**BI Kmovies: analisando dados sobre filmes de uma plataforma de *streaming***

ANTONIO ALVES DE SOUSA JUNIOR (IFPB, Campus Sousa), BIANCA KARLA AMORIM DE SOUSA MEDEIROS (IFPB, Campus João Pessoa), ISLEIMAR DE SOUZA OLIVEIRA (IFPB, Campus Cajazeiras), JANDERSON FERREIRA DUTRA (IFPB, Campus Cajazeiras), DAMIRES YLUSKA SOUZA FERNANDES (IFPB, Campus João Pessoa)

**E-mails:** [antonio.sousa@ifpb.edu.br, bianca.amorim@academico.ifpb.edu.br, isleimar.oliveira@ifpb.edu.br, janderson.dutra@ifpb.edu.br, damires@ifpb.edu.br](mailto:fulano@academico.ifpb.edu.br)

**Área de conhecimento:** 1.03.03.03-0 Banco de Dados.

**Palavras-Chave**: ETL; Visualização de dados; Netflix; IMDb; TMDB; MongoDB.

1. **Introdução**

A Netflix é um dos principais serviços de streaming disponíveis no mercado, possuindo usuários conectados ao redor do mundo, através de uma assinatura mensal e com acesso à internet. As sugestões personalizadas ajudam o usuário a encontrar séries e filmes de seu interesse, através de um sistema próprio e complexo de recomendações (OLIVEIRA, 2019). Apesar da relevância do sistema de recomendação do serviço, muitos assinantes da plataforma preferem consultar as opiniões da crítica e de outros usuários antes de assistirem a algum conteúdo (LADEIRA, 2019; SIGILIANO et al., 2016; JANNACH et al., 2010).

Nesse contexto, duas plataformas de consulta de séries e filmes se destacam: a The Movie Database (TMDb) e o Internet Movie Database (IMDb). Ambas são alimentadas por indivíduos interessados neste tipo de conteúdo, porém ainda há carência de aplicações que integrem as informações dessas duas plataformas. Olney (2010) realizou um estudo sobre classificação do conjunto de dados da Netflix utilizando um modelo probabilístico para mostrar que uma análise baseada na preferência do usuário para prever o gênero tem mais sucesso do que uma abordagem baseada nas características das produções. O dataset foi extraído da API (Application Programming Interface) da Netflix, disponível gratuitamente online na época do estudo, contendo dados de 17.770 filmes no conjunto, 480.189 usuários e 100.480.507 avaliações. Para cada filme há uma identificação exclusiva, o título e o ano de lançamento, além de uma pasta para cada id de filme, com o conjunto com id do usuário, classificação e data da classificação atribuídas por vários usuários.

Considerando esse cenário, esse trabalho apresenta resultados de análises de dados sobre filmes da plataforma de streaming mundial Netflix. Para isso, dados sobre avaliação, orçamento, bilheteria, diretor e elenco foram coletados por meio das APIs do TMDb e do IMDb. O trabalho provê análises sobre o desempenho do conteúdo e compara as abordagens de avaliação da mídia na plataforma de origem e na comunidade.

Uma aplicação web para visualização de dados intitulada BI Kmovies foi implementada para o público de consumidores de conteúdos da plataforma de streaming e qualquer pessoa interessada em informações adicionais sobre os conteúdos de mídia (BOAGLIO, 2020). Os dados obtidos nas três fontes foram integrados na plataforma do MongoDB. A base de dados gerada pode ser usada para extrair outras informações e prover novas análises dos dados. Todos os arquivos e orientações para uso dos recursos gerados estão disponíveis no repositório[[1]](#footnote-1) do projeto.

1. **Materiais e Métodos**

Este trabalho foi desenvolvido seguindo um processo adaptado de engenharia de dados (coleta, seleção, integração, preparação) e carga (persistência) em um BD NoSQL (MongoDB), conforme mostra a Figura 1.

Esse modelo de processo possui basicamente três etapas que foram seguidas: (i) coleta dos dados em três fontes; (ii) processo de engenharia dos dados (ETL - Extract; Transform; Load) e (iii) desenvolvimento de uma aplicação de visualização de dados.

Neste trabalho foram usadas três fontes de dados. A primeira foi extraída da plataforma Kaggle, obtendo-se o arquivo netflix\_titles.csv (A) que contém dados do conteúdo da Netflix; este dataset é atualizado mensalmente, e possui licença aberta para uso. A segunda fonte foi a TMDb (B), cujos dados foram obtidos através do acesso à API desta plataforma, usando a linguagem de programação Python. Para que fosse possível a integração das fontes A e B, foram usados os datasets (formato TSV) disponibilizados no IMDb (C), pois através do identificador universal deste dataset é possível consultar títulos no TMDb.

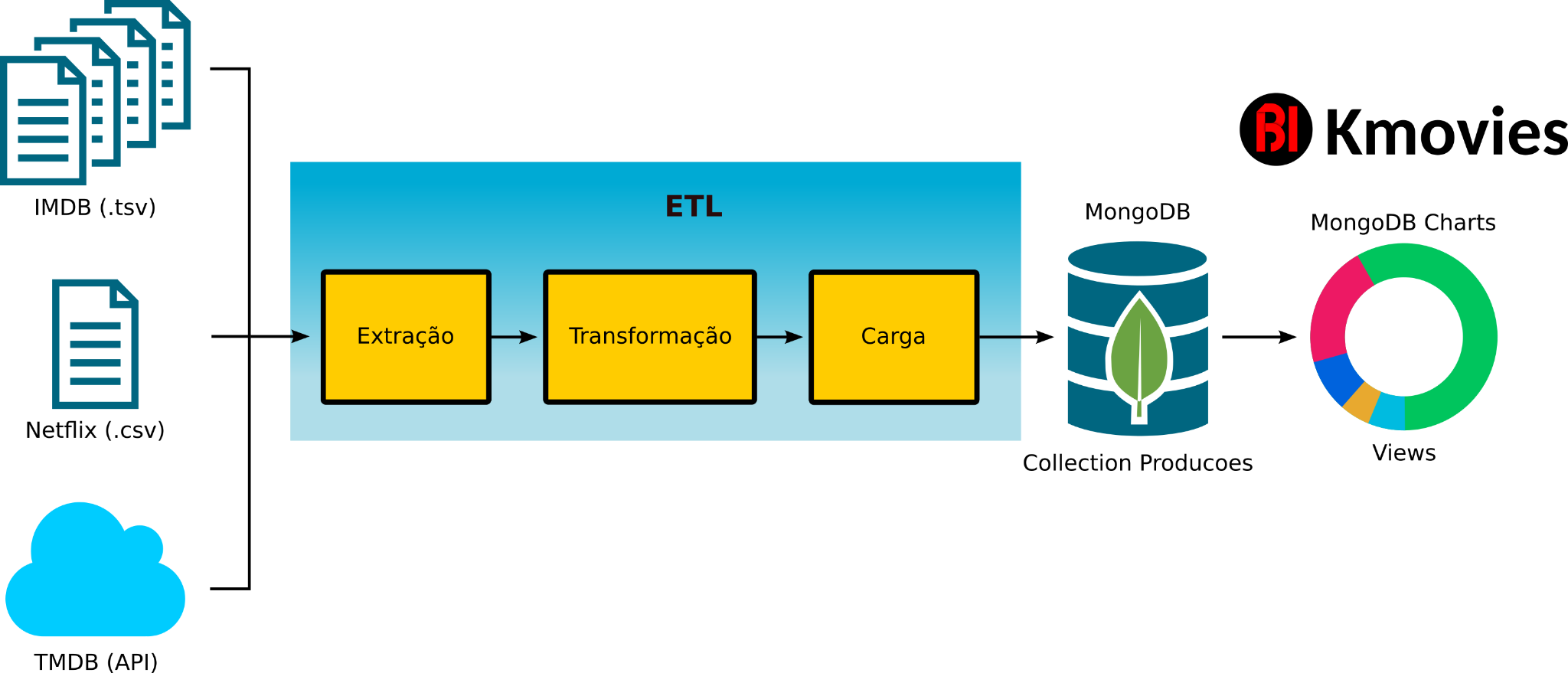


Figura 1 - Pipeline de ETL e desenvolvimento da aplicação | Fonte: os autores. BI KMovies (2021)

Com o estudo das três fontes foi possível integrar os dados numa única coleção. Para isso, foi necessário converter, via implementação em Python, sete datasets em TSV da fonte C e um dataset em CSV da fonte A para o formato JSON. Após esta etapa, os dados foram carregados no MongoDB. Para integrar a fonte A e C foi realizado um pipeline de agregação e foram usados os campos nome e ano de produção como critério de relacionamento entre as bases de dados. Com isto, a coleção originada com os dados da plataforma Kaggle foi modificada passando a ter como identificador único o valor do IMDb. A Figura 2 apresenta o processo para inserção do campo tconst da base do IMDB no campo netflix\_imdb da Netflix para possibilitar as requisições à API do TMDB.

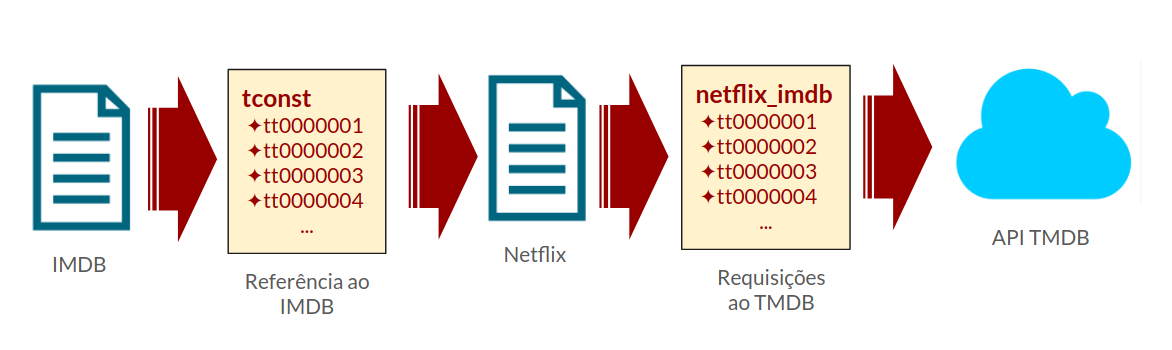
****

Figura 2 - Processo de inclusão de referências | Fonte: os atores. BI KMovies (2021)

A requisição direta a partir dos dados do IMDB é inviável, pois possui mais de um milhão de títulos registrados, tornando assim o processo de requisição de todos os itens, demasiadamente oneroso.

Para extração dos dados da fonte B foram realizadas requisições através do identificador imdb\_id e o token gerado pelo TMDb de forma automatizada, de acordo com a coleção netflix\_imdb. Por fim, ao término da etapa de integração dos dados, as bases Netflix, IMDB e TMDB geraram a base de dados ‘producoes’. Para atualização dos dados, o fluxo deve ser repetido sempre que houver alteração dos datasets originais.

1. **Resultados e Discussão**

Após o processo de extração, transformação e integração dos dados, foi feita sua carga por meio de uma coleção de documentos de filmes da Netflix no banco de dados MongoDB. Ao final, a coleção 'producoes' advinda do processo apresentado é composta por um total de 680 títulos e 1636 produtoras.

Foram criadas visualizações dos dados, utilizando o software MongoDB Charts, vinculado à coleção ‘producoes’, para melhor entendimento e análise desses dados. O objetivo era responder algumas perguntas formuladas sobre os dados, tais como: (i) Qual a média de avaliação dos filmes por gênero? (ii) Quais produções tiveram o maior orçamento? (iii) Quais produções tiveram a maior receita? (iv) Os filmes antigos, considerados clássicos pelos cinéfilos, são melhores, em termos de avaliação pública? (v) Os filmes brasileiros são considerados ruins pelos usuários?; entre outras.

Além destas perguntas, os dados permitem a obtenção de muitas outras análises a respeito das produções na Netflix, permitindo avaliar também o contexto de filmes independentes e documentários.

A Figura 3 ilustra uma visualização geral sobre as 680 produções e as 1.636 produtoras. Essa visualização permite responder a diversas das questões levantadas, como, por exemplo, a de cada filme ter uma ou mais produtoras, podendo também a produção do filme ter sido realizada em vários países e/ou localidades.

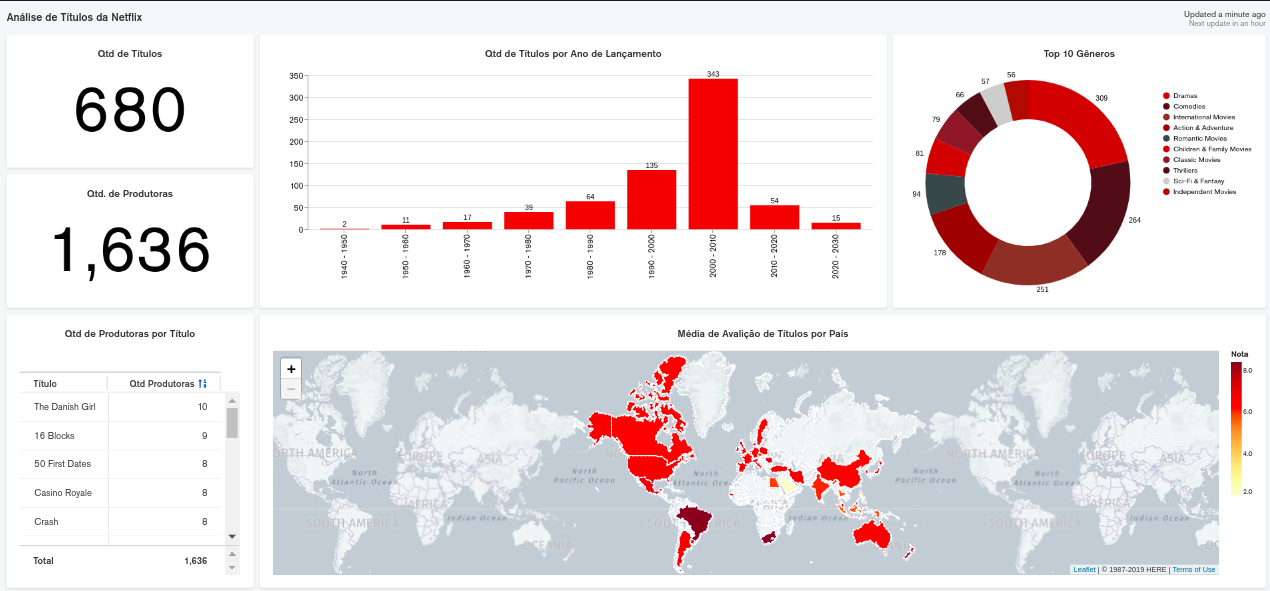
**

Figura 3 - Visualização de análise de produções | Fonte: os autores. BI KMovies (2021)

1. **Considerações Finais**

Neste trabalho foi apresentado o processo de desenvolvimento da BI KMovies - uma aplicação web cujo objetivo é consumir dados integrados que originaram a base de dados ‘producoes’. Visa exibir dados de streaming onde os usuários podem compreender visualmente informações referentes a, por exemplo, avaliação, orçamento, popularidade e receita de títulos disponíveis na Netflix. Outra contribuição importante é que essa base de dados pode ser usada para criação de novos repositórios de dados mais volumosos, possibilitando a obtenção de novas análises.

O maior esforço no projeto foi dedicado à etapa de transformação dos dados, em que as maiores dificuldades surgiram, devido à heterogeneidade das fontes, que apresentavam tipos de dados diferentes, sendo necessário a identificação e criação de chaves para relacioná-las de forma correta. Os obstáculos foram resolvidos no pipeline de integração da fonte da Netflix com o IMDB, o que tornou possível a extração dos dados da API do TMDB. A aplicação web mostra dados oriundos de diversas fontes (datasets e API) de modo a facilitar a análise das informações que o usuário busca sobre os filmes, além de gerar um rankeamento e fazer um comparativo graficamente entre os filmes disponíveis.

Como trabalhos futuros, pretende-se adicionar novas funcionalidades, tais como: gráficos para comparar as avaliações dos filmes nas plataformas IMDb e TMDB simultaneamente; visualizações para exibir o ator ou diretor mais popular ou melhor avaliado; e permitir ao usuário criar suas próprias visualizações a partir do conjunto de dados ‘producoes’.

**Referências**

BOAGLIO, Fernando. MongoDB: construa novas aplicações com novas tecnologias. Editora Casa do Código, 2020.

FORBES, Um ano depois do início da pandemia, plataformas de streaming contabilizam ganhos. Disponível em: <https://forbes.com.br/forbes-money/2021/03/um-ano-depois-do-inicio-da-pandemia-plataformas-de-streaming-contabilizam-ganhos/>. Acesso em: 09 mai. 2021.

IMDb. IMDb Datasets. 2021. Disponível em: https://www.imdb.com/interfaces/. Acesso em: 30 jun. 2021.

JANNACH, D.; ZANKER, M.; FELFERNIG, A.; FRIEDRICH, G. Recommender systems: an introduction. Cambridge University Press, 2010.

Kaggle. Netflix Movies and TV Shows. 2021. Disponível em: https://www.kaggle.com/shivamb/netflix-shows?select=netflix\_titles.csv. Acesso em: 01 abr. 2021.

LADEIRA, João Damasceno Martins. O algoritmo e o fluxo: Netflix, aprendizado de máquina e algoritmos de recomendações. Intexto, n. 47, p. 166-184, 2019.

MongoDB Charts. 2021. Disponível em: https://docs.mongodb.com/charts/saas/. Acesso em: 05 abr. 2021.

OLIVEIRA, Alan de Freitas. Comportamento de consumidores de serviços de streaming: um estudo de caso de usuários da Netflix no Brasil e em Portugal. 2019. Tese de Doutorado.

OLNEY, Andrew. Likability-Based Genres: Analysis and Evaluation of the Netflix Dataset. 2010.

SIGILIANO, Daiana; FAUSTINO, Eduardo. NETFLIX: Sistemas de Recomendação Inteligentes. http://dx.doi.org/10.15601/1983-7631/rt. v9n16p13-26. Revista Tecer, v. 9, n. 16, 2016.

THE Movie Database (TMDb). 2021. Disponível em: www.themoviedb.org/. Acesso em: 04 abr. 2021.

1. Disponível em: https://github.com/biancaamorimelo/bi-kmovies [↑](#footnote-ref-1)