: Painéis com os *rankings* por música

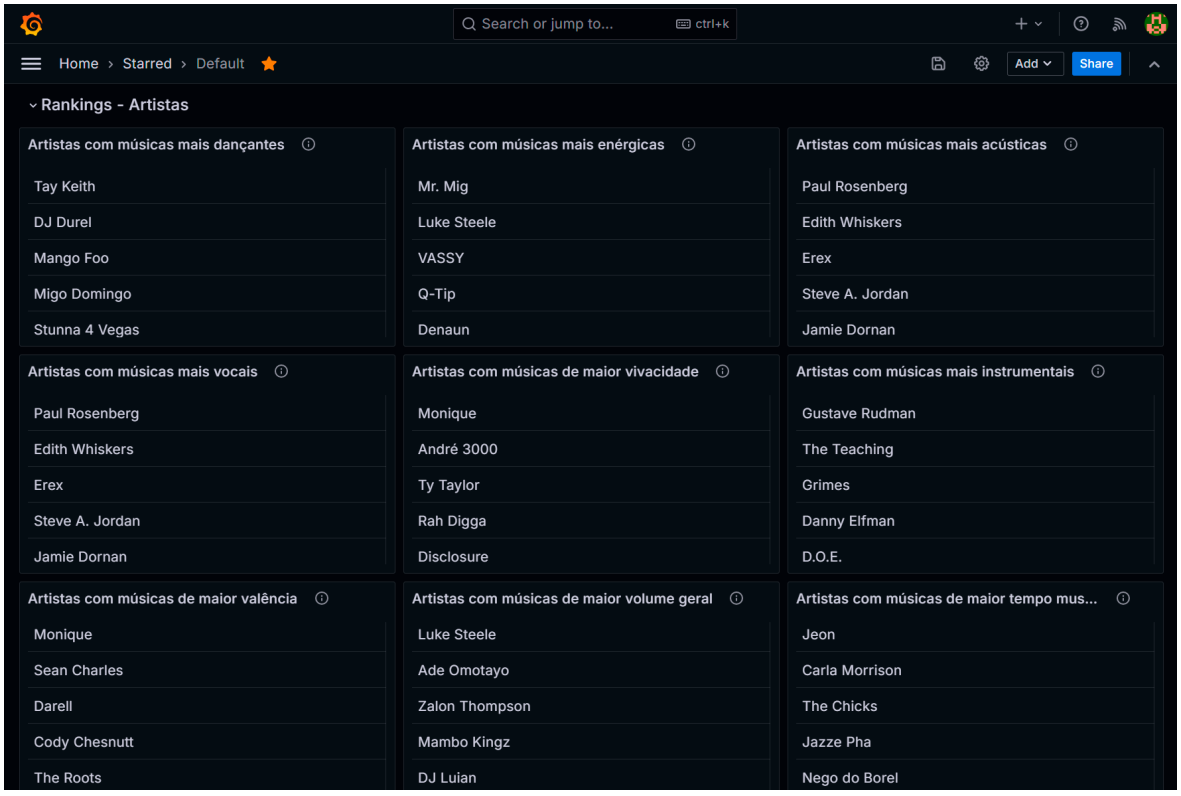


Figura 4.2: Painéis com os *rankings* por artista

Além disso, o *dashboard* exibe gráficos sobre as métricas em cada gênero, como mostra a Figura 4.3. Esses gráficos fornecem uma análise valiosa, ajudando o usuário a identificar os estilos musicais mais alinhados com suas preferências e aumentando as chances de encontrar recomendações que atendam aos seus gostos.

Figura 4.3: Painéis com as médias por gênero

Esses *insights* ajudam a identificar tendências baseadas no humor, como *playlists* animadas para exercícios ou conjuntos acústicos relaxantes para foco, adaptando a música para se adequar às preferências e atividades individuais.

Também é possível aprimorar a descoberta e a interatividade, oferecendo ferramentas para explorar faixas por humor, tom ou tempo musical. Os usuários podem aproveitar experiências gamificadas, como visualizar a energia de uma trilha em tempo real, enquanto DJs e criadores se beneficiam de transições perfeitas usando atributos como tempo musical e volume. Ao aproveitar esses recursos, o *Spotify* aprofunda o engajamento, tornando o *streaming* de música mais intuitivo e agradável.

CAPÍTULO 5

Conclusão

Neste trabalho, foi explorado como o entendimento de um cenário *Big Data* pode orientar a aplicação de técnicas adequadas para lidar com as dificuldades inerentes ao processamento de grandes volumes de dados, como a abordagem Extração, Transformação e Carga (ETL - do inglês *Extract, Transform and Load*).

Optou-se por desenvolver o projeto via código, uma decisão motivada pela intenção de evitar custos com ferramentas pagas e pelo aproveitamento do conhecimento prévio em bibliotecas de manipulação de dados. Essa abordagem permitiu direcionar o tempo de aprendizado para as novas tecnologias empregadas no projeto, como a API do *Spotify* e o *Grafana*, que apresentaram desafios ao longo do desenvolvimento.

A respeito da API do *Spotify*, apesar de fornecer dados detalhados, ela possui uma documentação não muito clara sobre alguns aspectos da ferramenta. Na prática, foi descoberto que existe um sistema de limite de uso dinâmico para o número de requisições realizadas, que não é fixo e varia constantemente, dependendo do horário e da carga nos servidores do *Spotify*. Atingir o limite de requisições repetidamente num curto intervalo de tempo causava bloqueios temporários na credencial de acesso utilizada, que se tornavam cada vez mais longos.

Diante dessas dificuldades, apesar dos dados serem extraídos de duas maneiras, o tipo de extração realizado foi o completo, de modo que o *dataset* não foi atualizado posteriormente, aumentando seu tamanho. Assim, todo o conjunto de dados foi manipulado de uma só vez, permitindo que todos os dados ficassem disponíveis para a carga conjunta e que fosse feito um carregamento inicial, já que o conjunto completo foi enviado ao banco de dados relacional.

Além disso, a documentação da ferramenta também se encontra desatualizada, uma vez que afirma retornar dados para atributos que aparentam ter sido desativados, como foi o caso do campo de gênero, que teve que ser deduzido manualmente durante a fase de transformação. Por conta disso, o volume de dados extraídos foi menor do que o desejável e a veracidade do banco de dados foi afetada devido as aproximações feitas para preencher campos como gênero.

Sobre a visualização de dados, algumas limitações do *Grafana* foram observadas durante a criação dos painéis do *dashboard*. Embora a ferramenta disponibilize vários recursos de manipulação e transformação das informações contidas no banco de dados antes que sejam exibidas para o usuário, na prática os mesmos se mostraram muito engessados, tornando a criação de determinadas visualizações mais complicada, ou, em alguns casos, impossível. Essas limitações puderam ser contornadas através da utilização de *plugins* de terceiros, em particular o *Business Text*, que amplia drasticamente a funcionalidade que pode ser incorporada em um painel do *dashboard*.

Apesar desses empecilhos, foi possível simular com sucesso um cenário de *Big Data*, extraindo valor significativo para análise. Através da visualização dos dados com a criação do *dashboard*, o usuário pode utilizar os filtros nos painéis para identificar as faixas e artistas mais indicados de acordo com as características sonoras das faixas, assim como observar distribuição dos gêneros de acordo com seus elementos sonoros, possibilitando entender tendências musicais.

Para pesquisas futuras, propõe-se aprimorar a usabilidade do *dashboard*, aproximando-o da experiência proporcionada pelo *Spotify*. Isso poderia ser alcançado substituindo o *Grafana* por aplicativos desenvolvidos especificamente para apresentar as informações do *dashboard*, permitindo que o usuário receba suas recomendações de forma remota e diretamente no dispositivo utilizado para reproduzir as músicas no *Spotify*.

Em síntese, esta pesquisa demonstrou a importância do tratamento adequado de dados para extrair valor e transformar informações complexas em soluções práticas. A aplicação do processo ETL mostrou-se eficaz na organização e análise do *Big Data*, ressaltando como estratégias bem planejadas podem gerar valores significativos e impactar positivamente a experiência do usuário final.

Referências Bibliográficas

[1] DUARTE, F. Amount of Data Created Daily (2024). Exploding Topics, 2024. Disponível em: https://explodingtopics.com/blog/data-generated-per-day. Acesso em: 17 de novembro de 2024.

[2] TALAGALA, N. Data as The New Oil Is Not Enough: Four Principles For Avoiding Data Fires. Forbes, 2022. Disponível em: https://www.forbes.com/sites/nishatalagala/2022/03/02/data-as-the-new-oil-is-not-enough-four-principles-for-avoiding-data-fires/. Acesso em: 31 de agosto de 2024.

[3] HEMN BARZAN ABDALLA et al. Big Data: Past, Present, and Future Insights. v. 22, p. 60–70, 26 jul. 2024. DOI: https://doi.org/10.1145/3685767.3685777

[4] EDGE DELTA. Breaking Down The Numbers: How Much Data Does The World Create Daily in 2024? Edge Delta, 2024. Disponível em: https://edgedelta.com/company/blog/how-much-data-is-created-per-day. Acesso em: 17 de novembro de 2024.

[5] KOSTAKIS, P.; KARGAS, A. Big-Data Management: A Driver for Digital Transformation? Information, v. 12, n. 10, p. 411, 1 out. 2021. DOI: https://doi.org/10.3390/info12100411

[6] LANEY, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety; META Group: Stamford, CT, USA, 2001.

[7] MANYIKA, J. et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 1 mai. 2011.

[8] SANDHU, A. K. Big data with cloud computing: Discussions and challenges. Big Data Mining and Analytics, v. 5, n. 1, p. 32–40, mar. 2022. DOI: https://doi.org/10.26599/BDMA.2021.9020016

[9] ROWLEY, J. The Wisdom hierarchy: Representations of the DIKW Hierarchy. Journal of Information Science, v. 33, n. 2, p. 163–180, 15 fev. 2007. DOI: https://doi.org/10.1177/0165551506070706

[10] WIKIMEDIA COMMONS. DIKW Pyramid. Disponível em: https://commons.wikimedia.org/wiki/File:DIKW_Pyramid.svg. Acesso em: 8 de outubro de 2024.

[11] ORACLE. O que é Big Data? Oracle, 2024. Disponível em: https://www.oracle.com/br/big-data/what-is-big-data/#use-cases. Acesso em: 9 de setembro de 2024.

[12] MADDODI, S.; K., K. PRASAD. Netflix Bigdata Analytics- The Emergence of Data Driven Recommendation . International Journal of Case Studies in Business, IT (IJCSBE), v. 3, n. 2581-6942, p. 41–51, 21 out. 2019. DOI: https://doi.org/10.5281/zenodo.3510316

[13] FOULADIRAD, M. et al. Entertaining Data: Business Analytics and Netflix. International Journal of Data Analysis and Information Systems, v. 10, n. 1, p. 13–21, jun. 2018.

[14] AHMED, A.; ABDULKAREEM, A. M. Big Data Analytics in the Entertainment Industry: Audience Behavior Analysis, Content Recommendation, and Revenue Maximization . Reviews of Contemporary Business Analytics, v. 6, n. 1, p. 88, 2023.

[15] NETFLIX TECHNOLOGY BLOG. Artwork Personalization at Netflix. Disponível em: https://netflixtechblog.com/artwork-personalization-c589f074ad76. Acesso em: 21 de setembro de 2024.

[16] AWS. O que é ETL? – Explicação sobre extrair, transformar e carregar. Disponível em: https://aws.amazon.com/pt/what-is/etl/. Acesso em: 9 de outubro de 2024.

[17] BISWAS, N.; KARTIK CHANDRA MONDAL. Integration of ETL in Cloud Using Spark for Streaming Data. Lecture notes in networks and systems, p. 172–182, 25 fev. 2021. DOI: https://dx.doi.org/10.1007/978-981-16-4435-1_18

[18] SINGH, M. M. Extraction Transformation and Loading (ETL) of Data Using ETL Tools. International Journal for Research in Applied Science and Engineering Technology, v. 10, n. 6, p. 4415–4420, 30 jun. 2022. DOI: https://doi.org/10.22214/ijraset.2022.44939

[19] KHAN, B. et al. An Overview of ETL Techniques, Tools, Processes and Evaluations in Data Warehousing. Journal on Big Data, v. 6, p. 1–20, 26 jan. 2024. DOI: https://doi.org/10.32604/jbd.2023.046223

[20] GOKILA, D.; BALASUBRAMANI, S. Impact of Normalization in Future. International Journal of Trend in Scientific Research and Development (IJTSRD), v. 3, n. 5, p. 153–156, ago. 2019. Disponível em: https://www.ijtsrd.com/papers/ijtsrd25128.pdf

[21] ALURA. Normalização em Banco de Dados - Estrutura. Disponível em: https://www.alura.com.br/artigos/normalizacao-banco-de-dados-estrutura?srsltid=AfmBOoqaSQoSb0KwVDGiEk1KTnUb-lqz52MoC7uiY2nndoDq7A14jsev. Acesso em: 22 de junho de 2024.

[22] BALACHANDAR PAULRAJ. Scalable ETL Pipelines for Telecom Billing Systems: A Comparative Study. Darpan International Research Analysis, v. 12, n. 3, p. 555–573, 20 set. 2024. DOI: https://doi.org/10.36676/dira.v12.i3.107

[23] QAISER, A. et al. Comparative Analysis of ETL Tools in Big Data Analytics. Pakistan Journal of Engineering and Technology, v. 6, n. 1, p. 7–12, 20 jan. 2023. Disponível em: https://journals.uol.edu.pk/pakjet/article/view/2266/1136

[24] KELLEHER, C.; WAGENER, T. Ten guidelines for effective data visualization in scientific publications. Environmental Modelling & Software, 26(6), 822–827, 19 jan. 2011. DOI: https://dx.doi.org/10.1016/j.envsoft.2010.12.006

[25] GOUNDAR, S.; BHARDWAJ, A.; SINGH, S.; SINGH, M.; H L, G. Big Data and Big Data Analytics: A Review of Tools and its Application. p. 8-10, 1 jan. 2021. DOI: https://doi.org/10.4018/978-1-7998-6673-2.ch001

[26] GRAFANA LABS. Grafana: Query, visualize, alerting observability platform. Disponível em: https://grafana.com/grafana/. Acesso em: 22 de agosto de 2024.

[27] STROHBACH, M. et al. Towards a Big Data Analytics Framework for IoT and Smart City Applications. Modeling and Processing for Next-Generation Big-Data Technologies, p. 257–282, 2015. DOI: https://dx.doi.org/10.1007/978-3-319-09177-8_11

[28] KUJAWSKI, M. Building an Efficient ETL/ELT Process for Data Delivery. Disponível em: https://medium.com/@mariusz_kujawski/building-an-efficient-etl-elt-process-for-data-delivery-9ee775375418. Acesso em: 6 de maio de 2024.

[29] ANUDEEP MEDICHARLA. Exploring the World of Music With the Spotify API and Postbot — Postman Blog. Disponível em: https://blog.postman.com/spotify-api-and-postbot-ai/. Acesso em: 02 de maio de 2024.

[30] COSTA, J. F. Using the Spotify API for data-driven analysis of my playlists (Part 1/2). Disponível em: https://ze1598.medium.com/using-the-spotify-api-for-data-driven-analysis-of-my-playlists-part-1-2-a4598ca7b96d. Acesso em: 15 de abril de 2024.

[31] COSTA, J. F. Using the Spotify API for data-driven analysis of my playlists (Part 2/2). Disponível em: https://ze1598.medium.com/using-the-spotify-api-for-data-driven-analysis-of-my-playlists-part-2-2-92e60331d038. Acesso em: 16 de abril de 2024.

[32] NASCIMENTO, J. C. D. Using Spotify data to predict which "Novidades da semana" songs would become hits. Disponível em: https://medium.com/@jcarolinedias1/using-spotify-data-to-predict-which-novidades-da-semana-songs-would-become-hits-e817ae0c091. Acesso em: 20 de abril de 2024.

[33] SPOTIFY. Web API Reference — Spotify for Developers. Disponível em: https://developer.spotify.com/documentation/web-api/reference/get-audio-features. Acesso em: 12 de junho de 2024.

[34] GITHUB. Processo de ETL e visualização de dados utilizando dados da Spotify Web API. Disponível em: https://github.com/mbc07/TCC-BigData-Spotify. Acesso em: 27 de novembro de 2024.