Atividade Orientada de Ensino: Análise de Dados Musicais Utilizando Técnicas de Visualização

Isadora S. da Silva, Mateus B. Cassiano

Resumo: Este trabalho explora a análise de dados musicais utilizando o processo de Extração, Transformação e Carga (ETL) para organizar e explorar *Big Data*. O estudo expande uma pesquisa anterior focada em análises musicais, agora com ênfase nas métricas de popularidade e duração das músicas por gênero. A análise foi realizada com o uso das bibliotecas Pandas, *Seaborn* e *Mathplotlib*, proporcionando uma visualização clara e eficaz dos padrões encontrados. O estudo evidencia a importância de técnicas analíticas para lidar com *Big Data* e gerar *insights* valiosos, demonstrando como a organização e visualização adequadas dos dados podem fornecer informações estratégicas e aprofundadas sobre as tendências musicais.

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

Abstract: This work explores musical data analysis using the Extract, Transform, and Load (ETL) process to organize and analyze Big Data. The study builds upon previous research focused on musical analyses, now emphasizing metrics such as song popularity and duration by genre. The analysis was conducted using the Pandas, Seaborn, and Matplotlib libraries, providing clear and effective visualizations of the identified patterns. The study highlights the importance of analytical techniques for managing Big Data and generating valuable insights, demonstrating how proper data organization and visualization can deliver strategic and in-depth information on musical trends.

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

1 Introdução

O conceito de *Big Data* surge para definir um conjunto de dados caracterizado pelo alto volume, velocidade e variabilidade, além de aspectos como veracidade e valor das informações geradas. Esses dados são produzidos continuamente por diversas fontes, como redes sociais, sensores, dispositivos IoT, transações financeiras, entre outros, formando um ambiente de crescimento exponencial [1]. Todavia, a grande quantidade de dados, por si só, não gera valor ou *insights* úteis se não for processada, organizada e analisada de maneira eficiente.

Nesse contexto, o processo de Extração, Transformação e Carga (ETL) desempenha um papel crucial. Ele permite a coleta de dados brutos de diferentes fontes, a aplicação de transformações para limpeza, organização e enriquecimento, e, por fim, a carga em sistemas de armazenamento ou análise [2]. Dessa forma, o ETL prepara os dados para que sejam utilizados em análises mais avançadas, viabilizando a geração de informações estratégicas e tomadas de decisão baseadas em dados. Esse processo é essencial para lidar com os desafios e a complexidade do *Big Data*.

Este trabalho é uma extensão do estudo intitulado "*Big Data* e Processamento de Dados: Uma Jornada para a Descoberta de *Insights*" [3], onde foram explorados dados musicais com foco na análise de parâmetros de elementos sonoros por gênero e artista. A pesquisa inicial utilizou o *Grafana* para criar uma visualização interativa e dinâmica. Agora, a proposta deste estudo é expandir a análise, utilizando técnicas alternativas de visualização dos dados, com o objetivo de demonstrar novas possibilidades de análise e enriquecer a interpretação dos dados existentes.

2 Fundamentação Teórica

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 [4].

As primeiras definições para considerar uma coleção de dados como *Big Data* se baseiam no conceito de "3 V's" proposto por Laney [5]. Com o passar do tempo o conceito de *Big Data* passou a abranger "5 V's", os dois V's subsequentes enfatizam sua utilidade prática e na aplicação [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.

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.

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.

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 [2]. Dessa forma, o ETL é fundamental em ambientes que lidam com o *Big Data*.

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) [7].

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.

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

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 [8]. A escolha de uma ferramenta de visualização envolve diversos fatores, como facilidade de uso, integração com fontes de dados, customização, escalabilidade, custo e suporte [9].

3 Implementação

A implementação deste estudo baseia-se em utilizar o conjunto de dados resultantes do processo ETL montado anteriormente em [3] para demonstrar novas formas de análise e gerar *insights* adicionais a partir dos mesmos dados.

O projeto foi desenvolvido localmente 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 para Manipulação dos Dados:** Pandas;
- **Bibliotecas para Visualização dos Dados:** *Seaborn* e *Mathplotlib*.

Primeiramente, foi realizada a integração e carregamento dos dados. Para isso, os dados já extraídos e organizados pelo processo de ETL foram carregados em um ambiente de análise, permitindo que todas as informações sobre as músicas, como gênero, popularidade e outras características, fossem centralizadas em um formato adequado para análise.

Na segunda etapa, o foco foi em organizar os dados de acordo com categorias relevantes, no caso, os gêneros musicais. Para isso, os dados foram agrupados para identificar padrões e tendências dentro de cada gênero, como a popularidade ou duração das músicas. Dessa forma, é possível entender as músicas que se destacam em cada gênero, identificando as mais populares e as mais longas.

Por fim, a última etapa consistiu na visualização dos dados, onde as informações extraídas e agrupadas foram representadas de maneira gráfica.

4 Resultados

A primeira análise realizada mostra como a música mais popular pertence ao gênero *pop*, enquanto a com menor popularidade é do gênero clássico, conforme mostra a Figura 1. Além disso, após agrupar as músicas por gênero, foi possível identificar qual foi a música mais popular de cada um.

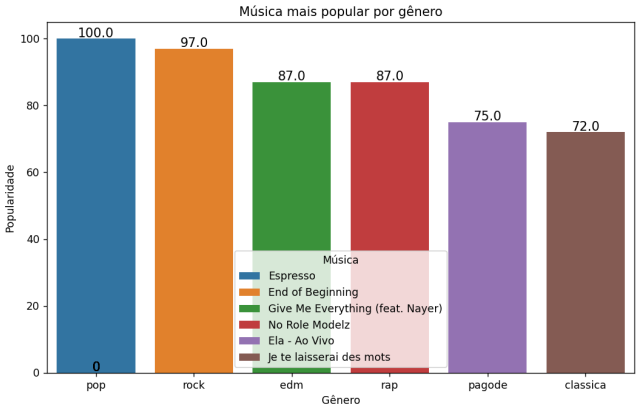

Figura 1: Gráfico de popularidade por gênero

De forma semelhante, a segunda análise apresenta a duração das músicas mais populares dentro de cada gênero. Como ilustrado na Figura 2, observa-se que a música mais longa pertence ao gênero *rap*. Por outro lado, na Figura 1, apesar de a música de *rap* não ser a mais popular, ela ainda é amplamente reconhecida. Isso demonstra que, embora a música de *rap* tenha uma duração consideravelmente maior em comparação com as de outros gêneros, isso não impactou negativamente sua popularidade.

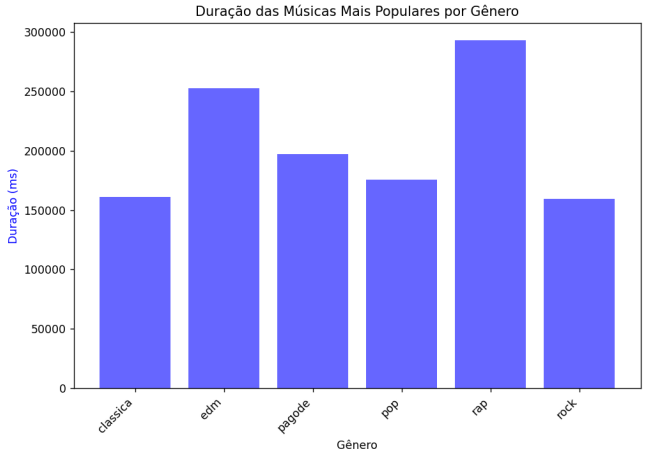

Figura 2: Gráfico de duração em milissegundos por gênero

Sendo assim, a visualização desses dados em um gráfico de barras permite observar de forma clara as músicas que se destacam em seus respectivos gêneros.

5 Conclusão

A conclusão deste estudo revela a importância de aplicar técnicas analíticas modernas para explorar grandes volumes de dados e extrair *insights* valiosos. Através do uso de um processo ETL bem estruturado, é possível organizar e processar dados brutos, transformando-os em informações que podem ser visualizadas e analisadas de maneira eficaz. A análise realizada sobre as músicas mais populares por gênero, associada a durabilidade destas faixas, trouxe uma compreensão mais profunda sobre os padrões de popularidade e duração musical nas diferentes categorias musicais.

Além disso, a visualização dos resultados em gráficos ajudou a ilustrar de forma clara as tendências musicais, permitindo que os dados fossem mais facilmente compreendidos e analisados. O uso de ferramentas como Pandas, *Seaborn* e *Mathplotlib* possibilitou uma abordagem prática e eficiente para manipulação e visualização dos dados, e os resultados obtidos não só validaram a importância da organização dos dados, mas também destacaram como a análise de grandes volumes de informações pode fornecer *insights* valiosos.

Por fim, o estudo demonstrou que, embora a complexidade do *Big Data* exija ferramentas e processos adequados, quando bem estruturados, esses dados podem ser convertidos em valor real, capaz de fornecer respostas e direcionamentos precisos. As metodologias aplicadas aqui são apenas um exemplo de como as técnicas de análise podem ser expandidas e utilizadas para gerar *insights* mais aprofundados em cenários com grandes volumes de dados.

Referências

[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] 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.

[3] DA SILVA, ISADORA S.; CASSIANO, MATEUS B.; Big Data e Processamento de Dados: Uma Jornada para a Descoberta de Insights. Trabalho de Conclusão de Curso. Universidade Federal de Mato Grosso do Sul, Faculdade de Computação, 2024.

[4] 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`

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

[6] 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`

[7] 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`

[8] 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

[9] 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