UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ESCOLA DE INFORMÁTICA APLICADA

Web Scraping, ETL e análise dos catálogos dos principais serviços de *streaming* de filmes no Brasil

Davi Gervásio Coutinho

Orientador
Pedro Nuno de Souza Moura

RIO DE JANEIRO, RJ — BRASIL
SETEMBRO DE 2024

Catalogação informatizada pelo autor

G871 Gervásio Coutinho, Davi
Web Scraping, ETL e análise dos catálogos dos principais serviços de streaming de filmes no Brasil / Davi Gervásio Coutinho. -- Rio de Janeiro, 2024.
72

Orientador: Pedro Nuno de Souza Moura.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Sistemas de Informação, 2024.

1. Web Scraping. 2. ETL. 3. Análise de dados. I. Nuno de Souza Moura, Pedro, orient. II. Titulo.

Web Scraping, ETL e análise dos catálogos dos principais serviços de *streaming* de filmes no Brasil

Davi Gervásio Coutinho

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado por:

_____

Prof. Dr. Pedro Nuno de Souza Moura (UNIRIO)

_____

Prof Dr. Jobson Luiz Massollar da Silva(UNIRIO)

_____

Prof. Dr Reinaldo Viana Alvares(UNIRIO)

RIO DE JANEIRO, RJ — BRASIL.

SETEMBRO DE 2024

Amigo eu nunca fiz bebendo leite
Amigo eu não criei bebendo chá
Eu sou da madrugada, me respeite
Que eu sei a hora de ir trabalhar

Zeca Pagodinho
Toda Hora

E onde a sorte há de te levar
Saiba, o caminho é o fim, mais que chegar
E queira o dia ser gentil
À tua mão aberta pra quem é

Little Joy
The Next Time Around

Agradecimentos

Aos meus pais, Ana Lúcia e Geraldo, por nunca ter me faltado nada, nada mesmo.

À minha irmã, Helena, por ser a pessoa que mais me conhece neste mundo.

Aos meus avós, Tatá e Petinha, por todo o carinho e amor.

Ao Prof. Dr. Pedro Nuno de Souza Moura, pela paciência e orientação neste trabalho.

A **todos** os meus amigos que sempre estiveram comigo e a **todos** que fiz nessa jornada — quem sabe que é, é.

A todos que cruzaram minha jornada profissional e me ensinaram tudo o que sei, em especial: Diogo Moreno, Fred Argolo, André Gorestin, Marcus Soliva, Alexander Jardim e Guilherme Martins.

Às duas melhores pessoas que conheço, Eduardo Brasil e Renato da Costa. Vocês são minha maior inspiração.

RESUMO

Este trabalho realiza a extração, análise e democratização dos dados dos catálogos de filmes dos principais serviços de *streaming* no Brasil por meio da aplicação de técnicas de *Web Scraping*, ETL (*Extract*, *Transform*, *Load*) e análise de dados. Para isso, foram selecionados os sete serviços de *streamings* mais populares no país e coletados os dados de seus respectivos filmes em diferentes portais existentes na internet. Após a coleta, os dados foram transformados e armazenados, realizando-se diversas análises, incluindo a identificação de tendências de gênero, comparações entre os catálogos e análise de notas e avaliações de diversas fontes de serviços de crítica. Em geral, este trabalho demonstrou a utilidade das técnicas de *Web Scraping*, ETL e análise de dados na aplicação a catálogos de serviços de streaming no Brasil. Dessa maneira, foi possível fornecer insights valiosos a respeito do mercado de serviços de streaming no país, além da própria democratização da base de dados para outros usuários.

Palavras-chave: Serviços de *streaming*, *Web Scraping*, Extração, Transformação, Carga, Dados, Análise, Filmes.

ABSTRACT

This work involves the extraction, analysis, and democratization of movie catalog data from the leading streaming services in Brazil through the application of Web Scraping, ETL (Extract, Transform, Load), and data analysis techniques. For this purpose, the seven most popular streaming services in the country were selected, and data on their respective movies were collected from various online portals. After collection, the data were transformed and stored, with several analyses conducted, including the identification of genre trends, catalog comparisons, and analysis of ratings and reviews from various sources of criticism services. Overall, this work demonstrated the usefulness of Web Scraping, ETL, and data analysis techniques when applied to streaming service catalogs in Brazil. This approach provided valuable insights into the streaming service market in the country, as well as the democratization of the data set for other users.

Keywords: Streaming services, Web Scraping, Extraction, Transformation, Load, Data, Analysis, Movies.

Sumário

Lista de figuras

1. Introdução

1.1. Motivação

A revolução digital nas últimas décadas reconfigurou de maneira substancial o panorama da indústria do entretenimento, especialmente no que diz respeito à forma como as pessoas consomem filmes e programas de televisão. Esse fenômeno é amplamente documentado na literatura especializada, que destaca como a convergência tecnológica e a internet alteraram os paradigmas de produção, distribuição e acesso ao conteúdo audiovisual. Manuel Castells, em *The Power of Communication* (2009), buscou abordar esta questão, afirmando que:

> Na sociedade em rede, a cultura está majoritariamente embutida nos processos de comunicação, particularmente no hipertexto eletrônico, com redes globais de negócios multimídia e a internet no seu núcleo. Assim, ideias podem ser geradas a partir de uma variedade de origens, e ligadas a interesses específicos e subculturas. (Castells, 2009, p. 46, tradução própria)[1]

Castells prossegue, delineando o impacto dessas mudanças na criação e na difusão de discursos:

> Na sociedade em rede, discursos são gerados, difundidos, disputados, internalizados e, por fim, incorporados à ação humana, no domínio da comunicação socializada construído em torno de redes locais-globais de comunicação multimodal digital, incluindo a mídia e a internet. O poder na sociedade em rede é o poder da comunicação. (Castells, 2009, p. 53, tradução própria)[2]

Essas observações de Castells são cruciais para entender como as tecnologias digitais não apenas transformaram a indústria do entretenimento, mas também reconfiguraram as dinâmicas culturais e sociais, influenciando diretamente a forma como o conteúdo audiovisual é produzido, distribuído e consumido.

Em particular, os serviços de *streaming* emergiram como agentes transformadores, democratizando o acesso a uma grande diversidade de obras

[1] "In the network society, culture is mostly embedded in the processes of communication, particularly in the electronic hypertext, with global multimedia business networks and the Internet at its core. So, ideas may be generated from a variety of origins, and linked to specific interests and subcultures."

[2] "In the network society, discourses are generated, diffused, fought over, internalized, and ultimately embodied in human action, in the socialized communication realm constructed around local–global networks of multimodal, digital communication, including the media and the Internet. Power in the network society is communication power."

cinematográficas e seriadas, oferecidas de forma conveniente e alinhadas às preferências individuais dos consumidores.

No Brasil, este cenário não é diferente. Dados recentes indicam que a adesão aos serviços de *streaming* alcançou marcas expressivas, com mais de 40% dos domicílios brasileiros utilizando essas plataformas em 2022 (Nery, 2023). Tal popularidade reflete uma mudança nos hábitos de consumo de entretenimento e aponta para uma nova configuração cultural e social mediada pela tecnologia. Além disso, a pandemia de COVID-19 acelerou significativamente a adoção desses serviços, pois o isolamento social levou as pessoas a buscarem novas formas de entretenimento e conteúdo digital (Matos, 2020).

O conceito de *streaming*, como discutido por Spilker e Colbjørnsen (2020), transcendeu ser apenas uma tecnologia de distribuição, tornando-se uma prática cultural e econômica profundamente arraigada na sociedade contemporânea.

> Em sua essência, o *streaming* parece significar uma forma ilimitada de distribuir e consumir conteúdo de mídia. Representantes corporativos e ativistas da Internet igualmente sublinham o que argumentam ser a natureza transformadora ou disruptiva do *streaming*. (Spilker; Colbjørnsen, 2020, p. 1211)[3]

No entanto, apesar da crescente penetração desses serviços, pouco se sabe sobre a composição e a diversidade dos catálogos disponíveis, especialmente no que tange à representatividade de gêneros, nacionalidades e faixas etárias, o que pode suscitar questionamentos acerca da qualidade e da abrangência do conteúdo oferecido.

Nesse contexto, faz-se imperativo um estudo sistemático e detalhado dos catálogos dos principais serviços de *streaming* operantes no Brasil. Uma análise aprofundada das obras disponibilizadas por essas plataformas pode revelar *insights* significativos sobre tendências de mercado, preferências do público e até mesmo lacunas na oferta de conteúdo. Tal investigação assume uma relevância particular tanto para o público consumidor, ávido por conteúdos que ressoem com suas preferências pessoais e culturais, quanto para acadêmicos, profissionais da indústria do entretenimento e formuladores de políticas públicas interessados na promoção da diversidade e da inclusão cultural.

[3] "At its core, streaming seems to signify a limitless way of distributing and consuming media content. Corporate representatives and Internet activists alike underline what they argue is the transformative or disruptive nature of streaming."

A metodologia proposta para este estudo, que combina técnicas de *Web Scraping*, processos de ETL (*Extract*, *Transform*, *Load*) e análise de dados, representa uma abordagem eficaz para o mapeamento e a interpretação do vasto universo de dados gerados pelos serviços de *streaming*. Este enfoque metodológico permite a coleta de informações abrangentes e atualizadas, assim como a realização de análises qualitativas e quantitativas capazes de fornecer uma visão holística e crítica dos catálogos em questão. Portanto, este trabalho representa uma contribuição da área de Sistemas de Informação para o campo de estudos de mídia e comunicação, além de oferecer subsídios para a reflexão sobre as dinâmicas de consumo cultural na era digital.

1.2. Objetivos

Este trabalho de conclusão de curso está embasado na premissa de que a compreensão aprofundada dos catálogos dos serviços de *streaming* de filmes no Brasil, mediante a aplicação de técnicas avançadas de coleta e análise de dados, pode revelar *insights* significativos sobre as dinâmicas do mercado de entretenimento digital. Nesse sentido, delineiam-se os seguintes objetivos:

Objetivo Geral:

- Investigar a aplicabilidade das técnicas de *Web Scraping* e ETL (*Extract*, *Transform*, *Load*) na coleta, organização e análise dos dados dos catálogos de filmes oferecidos pelos principais serviços de *streaming* no Brasil, com o intuito de extrair tendências, padrões e particularidades que possam informar tanto consumidores quanto *stakeholders* da indústria cinematográfica.

Objetivos Específicos:

1. Mapear e caracterizar os catálogos dos principais serviços de *streaming* disponíveis no Brasil, identificando a diversidade de gêneros, origens (nacionais e internacionais) e classificações indicativas sugeridas;
2. Avaliar a representatividade e a diversidade dos catálogos em termos de inclusão de produções nacionais e independentes, bem como a presença de filmes que atendam a nichos específicos de público;
3. Contribuir para a literatura acadêmica na área de Sistemas de Informação e Estudos de Mídia;

4. Disponibilizar a base de dados compilada e processada, bem como o código utilizado neste estudo, para a comunidade acadêmica e profissional, possibilitando a realização de futuras pesquisas, o desenvolvimento de ferramentas de recomendação personalizada e a realização de análises comparativas entre diferentes plataformas de *streaming*.

1.3. Organização do texto

O presente trabalho está estruturado em capítulos, visando uma exploração clara e coerente dos métodos, análises e descobertas realizadas no âmbito da aplicação de técnicas de *Web Scraping* e ETL para a análise de dados de catálogos de filmes em serviços de *streaming* no Brasil. A estrutura do documento segue uma lógica sequencial, que facilita a compreensão do leitor sobre a complexidade e os resultados obtidos pela pesquisa. A organização textual é delineada conforme descrito a seguir:

Capítulo 1 — Introdução: este capítulo inicial estabelece o contexto e a relevância do estudo, apresentando a problemática investigada, os objetivos propostos e a justificativa para a escolha do tema. A introdução também delineia a estrutura geral do trabalho, fornecendo uma visão geral do que será abordado nos capítulos subsequentes;

Capítulo 2 — Conceitos preliminares: dedicado à fundamentação teórica, este capítulo aborda os conceitos essenciais que embasam a pesquisa, incluindo uma revisão sobre *Web Scraping* e ETL (*Extract*, *Transform*, *Load*). Além disso, discutem-se as técnicas e ferramentas utilizadas, buscando apresentar ao leitor um entendimento dos conceitos necessários para o desenvolvimento do trabalho;

Capítulo 3 — Plataformas exploradas: este capítulo apresenta um panorama das plataformas de *streaming* analisadas, bem como dos critérios e fontes utilizados para a avaliação dos filmes. A seleção das plataformas e a justificativa de sua inclusão são discutidas, evidenciando a relevância de cada uma para o escopo da pesquisa;

Capítulo 4 — Arquitetura do projeto: descreve a infraestrutura metodológica e tecnológica empregada na realização do trabalho, detalhando o processo de desenvolvimento do código-fonte e a arquitetura geral do sistema. Este capítulo

esclarece desde a extração dos dados até a organização e armazenamento destes em um Data Lake no banco de dados, ilustrando as etapas do processo;

Capítulo 5 — Organização do banco de dados: este capítulo aborda a organização e processamento dos dados no Data Lake utilizando o Google BigQuery. O processo inclui a importação, transformação, tradução e integração dos dados de várias fontes, resultando em uma estrutura padronizada e coesa para análises futuras. As etapas garantem a consistência e qualidade dos dados, facilitando uma abordagem centralizada e eficaz para o tratamento e análise das informações;

Capítulo 6 — Análise dos dados: focado nos resultados da investigação, o quinto capítulo expõe as análises realizadas a partir da base de dados construída. São apresentados gráficos, tabelas e demais instrumentos analíticos que revelam *insights* sobre os catálogos de filmes, incluindo tendências de gênero, preferências dos usuários e comparações entre as plataformas;

Capítulo 7 — Conclusões: reúne as considerações finais do estudo, enfatizando as principais contribuições e descobertas. Este capítulo também sugere direções para trabalhos futuros, indicando possibilidades de aprofundamento e expansão do conhecimento produzido.

2. Conceitos preliminares

Este capítulo visa a elucidar os conceitos fundamentais de *Web Scraping* e ETL (*Extract, Transform, Load*), técnicas cruciais empregadas neste trabalho para a coleta e processamento de dados dos catálogos de filmes disponíveis em serviços de *streaming* no Brasil. O leitor que já está familiarizado com esses conceitos pode pular este capítulo sem prejuízo para o entendimento do trabalho.

2.1 *Web Scraping*

No mundo digital em que vivemos, a vastidão e o crescimento constante das informações disponíveis na *web* apresentam tanto desafios quanto oportunidades. Dados dispersos pela internet podem revelar *insights* valiosos para uma ampla gama de campos, desde a pesquisa acadêmica até análises de mercado e desenvolvimento de produtos. Entretanto, a coleta manual desses dados pode ser demorada, sujeita a erros e impraticável para grandes volumes. Nesse cenário, o *Web Scraping* surge como uma técnica inovadora para a coleta automatizada de dados de páginas da *web*, permitindo a extração de dados relevantes, estruturadas e não estruturadas, de páginas *web*, por meio de programação e automação.

Ryan Mitchell, em seu livro *Web Scraping with Python: Collecting More Data from the Modern web* (2018), destaca a relevância do Web Scraping, afirmando que "se programar é magia, *Web Scraping* é feitiçaria: a aplicação da magia para feitos particularmente impressionantes e úteis — e surpreendentemente fáceis" (p. ix, tradução própria).[4] Essa é uma abordagem que simplifica a coleta de grandes volumes de dados e se estabelece como uma prática fundamental no contexto da pesquisa digital, marcada pela ascensão da internet e a ubiquidade dos dados digitais na vida social.

A capacidade singular do *Web Scraping* em extrair informações estruturadas de fontes on-line é particularmente enfatizada na literatura técnica sobre recuperação de informações. Mitchell explica que:

> Na teoria, *web scraping* é a prática de coleta de dados por qualquer meio que não seja um programa interagindo com uma API (ou, obviamente, através de um humano usando um navegador). Isso é mais comumente realizado escrevendo um programa automatizado que consulta um servidor *web*, solicita dados (geralmente na forma de HTML e outros arquivos que compõem páginas *web*), e então analisa esses dados para

[4] "If programming is magic, web scraping is wizardry: the application of magic for particularly impressive and useful — yet surprisingly effortless — feats."

> extrair as informações necessárias. (Mitchell, 2018, p. x, tradução própria).[5]

Essa observação sublinha o potencial do *Web Scraping* não apenas como um método para reunir dados massivos, mas também como uma técnica fundamental para a coleta automatizada de informações da *web*, viabilizando a análise e interpretação de grandes volumes de dados de maneira eficiente.

Para este trabalho, escolhemos a biblioteca Scrapy de Python devido à sua eficiência em coletar dados de páginas *web* complexas. Scrapy não só facilita a extração rápida e precisa de textos, imagens, *links* e tabelas, mas também se adapta bem à análise de estruturas HTML e ao uso de APIs (*Application Programming Interfaces*) para acessar dados no *back end* dos sites.

Esta escolha reflete a necessidade de uma ferramenta versátil que possa atender aos variados desafios de coleta de dados on-line, ajustando-se tanto à complexidade dos projetos quanto às especificidades dos dados desejados. O Scrapy se destaca por sua capacidade de adaptação a diferentes estruturas de páginas e APIs, garantindo uma coleta de dados abrangente e precisa.

A implementação de Scrapy e a exploração de suas técnicas de extração de dados serão detalhadas nas seções seguintes, destacando como ela se alinha aos objetivos da pesquisa, enfatizando sua importância na coleta e análise de grandes volumes de dados.

2.1.1. *Scrapy*

Scrapy[6] é uma biblioteca Python de código aberto, reconhecida por sua robustez e desempenho na automação de tarefas recorrentes de *Web Scraping*, facilitando a criação de *spiders* eficientes. Esses *spiders* automatizam a navegação por páginas *web* especificadas no seu código fonte, extraindo os dados relevantes e armazenando-os.

Uma das principais vantagens do Scrapy reside em sua capacidade de automatizar a coleta de URLs, compará-las com regras predefinidas, assegurar a unicidade dos URLs, normalizar URLs relativas quando necessário e aprofundar-se nas páginas através de processos recursivos. Ryan Mitchell, em seu livro *Web Scraping with Python: Collecting More Data from the Modern web* (2018), destaca a versatilidade do Scrapy, afirmando:

---

[5] "In theory, web scraping is the practice of gathering data through any means other than a program interacting with an API (or, obviously, through a human using a web browser). This is most commonly accomplished by writing an automated program that queries a web server, requests data (usually in the form of HTML and other files that compose web pages), and then parses that data to extract needed information."

[6] Disponível em: <https://github.com/scrapy/scrapy>. Acesso em: 19/07/2023.

> Scrapy é uma ferramenta poderosa que lida com muitos problemas associados ao rastreamento da web. Ela automaticamente reúne todos os URLs e os compara contra regras predefinidas, certifica-se de que todos os URLs sejam únicos, normaliza URLs relativas quando necessário, e recorre para ir mais profundamente nas página. […] Scrapy é uma biblioteca extremamente grande e abrangente com muitos recursos. Seus recursos trabalham juntos de maneira coesa, mas têm muitas áreas de sobreposição que permitem aos usuários desenvolver facilmente seu próprio estilo particular dentro dele. (Mitchell, 2018, p. 80-81, tradução própria)[7]

Esse enfoque em automatizar e gerenciar complexidades inerentes ao *Web Scraping* destaca o Scrapy como uma ferramenta de coleta de dados e também como uma solução integral para enfrentar desafios de rastreamento web. Sua arquitetura modular e o suporte a uma ampla gama de funcionalidades tornam o Scrapy adequado tanto para projetos simples quanto para aqueles de alta complexidade, permitindo personalização extensiva e adaptabilidade.

Distinguem-se as seguintes vantagens do Scrapy:

1. Robustez e Desempenho: Projetado para lidar com grandes volumes de dados e tarefas de *scraping* complexas, o Scrapy oferece recursos de gerenciamento de memória e tolerância a falhas, maximizando a eficiência através da execução assíncrona de solicitações;
2. Twisted: Baseado no framework assíncrono Twisted, Scrapy beneficia-se de melhorias significativas no desempenho e na eficiência, especialmente no manejo de chamadas HTTP assíncronas que aceleram a extração de dados;
3. Arquitetura Modular: Sua estrutura modular permite aos desenvolvedores personalizar e reutilizar componentes de forma eficiente. Com módulos responsáveis por tarefas específicas como *spiders*, pipelines e middlewares, Scrapy facilita a manutenção e amplia a escalabilidade do projeto;
4. Suporte a Requisições HTTP: Oferece suporte completo a solicitações e respostas HTTP, incluindo envio de cabeçalhos personalizados, autenticação e manipulação de cookies. Compatível com proxies, permite alterar o endereço IP para contornar restrições ou evitar bloqueios;

[7] "Scrapy is a powerful tool that handles many problems associated with crawling the web. It automatically gathers all URLs and compares them against predefined rules, makes sure all URLs are unique, normalizes relative URLs where needed, and recurses to go more deeply into pages. […] Scrapy is an extremely large and sprawling library with many features. Its features work together seamlessly, but have many areas of overlap that allow users to easily develop their own particular style within it."

5. Pipeline de Processamento de Dados: Proporciona um sistema flexível para o processamento de dados coletados, permitindo a limpeza, validação e armazenamento em diversos formatos como CSV, JSON, ou diretamente em bancos de dados, facilitando a integração do *Web Scraping* com etapas subsequentes de análise;
6. Suporte a Middleware: Inclui uma camada de middleware que possibilita a personalização do comportamento dos requests e responses, agregando funcionalidades como autenticação e manipulação de erros, o que torna o processo de *scraping* adaptável a vários cenários;
7. Escalabilidade: Scrapy suporta a execução simultânea de múltiplas *spiders*, otimizando o tempo de coleta de dados. Também permite o agendamento automático de tarefas de *scraping*, facilitando a programação e o controle das atividades de coleta em intervalos definidos.

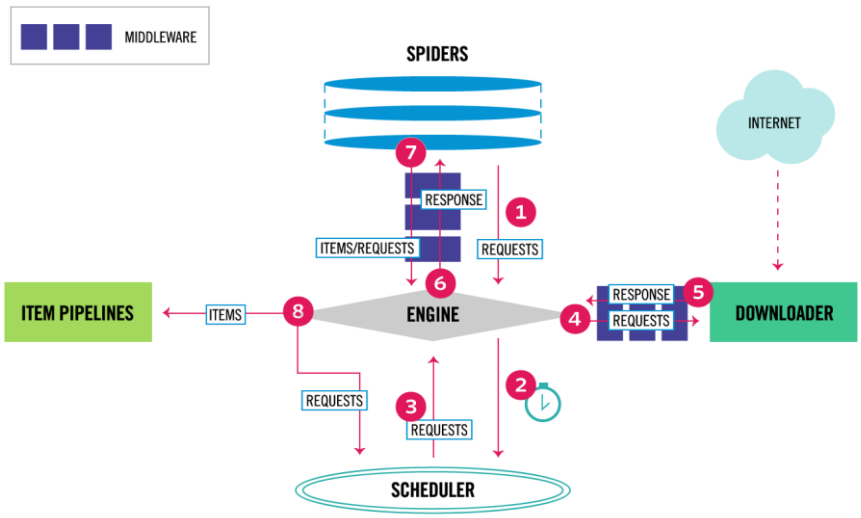
Em síntese, o Scrapy é uma escolha sólida para este trabalho, dada sua capacidade comprovada de otimizar o processo de *Web Scraping*. Sua arquitetura modular, junto aos recursos avançados e desempenho excepcional, torna o Scrapy uma ferramenta indispensável para a coleta e análise de dados automatizada na web, especialmente para investigar os catálogos dos principais serviços de *streaming* de filmes no Brasil.

A Figura 1 ilustra a estrutura geral da biblioteca Scrapy, em que é possível ver os módulos *Spiders, Scheduler, Downloader, Item Pipelines, Middleware,* e *Engine*. Estes módulos interagem entre si para realizar o processo de web scraping, coordenando a coleta, o processamento e a armazenagem de dados a partir de websites. Usando os números e os componentes principais indicados na figura:

1. *Requests* (Requisições): As requisições são enviadas das spiders para a engine. As spiders geram requisições HTTP que buscam páginas da web para extração de dados. Estas requisições são passadas à engine para serem processadas e enviadas ao próximo estágio.
2. *Scheduler* (Agendador): O agendador recebe as requisições da *engine* e organiza quais delas serão enviadas ao *downloader* para serem executadas. Ele funciona como uma fila, armazenando as requisições até que estejam prontas para serem baixadas.

3. *Requests* (Requisições): As requisições que o scheduler decide processar são enviadas de volta para a engine e depois repassadas ao downloader. Essa etapa representa a preparação e o envio das requisições organizadas pelo scheduler.
4. *Response*/*Requests* (Respostas/Requisições): Quando o *downloader* baixa a página da web solicitada pela requisição, ele envia a resposta de volta para a engine. A resposta contém o conteúdo da página baixada (HTML ou outro tipo de dado), que será passada para as spiders para extração de informações.
5. *Downloader* (Baixador): O downloader é responsável por baixar as páginas da web baseadas nas requisições enviadas pelas spiders. Ele converte as requisições HTTP em respostas que contenham os dados das páginas da web solicitadas.
6. *Engine* (Motor): A engine é o núcleo do Scrapy, que coordena o fluxo de dados entre todos os componentes. Ela faz a mediação entre as spiders, scheduler, downloader e item pipelines, gerenciando tanto requisições quanto respostas.
7. *Response* (Respostas): As respostas baixadas pelo downloader são entregues de volta às spiders através da engine. As spiders processam essas respostas para extrair os dados relevantes, que podem ser enviados para o próximo estágio.
8. *Items* (Itens): Os itens extraídos pelas spiders são enviados para o item pipeline, onde os dados são processados e, eventualmente, armazenados em uma base de dados ou exportados para outro formato.

Figura 1: Arquitetura da biblioteca Scrapy



Fonte: scrapy.org, 2024

2.2 ETL (*Extract, Transform, Load*)

No universo dos dados, os processos de extração, transformação e carregamento de dados, conhecidos como ETL, desempenham um papel fundamental na obtenção de informações significativas e acionáveis a partir de variadas fontes de dados. Segundo Kimball e Ross, "o sistema extração, transformação e carregamento (ETL) do ambiente DW/BI consiste em uma área de trabalho, estruturas de dados instanciadas e um conjunto de processos" (Kimball; Ross, 2013, p. 19, tradução própria)[8]. Ou seja, o ETL é um conjunto de procedimentos que envolve a coleta de dados brutos de diferentes fontes, a transformação desses dados em um formato consistente e útil e, por fim, o carregamento desses em um repositório final para análise e tomada de decisões.

Para Kimball e Ross, a complexidade e a importância do ETL podem ser comparadas à cozinha de um restaurante, onde matérias-primas são transformadas em pratos apetitosos:

> O sistema ETL é análogo à cozinha de um restaurante. A cozinha do restaurante é um mundo à parte. Chefs talentosos pegam matérias-primas e as transformam em refeições apetitosas e deliciosas para os clientes do restaurante. Mas muito antes de uma cozinha comercial entrar em operação, uma quantidade significativa de planejamento é investida no design do layout do espaço de trabalho e seus componentes. A cozinha é organizada com vários objetivos de design em mente. Primeiro, o layout deve ser altamente **eficiente**. […] Entregar qualidade **consistente** da cozinha do restaurante é o segundo objetivo importante. […] Por fim, o resultado da cozinha, as refeições entregues aos clientes do restaurante, também deve ser de alta **integridade**. (Kimball; Ross, 2013, p. 23, grifos e tradução própria)[9]

Assim como a cozinha precisa operar com eficiência e produzir refeições de qualidade de forma consistente, o sistema ETL deve processar os dados garantindo sua qualidade, consistência e integridade, preparando-os para análises confiáveis e suporte à decisão:

> O sistema ETL do *data warehouse* se assemelha à cozinha do restaurante. Os dados de origem são transformados magicamente em informações significativas e apresentáveis. O sistema ETL de bastidores deve ser planejado e arquitetado muito antes de qualquer dado ser extraído da fonte. Assim como a cozinha, o sistema ETL é

[8] "The *extract, transformation, and load* (ETL) system of the DW/BI environment consists of a work area, instantiated data structures, and a set of processes."
[9] "The ETL system is analogous to the kitchen of a restaurant. The restaurant's kitchen is a world unto itself. Talented chefs take raw materials and transform them into appetizing, delicious meals for the restaurant's diners. But long before a commercial kitchen swings into operation, a signifi cant amount of planning goes into designing the workspace layout and components."

> projetado para garantir o *throughput*. Ele deve transformar eficientemente os dados de origem brutos no modelo alvo, minimizando movimentos desnecessários.
> Obviamente, o sistema ETL também está altamente preocupado com a **qualidade**, **integridade** e **consistência** dos dados. Os dados que chegam são verificados quanto à qualidade razoável assim que entram. As condições são continuamente monitoradas para garantir que as saídas do ETL sejam de alta integridade. Regras de negócios para derivar métricas consistentemente e atributos de valor agregado são aplicadas uma vez por profissionais qualificados no sistema ETL, em vez de depender de cada usuário para desenvolvê-las independentemente. (Kimball; Ross, 2013, p. 24, grifos e tradução própria)[10]

O objetivo do ETL, portanto, transcende a simples coleta de dados, garantindo que tais dados brutos, crus e defeituosos sejam não apenas coletados, mas também limpos, organizados e transformados de maneira eficiente, confiável e íntegra, a fim de fornecer uma visão precisa e coerente dos dados para análise posterior. Isso é especialmente importante quando se lida com conjuntos de dados heterogêneos, provenientes de diferentes sistemas, formatos e estruturas.

O processo de ETL geralmente envolve as seguintes etapas:

1. **Extração (*Extract*):** Nesta etapa, os dados são coletados de diferentes fontes — bancos de dados, arquivos CSV, planilhas, APIs, serviços web, entre outros. Essa fase pode envolver tanto a coleta completa quanto seletiva de dados, dependendo das necessidades do projeto. É importante garantir a integridade dos dados extraídos, verificando suas qualidade, precisão e consistência;
2. **Transformação (*Transform*):** Após a extração, os dados brutos passam por uma fase de transformação, na qual são manipulados e convertidos para se adequar ao propósito analítico. Nessa etapa, as operações de limpeza, padronização, normalização, agregação, filtragem e enriquecimento dos dados podem ser aplicadas. A transformação visa garantir a consistência, a coerência e a qualidade dos dados, permitindo análises precisas e confiáveis;

[10] "The data warehouse's ETL system resembles the restaurant's kitchen. Source data is magically transformed into meaningful, presentable information. The back room ETL system must be laid out and architected long before any data is extracted from the source. Like the kitchen, the ETL system is designed to ensure throughput. It must transform raw source data into the target model effi ciently, minimizing unnecessary movement.

Obviously, the ETL system is also highly concerned about data quality, integrity, and consistency. Incoming data is checked for reasonable quality as it enters. Conditions are continually monitored to ensure ETL outputs are of high integrity. Business rules to consistently derive value-add metrics and attributes are applied once by skilled professionals in the ETL system rather than relying on each patron to develop them independently."

3. **Carregamento (*Load*):** Na etapa final, os dados transformados são carregados em um repositório final, como um banco de dados, um DL (*data lake*) ou uma aplicação de análise. Essa etapa envolve a organização estruturada dos dados transformados para facilitar a recuperação e o acesso eficiente. O método de carregamento, seja em tempo real ou em lotes, é escolhido com base nas especificidades do projeto.

Kimball e Ross enfatizam a necessidade de um sistema ETL bem projetado e executado, comparável à eficiência e à habilidade observadas na cozinha de um restaurante de alta qualidade. Essa analogia destaca o papel crítico do ETL não apenas na preparação dos dados para análise, mas também na garantia da qualidade e integridade dos dados no ambiente de DW/BI. Portanto, o processo ETL é indispensável para garantir a confiabilidade dos dados analisados e suportar efetivamente a tomada de decisões informadas em projetos de *business intelligence*, análise de dados e mineração de dados.

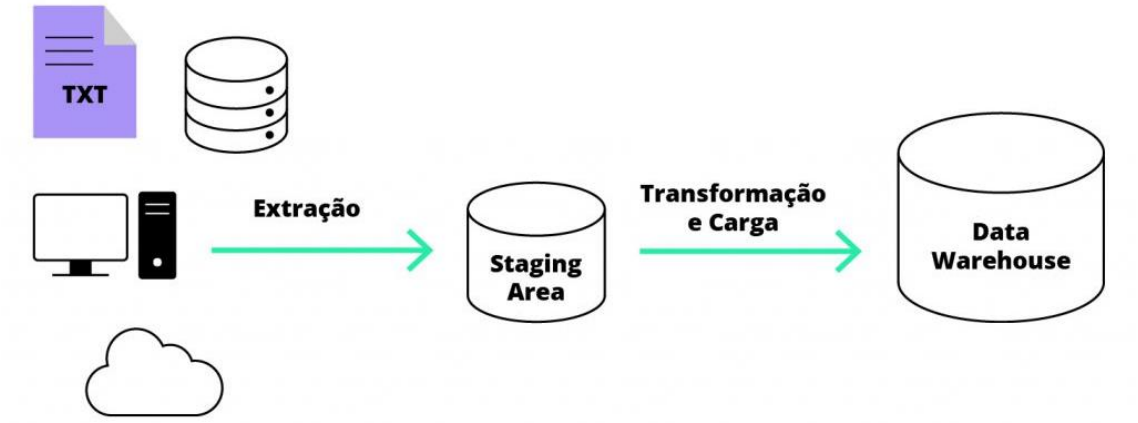


Figura 2: Processo de ETL

A Figura 2 ilustra um processo típico de ETL. Nesse exemplo, os dados são inicialmente extraídos de diversas fontes, como arquivos e bancos de dados, e são movidos para uma Staging Area. Nesta etapa, os dados são preparados e, em seguida, passam por um processo de transformação antes de serem carregados no Data Lake. Esse fluxo ilustra as principais etapas de um processo de ETL, essencial para a integração e preparação de dados em ambientes de Business Intelligence.

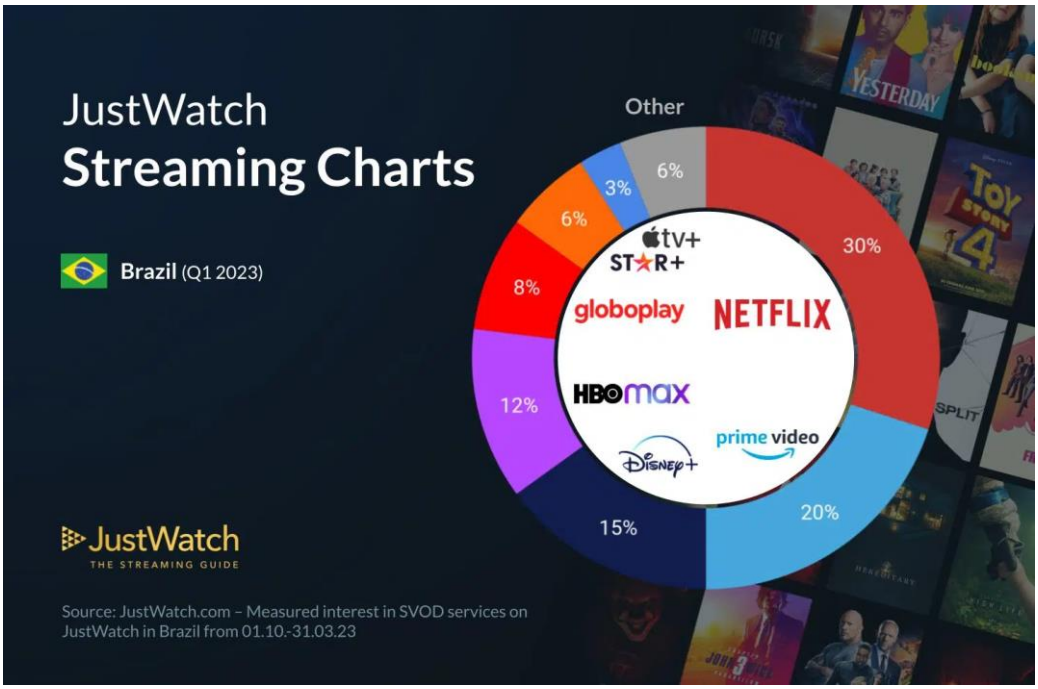
3. Plataformas exploradas

3.1. Plataformas de *streaming*

As plataformas de *streaming* transformaram a maneira como consumimos filmes e séries, tornando-se uma parte essencial do entretenimento digital no Brasil. A diversidade de conteúdo e a conveniência oferecida por esses serviços redefiniram os hábitos de consumo dos brasileiros. Uma pesquisa realizada pelo JustWatch, site que funciona como um buscador universal de filmes e séries, entre janeiro e março de 2023, destaca as preferências dos consumidores no país, ilustrando a competitividade e dinâmica deste mercado.

A Figura 3 demonstra o levantamento em um gráfico de pizza. A Netflix continua liderando com 30% do mercado, seguida pela Amazon Prime Video (20%), Disney+ (15%) e HBO Max (12%). O Globoplay, Star+ e Apple TV+ completam a lista com 8%, 6% e 3%, respectivamente. Notavelmente, o segmento de 'Outros' viu um aumento de um ponto percentual, sugerindo uma leve diversificação nas preferências dos usuários.

Figura 3: Gráfico da pesquisa do primeiro trimestre de 2023 com a porcentagem de usuários por serviço de *streaming* no mercado brasileiro



Fonte: macmagazine, 2023.

Bruno Cardoso aponta que a “Apple TV+ voltou a perder terreno […] Após ser ultrapassado pelo Star+ no último levantamento do JustWatch, o serviço da Maçã tem

agora ‘apenas’ metade da fatia de mercado da opção da Disney” (Cardoso, 2023). Esta análise reflete as flutuações e desafios enfrentados pelas plataformas em manter e expandir sua base de usuários no Brasil.

A pesquisa indica também uma tendência de crescimento para o HBO Max, potencialmente desafiando a posição do Disney+, que mostra sinais de declínio. Essas mudanças refletem um mercado em constante evolução, onde estratégias de conteúdo, preços e marketing são cruciais.

Esses *insights* esboçam o atual panorama das plataformas de *streaming* no Brasil e sublinham a importância de analisar as tendências do mercado para compreender as preferências dos consumidores. Isso nos permite discutir o impacto desses serviços no cenário do entretenimento digital brasileiro de maneira mais informada.

3.2. Plataforma de catálogo

A falta de transparência e acessibilidade dos catálogos das plataformas de *streaming* constituem um desafio significativo para pesquisas no âmbito do entretenimento digital. Embora sejam concorrentes diretos no mercado, estas plataformas compartilham uma estratégia comum de não divulgar publicamente seus catálogos completos. Essa prática obriga pesquisadores e usuários a recorrerem a fontes alternativas para obter informações detalhadas sobre a disponibilidade de conteúdo específico.

Neste contexto, o site JustWatch[11] emerge como uma ferramenta valiosa para a transposição dessas barreiras informativas. Atuando como o unico agregador de conteúdo de *streaming*, o JustWatch oferece uma visão abrangente dos catálogos das principais plataformas de *streaming*, adaptada à região geográfica do usuário. As informações disponiveis na ferramenta são sempre atualizadas e mudam de acordo com a disponibilidade dos filmes nos diferentes serviços. Por essa razão, escolheu-se o JustWatch como a principal fonte de dados para este trabalho, permitindo uma análise precisa e atualizada dos filmes disponíveis nas plataformas mais populares do Brasil.

A metodologia de coleta de dados empregada envolveu a utilização do JustWatch para extrair informações pertinentes aos catálogos de interesse. Isso incluiu a catalogação de filmes para cada uma das plataformas de streaming, além da coleta de dados complementares que enriquecem a análise, tais como classificação indicativa, duração, gênero e avaliações.

[11] Disponível em: < https://www.justwatch.com/br>. Acesso em: 30/09/24.

Utilizar o JustWatch como ferramenta de coleta de dados assegura a precisão e atualidade das informações, fundamentais para um estudo acadêmico rigoroso. Essa abordagem permite investigar as estratégias de conteúdo das plataformas e suas implicações para o mercado de entretenimento digital no Brasil, destacando a dinâmica competitiva e as tendências de consumo no país.

A Figura 4 mostra a tela de busca de filmes do JustWatch. Nela é possivel filtrar os filmes que estão presentes nos diferentes catalogos dos serviços de streaming. Nesse caso estão filtrados os serviçoes de streamming Netflix, Amazon Prime Video, HBO Max, Disney +, Globo Play, Star + e Apple TV +.

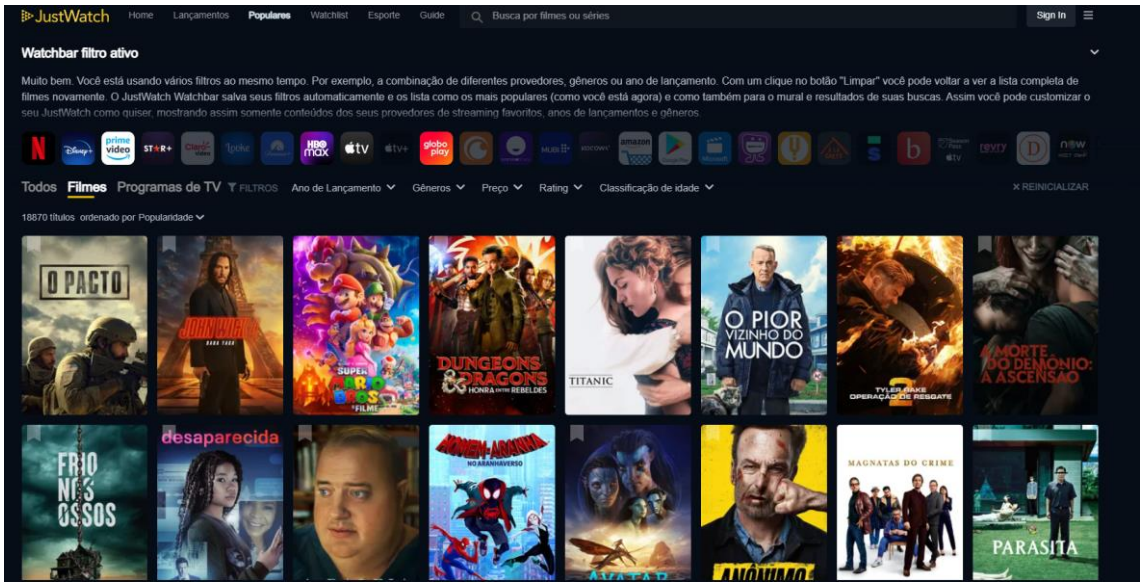


Figura 4: Página inicial do site JustWatch

3.3. Plataformas de avaliação

Uma análise profunda das avaliações de filmes requer acesso a dados diversificados e confiáveis, razão pela qual selecionamos múltiplas plataformas renomadas para extrair informações valiosas. Cada uma dessas fontes contribui com perspectivas únicas sobre a recepção dos filmes pelo público e pela crítica especializada:

1. JustWatch: Além das informações sobre os catálogos, o JustWatch oferece opiniões diretas dos usuários, refletindo a popularidade e a recepção dos títulos em seu catálogo;
2. IMDB (*Internet Movie Database*): Reconhecido globalmente, o IMDB fornece uma ampla base de avaliações de usuários, essencial para entender a aceitação internacional dos filmes. Os dados de avaliação dessa plataforma foram extraidos diretamente do Justwatch;

3. TMDB (*The Movie Database*): O TMDB complementa com notas dos usuários, enriquecendo a análise com mais uma camada de feedback do público. Os dados de avaliação dessa plataforma foram extraidos diretamente do Justwatch;
4. Rotten Tomatoes: Rotten Tomatoes é um site que se distingue pela agregação de críticas profissionais,além das opiniões do público, oferecendo uma medida da qualidade cinematográfica também baseada em opiniões especializadas;
5. Letterboxd: O Letterboxd é uma rede social dedicada a filmes que funciona como uma comunidade de cinéfilos, cujas classificações e comentários oferecem uma perspectiva apaixonada e pessoal sobre os filmes;
6. Filmow: O Filmow é uma plataforma brasileira que traz um olhar local sobre as obras, permitindo avaliar a recepção de títulos internacionais e nacionais pelo público brasileiro.

A combinação dessas plataformas garante uma base de dados robusta e multidimensional, permitindo uma avaliação abrangente do panorama cinematográfico atual. Através da análise dessas avaliações, pode-se identificar tendências de gosto, padrões de recepção crítica e diferenças culturais na apreciação dos filmes. Este processo tem potencial para enriquecer a compreensão da indústria cinematográfica e apoiar uma tomada de decisão mais informada pelos espectadores, produtores e distribuidores de conteúdo audiovisual.

Assim, a metodologia adotada para a seleção e extração de dados dessas plataformas é fundamental para alcançar os objetivos deste trabalho, buscando entender como as diversas obras são percebidas em diferentes contextos e por diferentes audiências.

4. Arquitetura da aplicação

Neste capítulo é descrita a arquitetura da aplicação desenvolvida, que é composta por diferentes camadas de um processo de *Web Scraping* usando a biblioteca scrapy e um processo de ETL utilizando o python.

4.1. Camada de extração (*extract*):

Na camada de extração da aplicação, foram desenvolvidas cinco *spiders* distintas, cada uma com uma lógica de navegação específica para as diferentes fontes de dados. Essas *spiders* atuam como "*crawlers*", explorando a *web* de a partir do codigo HTML e de requisições http em busca das informações desejadas, e são direcionadas a fontes de dados específicas das plataformas de *streaming* e fontes de dados de avaliações de filmes abordadas neste trabalho.

A lógica de cada *spider* foi cuidadosamente personalizada para atender às particularidades da fonte de dados em questão. Isso implica na definição de um conjunto único de passos para acessar e coletar os dados relevantes. Esses passos podem incluir a navegação por páginas *web*, o preenchimento de formulários, a interação com elementos dinâmicos e a extração de informações em diversos formatos, como texto, tabelas ou imagens.

O *framework* Scrapy proporciona uma estrutura robusta e flexível para a realização dessas tarefas. Ele permite a definição precisa de como as requisições devem ser feitas e como os dados devem ser extraídos. Em alguns casos, os dados são extraídos diretamente do DOM (*Document Object Model*) da página HTML; em outros, a extração é realizada através de requisições HTTP diretamente às APIs que fornecem os dados exibidos nos *websites*.

Quando uma *spider* faz uma requisição a uma página da *web*, o Scrapy gerencia a comunicação com o servidor e a obtenção do conteúdo HTML. Em seguida, a lógica da *spider* é ativada para processar o HTML e extrair as informações relevantes, como títulos, descrições e URLs. Esses dados são, então, estruturados de forma sistemática e preparados para a próxima etapa do processo, que é a transformação. Os dados extraídos são, então, escritos em arquivos CSV com o mesmo nome de sua respectiva spider e são salvos na *landing zone*, nesse caso, no servidor local de execução da aplicação.

A fim de evitar registros duplicados, foi configurado nos próprios parâmetros da biblioteca scrapy para evitar request de filmes que que já foram extraídos anteriormente.

Dessa forma assegura-se que não existirão erros ao tentar fazer *joins* das tabelas no banco de dados.

Esta camada inicial da aplicação é crucial, pois é responsável pela coleta dos dados brutos das diversas fontes. As informações extraídas são posteriormente processadas e transformadas na camada de transformação, em que são limpas, padronizadas e enriquecidas para serem adequadamente utilizadas nas etapas subsequentes de análise e carregamento.

A eficácia da camada de extração depende da precisão e da robustez das *spiders* desenvolvidas. Portanto, um design e uma implementação cuidadosos são essenciais para garantir que os dados coletados sejam de alta qualidade e adequados para as análises pretendidas. A flexibilidade do Scrapy e sua capacidade de lidar com diferentes formatos e estruturas de dados tornam-no uma ferramenta ideal para essa fase da aplicação, contribuindo significativamente para o sucesso da coleta de dados automatizada.

4.2. Camada de transformação (*transform*)

A camada de transformação desempenha um papel fundamental na estrutura do projeto, sendo responsável pelo refinamento e preparação dos dados coletados na camada de extração para a inserção na base de dados, a próxima etapa do processo.

Uma técnica comum de transformação de dados envolve o uso de consultas XPath. Quando os dados são extraídos diretamente de páginas HTML, o Scrapy permite a aplicação de consultas XPath para selecionar os elementos específicos que contêm as informações desejadas. Isso é particularmente útil para extrair alguns tipos de dados, como títulos, notas e datas, de maneira precisa e eficiente.

Além disso, quando os dados são obtidos por meio de chamadas de API(*Application Programming Interface*) , o Scrapy fornece acesso direto ao objeto *response*, facilitando a extração de dados no formato JSON ou XML. Utilizando métodos como *get*, é possível acessar os dados diretamente da resposta da API, tornando o processo de extração eficaz e direto.

Na camada de transformação é crucial tratar os dados para garantir que estejam no formato adequado para o estágio de carregamento. Isso inclui a limpeza de caracteres especiais, como vírgulas, que podem comprometer a integridade dos arquivos CSV criados nesta etapa na chamada *landing zone***,**onde os dados extraídos são salvos em arquivos CSV. Nessa etapa, acontece a normalização de dados, como a remoção de

acentos e espaços em branco. Isso é essencial para garantir a consistência dos dados, facilitando operações de junção (*join*) no banco de dados.

Por exemplo, no campo de título original do filme, foi realizada a normalização e criada uma nova coluna no arquivo CSV com os dados normalizados. A nova coluna além de deixar o titulo em caixa baixa, remove acentos entre outros caracteres. Esse cuidado permite juntar os dados de todas as fontes em uma tabela final, que será utilizada para análises futuras.

Portanto, a camada de transformação é responsável por refinar os dados coletados, tornando-os adequados para armazenamento e análise posterior. Esta etapa é crucial para assegurar a qualidade e a integridade dos dados ao longo de todo o processo, garantindo que os dados estejam prontos para fornecer *insights* valiosos na fase de análise.

4.3. Camada de carga (*load*)

A camada de carga representa a fase final da aplicação proposta, na qual os dados refinados na camada de transformação são enviados e armazenados nas soluções de nuvem oferecidas pelo Google Cloud Platform (GCP).[12] Esta etapa é crucial para garantir que os dados estejam organizados e acessíveis para análises futuras.

Primeiramente, o arquivo CSV gerado na etapa de transformação é enviado para o Google Cloud Storage. O Cloud Storage é um serviço de armazenamento de objetos escalável que permite armazenar, recuperar e gerenciar dados em uma infraestrutura global do Google. No projeto desenvolvido, cada arquivo é colocado em uma pasta nomeada com base na data de execução da *spider*, proporcionando organização e rastreabilidade dos dados ao longo do tempo

Após o armazenamento no Cloud Storage, . tabelas foram criadas s e os dados inseridos no Google BigQuery[13]. O BigQuery é um serviço de armazenamento e análise de dados em larga escala que permite executar consultas SQL complexas em grandes volumes de dados de maneira rápida e eficiente. A utilização do BigQuery possibilita que os dados sejam analisados de forma eficiente, fornecendo *insights* valiosos para a tomada de decisões.

A combinação desses serviços do GCP oferece uma solução escalável e poderosa para armazenar e analisar os dados coletados. O Google Cloud Storage fornece um armazenamento durável e acessível, enquanto o Google BigQuery permite a realização

[12] Disponível em: <https://console.cloud.google.com/>. Acesso em: 30/09/2024.
[13] Disponível em: <https://console.cloud.google.com/bigquery>. Acesso em: 30/09/2024.

de consultas e análises avançadas sobre esses dados. Essa abordagem não só facilita o gerenciamento dos dados, mas também garante que eles estejam disponíveis para análises complexas e detalhadas.

Além disso, o BigQuery oferece recursos avançados de segurança e controle de acesso, garantindo que os dados estejam protegidos contra acessos não autorizados. A capacidade de gerenciar permissões detalhadas e monitorar o uso dos dados é essencial para manter a integridade e a confidencialidade das informações armazenadas.

Em resumo, a camada de carregamento é responsável por enviar os dados coletados para a infraestrutura de nuvem do GCP, em que eles podem ser armazenados, consultados e analisados de forma eficiente. Essa abordagem aproveita os recursos da nuvem para proporcionar escalabilidade, segurança e acessibilidade, assegurando que os dados estejam bem-organizados e prontos para futuras análises.

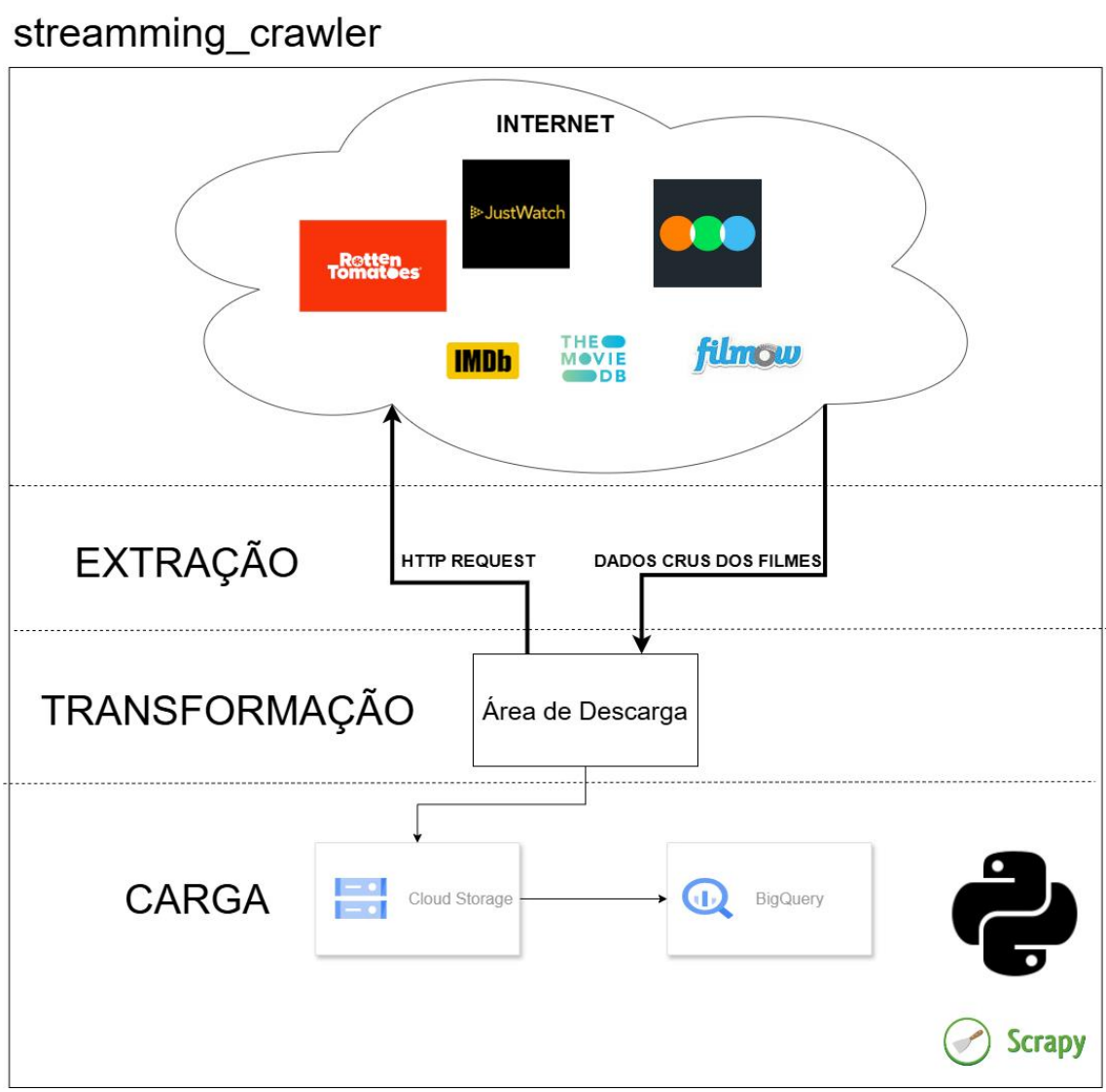


Figura 5: Arquitetura da aplicação

A Figura 5 ilustra a arquitetura do projeto que foi implementado em python com a biblioteca scrapy. A imagem é dividida em quatro partes. A camada da *internet* representa as fontes dos dados. Na camada de extração, o *crawler* envia requisições HTTP para essas fontes e coleta os dados crus dos filmes. Os dados brutos coletados e escritos em um arquivo CSV e então são enviados para a Área de Descarga, onde passam por processos de limpeza e transformação. Após a transformação, os dados processados são carregados e salvos no Cloud Storage e finalmente inseridos no BigQuery.

As *spiders* da aplicação foram executadas algumas vezes durante o processo de desenvolvimento, porém, os dados foram completamente extraídos cerca de 7 vezes entre o período de 21 de maio de 2023 até 10 de maio de 2024. Cada *spider* tinha seu próprio tempo de execução dependendo do volume de dados que precisava ser extraído.

O arquivo que executa a aplicação pode ser encontrado no repositório do Github.[14] Para rodar uma *spider* basta rodar no terminal o comando python3 main.py <nome_da_*spider*>. Dessa forma, todo o processo representado na Figura 5 é executado para aquela *spider*/fonte específica.

[14] Disponível em: < https://github.com/Davi98/streaming_crawler>. Acesso em: 30/09/2024.

5. Organização do banco de dados

Este capítulo descreve a organização e o processamento dos dados no *Data Lake*, utilizando o Google BigQuery como plataforma principal. O fluxo de dados compreende a importação, transformação, tradução e integração dos dados provenientes de diferentes fontes, resultando em uma estrutura coesa e pronta para análise.

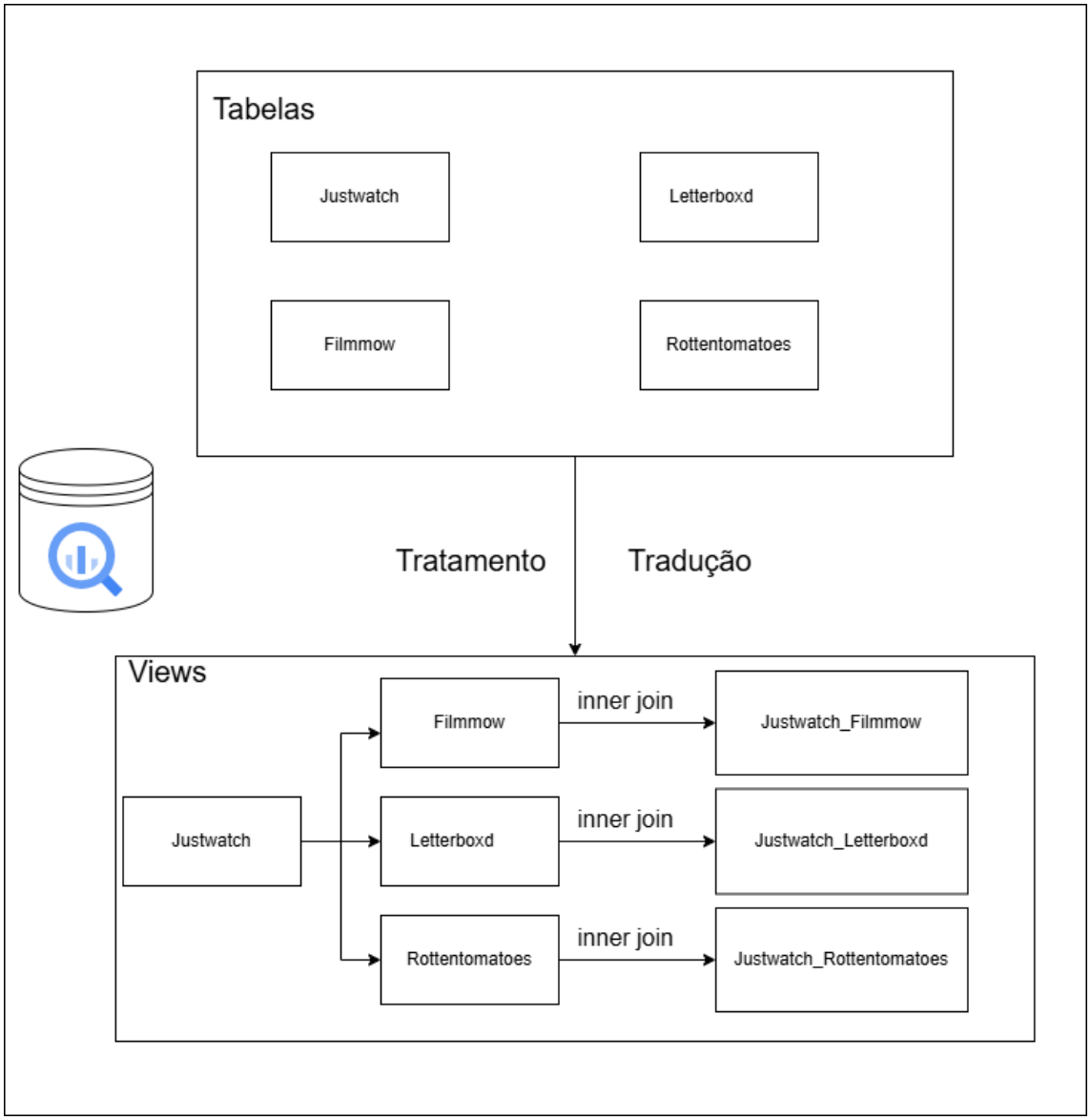


Figura 6: Arquitetura do Data Lake

A Figura 6 representa a organização do banco de dados Bigquery. A primeira etapa do processo é a importação das tabelas, em que os dados são carregados para o BigQuery e armazenados em um conjunto de dados denominado *data*. Dentro deste conjunto, existem tabelas específicas para cada uma das seguintes fontes de dados: JustWatch, Letterboxd, Filmow e Rotten Tomatoes.

Após a importação, cada tabela passa por um processo de transformação utilizando SQL. Durante esta fase, os dados são limpos, removendo caracteres indesejáveis, como caracteres especiais ou aspas, e corrigindo possíveis erros de formatação. Além disso, alguns tipos de dados são ajustados para garantir consistência, como a padronização de colunas de texto, a conversão de datas para um formato unificado e a transformação de números para os tipos corretos. Todas as colunas são traduzidas para o português, incluindo a conversão de abreviações para palavras completas. O resultado deste processamento são visões, que são representações transformadas e traduzidas das tabelas originais, substituindo as tabelas brutas e facilitando o uso posterior.

Algumas transformações nessa etapa foram necessárias após a inserção dos dados nas tabelas do BigQuery, como por exemplo a substituição de espaços em branco para facilitar o join, tradução das colunas e deixar todas as notas na mesma base (10) e até conversão de tipo de dados.

Como o objetivo deste trabalho é obter as avaliações dos filmes em diferentes plataformas voltadas ao propósito de avaliar serviços de *streaming*, realizou-se a junção da tabela JustWatch usando operações de *inner join* com as visões de cada um dos serviços. Esses *joins* utilizam o título original do filme e o ano de lançamento como chaves de junção, assegurando que os dados sejam corretamente alinhados. As operações de *inner join* resultam em novas visões integradas: Filmow é juntado com JustWatch, resultando em Justwatch_Filmow; Letterboxd é juntado com JustWatch, resultando em Justwatch_Letterboxd; e Rotten Tomatoes é juntado com JustWatch, resultando em Justwatch_Rottentomatoes.

O produto desse processo é um conjunto de visões integradas, em que cada visão representa uma combinação dos dados de JustWatch com outra fonte. Essas visões integradas são utilizadas para consultas e análises posteriores. Essa estrutura permite uma abordagem centralizada e padronizada para o tratamento dos dados, garantindo que todas as colunas estejam traduzidas e formatadas de maneira consistente em português. A limpeza dos dados e a mudança de tipos asseguram a qualidade e a integridade dos dados, enquanto o uso do título original e ano de lançamento para os *joins* garante a correta associação dos dados entre as diferentes fontes. Isso tende a facilitar o trabalho analítico o trabalho analítico e a obtenção de *insights* precisos.

Na Figura 7 é possivel observar o esquema lógico do banco de dados do projeto. O diagrama ilustra a modelagem de um banco de dados que agrega as informações de filmes provenientes de diversas plataformas de streaming e agregadores de avaliações,

como Justwatch, Letterboxd, Filmow e RottenTomatoes. A tabela Justwatch desempenha um papel central ao armazenar dados de filmes que estão disponíveis em múltiplos serviços de streaming. Para cada um dos sete serviços estudados, a tabela inclui campos booleanos que indicam se um filme está presente ou não em uma determinada plataforma, marcando true ou false de acordo com a disponibilidade. Essa abordagem elimina a necessidade de criar tabelas separadas para cada serviço, o que resultaria em registros duplicados, e simplifica tanto a modelagem quanto a manutenção do banco de dados. Além disso, a tabela permite identificar com precisão filmes exclusivos de cada plataforma, fornecendo uma visão detalhada e organizada da distribuição de conteúdos entre os serviços analisados.

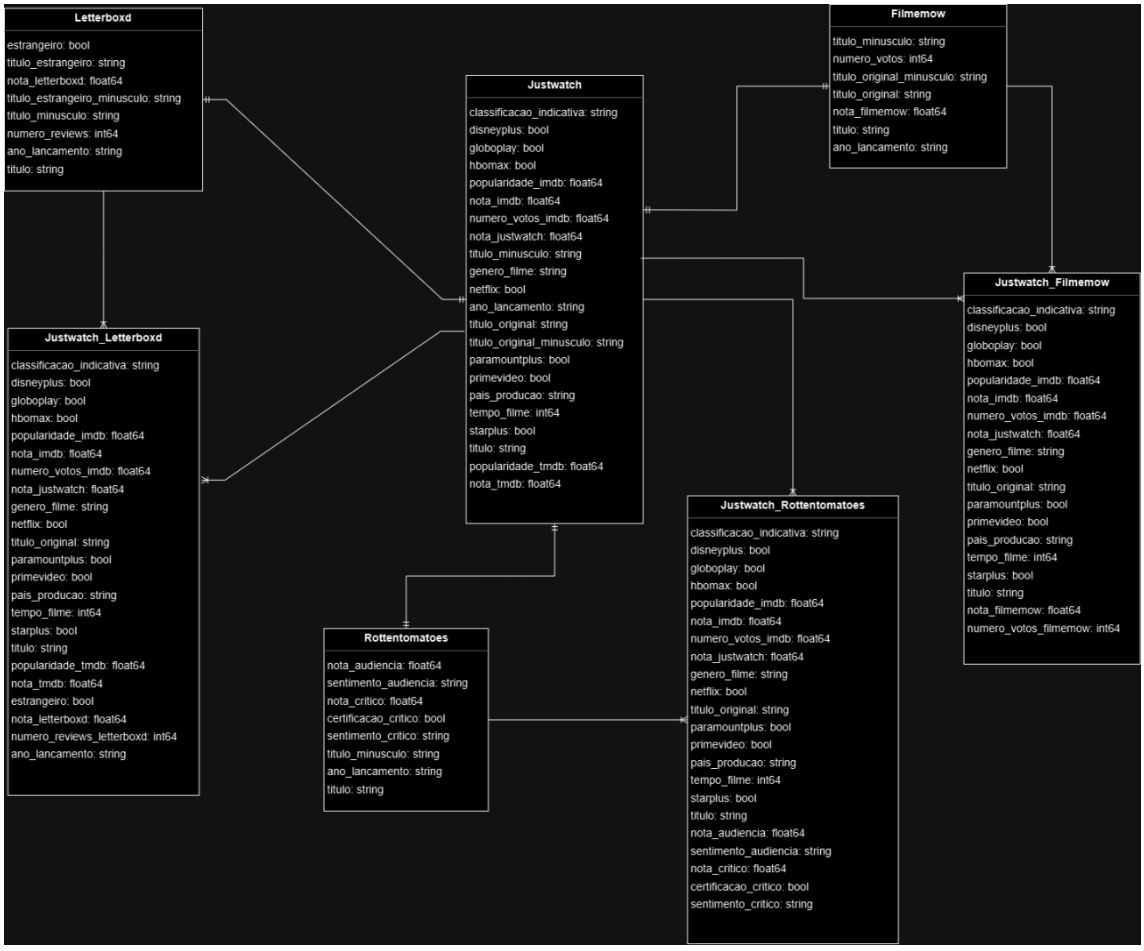


Figura 7: Modelo Relacional

```sql
SELECT
  CASE
    WHEN age_certification = 'L' THEN "0"
  ELSE
  age_certification
END
  AS classificacao_indicativa,
  disneyplus,
  globoplay,
  hbomax,
  imdbPopularity AS popularidade_imdb,
  imdbScore AS nota_imdb,
  imdbVotes AS numero_votos_imdb,
  ROUND(justwatchScore * 10, 1) AS nota_justwatch,
  lower_title AS titulo_minusculo,
  CASE
    WHEN movie_genre = 'act' THEN 'Ação'
    WHEN movie_genre = 'ani' THEN 'Animação'
    WHEN movie_genre = 'cmy' THEN 'Comédia'
    WHEN movie_genre = 'doc' THEN 'Documentário'
    WHEN movie_genre = 'drm' THEN 'Drama'
    WHEN movie_genre = 'fml' THEN 'Família'
    WHEN movie_genre = 'fnt' THEN 'Fantasia'
    WHEN movie_genre = 'hrr' THEN 'Horror'
    WHEN movie_genre = 'msc' THEN 'Musical'
    WHEN movie_genre = 'trl' THEN 'Thriller'
    WHEN movie_genre = 'scf' THEN 'Ficção Científica'
    WHEN movie_genre = 'crm' THEN 'Crime'
    WHEN movie_genre = 'hst' THEN 'História'
    WHEN movie_genre = 'rma' THEN 'Romance'
    WHEN movie_genre = 'spt' THEN 'Esporte'
    WHEN movie_genre = 'war' THEN 'Guerra'
    WHEN movie_genre = 'wsn' THEN 'Western'
    WHEN movie_genre = 'eur' THEN 'Europeu'
    WHEN movie_genre = 'rly' THEN 'Realidade'
  ELSE
  movie_genre
END
  AS genero_filme,
  netflix,
  CAST(original_release_year AS STRING) AS ano_lancamento,
  original_title AS titulo_original,
  LOWER(REPLACE(original_title, ' ', '')) AS titulo_original_minusculo,
  paramountplus,
  primevideo,
  CASE
    WHEN production_countrie IS NULL OR production_countrie = '' THEN 'US'
  ELSE
  production_countrie
END
  AS pais_producao,
  runtime AS tempo_filme,
  starplus,
  title AS titulo,
  tmdbPopularity AS popularidade_tmdb,
  ROUND(tmdbScore,1) AS nota_tmdb,
FROM
  `streamingsdata.data.justwatch`
```

Figura 8: *View* do JustWatch

A Figura 8 ilustra essas etapas, e o *script* SQL realiza as seguintes operações:

1. Tradução das colunas para português: no *script* SQL, as colunas e os atributos são traduzidos para português para facilitar a análise dos dados posteriormente;
2. Mapeamento de “L” para “0” na Classificação Indicativa: para facilitar a ordenação e compreensão dos dados, a classificação indicativa “L” (Livre para todos os públicos) foi mapeada para “0”. Isso permite uma ordenação mais clara e consistente, especialmente em casos de visualizações ou diagramas que dependem da classificação numérica. As demais classificações já estavam represetandas como números;
3. Mudança do gênero do filme de abreviação em inglês para palavra em português: O *script* SQL realiza a tradução dos códigos de gênero dos filmes, que anteriormente estavam em abreviações em inglês, para palavras em português. Por exemplo, “*act*” foi traduzido para “Ação”, “*ani*” para “Animação”, e assim por diante. Isso torna os dados e análises mais acessíveis e compreensíveis para usuários que falam português;
4. Atribuição do país de produção como Estados Unidos para filmes sem informação: como a coluna que representa o país de produção só é preenchida quando o filme é estrangeiro em relação às plataformas que se localizam nos Estados Unidos, foi atribuído o código dos Estados Unidos (US) para os casos em que essa informação está ausente. Isso padroniza os dados e facilita análises que dependem dessa informação;
5. Padronização das notas na mesma base: todas as notas de diferentes fontes de avaliação foram padronizadas para uma mesma base de 10 pontos e arredondadas para uma casa decimal Por exemplo, a nota do JustWatch, que varia continuamente de 0 a 1, foi multiplicada por 10 para alinhar com outras escalas de avaliação. Essa padronização permite comparações entre as notas de diferentes fontes e simplifica a análise dos dados;
6. Transformação do ano para *string*: o ano de lançamento foi transformado para o tipo *string*, pois não é utilizado para cálculos numéricos e o tipo string é mais indicado para operações de junção (*join*) entre tabelas.

Essas transformações garantem que os dados estejam limpos, padronizados e prontos para análises posteriores. A implementação dessas etapas é fundamental para assegurar a consistência e a qualidade dos dados no *Data Lake*.

```sql
SELECT
  FOREIGN AS estrangeiro,
  REPLACE(REPLACE(foreign_title, '', ''), '', '') AS titulo_estrangeiro,
  ROUND(letterboxd_grade*2,1) AS nota_letterboxd,
  REPLACE(REPLACE(REPLACE(lower_foreign_title, '', ''), ' ', ''), '', '') AS
  titulo_estrangeiro_minusculo,
  lower_title AS titulo_minusculo,
  number_reviews AS numero_reviews,
  CAST(original_release_year AS STRING) AS ano_lancamento,
  title AS titulo
FROM
  `streamingsdata.data.letterboxd`
```

Figura 9: *View* do Letterboxd

A tabela Letterboxd também passou por um processo de tradução e padronização. Na Figura 9, o *script* SQL utilizado realiza algumas operações importantes:

1. Tradução e limpeza de títulos: o *script* remove aspas simples dos títulos dos filmes e traduz as colunas para português. Por exemplo, "*foreign*" foi traduzido para "estrangeiro", e "*title*" para "titulo";
2. Padronização das notas: as notas foram convertidas de uma de uma escala de 1 a 5 para o intervalo contínuo de 1 a 10 e arredondadas para uma casa decimal, garantindo consistência na análise;
3. Transformação do ano para *string*: o ano de lançamento foi transformado em *string* para facilitar operações de junção e análise textual.

```sql
SELECT
  audienceScore/10 AS nota_audiencia,
  CASE
    WHEN audienceScore_sentiment = 'positive' THEN 'positivo'
    WHEN audienceScore_sentiment = 'negative' THEN 'negativo'
  ELSE
  audienceScore_sentiment
END
  AS sentimento_audiencia,
  criticsScore/10 AS nota_critico,
  criticsScore_certifiedAttribute AS certificacao_critico,
  CASE
    WHEN criticsScore_sentiment = 'positive' THEN 'positivo'
    WHEN criticsScore_sentiment = 'negative' THEN 'negativo'
  ELSE
  criticsScore_sentiment
END
  AS sentimento_critico,
  lower_title AS titulo_minusculo,
  CAST(original_release_year AS STRING) AS ano_lancamento,
  title AS titulo
FROM
  `streamingsdata.data.rottentomatoes`
```

Figura 10: *View* do Rottentomatoes

A tabela Rottentomatoes foi traduzida e padronizada utilizando o *script* SQL apresentado na Figura 10. As principais operações incluíram:

1. Padronização de notas e sentimentos: as notas da audiência e dos foram convertidas para uma escala contínua de 1 a 10 pontos. Por sua vez, os sentimentos (percepção obtida ao assistir) da audiência e dos críticos foram traduzidos para "positivo" e "negativo".
2. Transformação do ano para *string*: o ano de lançamento foi transformado para *string*, a fim de facilitar operações de junção e análises textuais.

```sql
SELECT
lower_title as titulo_minusculo,
number_votes as numero_votos,
original_lower_title as titulo_original_minusculo,
original_title as titulo_original,
score*2 as nota_filmow,
title as titulo,
CAST(year AS STRING) AS ano_lancamento
FROM `streamingsdata.data.filmow`
```

Figura 11: *View* do Filmow

Analogamente, o *script* SQL representado na Figura 11 para a tabela Filmow realizou as seguintes operações:

1. Tradução de colunas: os nomes das colunas foram traduzidos para o português, como "*lower_title"* para "titulo_minusculo" e "*number_votes"* para "numero_votos";
2. Padronização de notas: as notas foram convertidas de uma escala de 1 a 5 para uma escala contínua de 1 a 10, utilizando o cálculo *score* * 2, em que score representa a nota atual do filme, e renomeada para "nota_filmemow";
3. Transformação do ano para *string*: o ano de lançamento foi transformado para *string*, tornando mais fáceis as operações de junção e futuras análises textuais.

```sql
FROM
  `streamingsdata.views.Justwatch` jw
INNER JOIN
  `streamingsdata.views.Rottentomatoes`rtt
ON
  jw.ano_lancamento = rtt.ano_lancamento
  AND jw.titulo_original_minusculo = rtt.titulo_minusculo;
```

Figura 12: *Join*

O *script* SQL na Figura 12 realiza uma operação de *inner join* entre as visões Justwatch e Rottentomatoes. As junções (*joins*) foram feitas com base nos campos de título em minúsculas e do ano de lançamento do filme, a fim de garantir a correta associação dos dados entre as duas fontes.

```sql
FROM
  `streamingsdata.views.Justwatch` AS jw
JOIN
  `streamingsdata.views.Letterboxd` AS lb
ON
  (jw.titulo_minusculo = lb.titulo_minusculo
    AND NOT lb.estrangeiro)
  OR (jw.titulo_original_minusculo = lb.titulo_estrangeiro_minusculo
    AND lb.estrangeiro)
  AND jw.ano_lancamento = lb.ano_lancamento;
```

Figura 13: Junção de dados das *views* Justwatch e Letterboxd

O *script* SQL mostrado na Figura 13 realiza um *join* entre as visões Justwatch e Letterboxd. Neste, houve a adição de uma lógica para lidar com filmes estrangeiros. Quando o título em minúsculas (“titulo_minusculo”) da visão Letterboxd está vazio, a junção é feita utilizando o título original (“titulo_original_minusculo”). Além disso, a junção é feita utilizando o ano de lançamento.

```sql
FROM
  `streamingsdata.views.Justwatch` jw
INNER JOIN
  `streamingsdata.views.Rottentomatoes`rtt
ON
  jw.ano_lancamento = rtt.ano_lancamento
  AND jw.titulo_original_minusculo = rtt.titulo_minusculo;
```

Figura 14: Junção de dados das *views* Justwatch e Rottentomatoes

Finalmente, a Figura 14 exibe um script SQL que realiza um *inner join* entre as *views* Justwatch e Rottentomatoes, utilizando o título original em minúsculas e o ano de lançamento como chaves de junção para garantir a correspondência correta dos dados entre as duas fontes.

6. Análise dos dados

Após a definição do banco de dados e do desenvolvimento do projeto foi feita uma validação do projeto por meio da coleta e da análise de catálogos de filmes. Para tal, foram rodados spiders no período de 10 de abril a 10 de maio de 2024. Os dados e os gráficos das analises desse trabalho também estão presentes no repositorio do GitHub.[15]

Por meio do projeto desenvolvido, foi possível realizar análise dos dados coletados dos catálogos de filmes dos principais serviços de *streaming* no Brasil. Essas analises revelaram *insights* significativos sobre a qualidade e a recepção dos filmes disponíveis nessas plataformas. Através de técnicas de *Web Scraping* utilizando a biblioteca Scrapy, foi possível extrair dados detalhados de diferentes fontes on-line. Esses dados passaram por um processo de ETL (*Extract*, *Transform*, *Load*) para serem limpos, transformados e carregados em um banco de dados relacional, onde foram organizados e preparados para análise.

A ferramenta Metabase[16] desempenhou um papel crucial na visualização desses dados. Trata-se de uma plataforma de código aberto que permite a criação de *dashboards* interativos e relatórios detalhados, facilitando a análise comparativa entre as plataformas de *streaming*. Sua interface intuitiva e a capacidade de integração com diversos SGBDs (Sistemas de Gerenciamento de Banco de Dados Relacionais) tornam a Metabase uma escolha ideal para transformar dados complexos em informações claras e utilizáveis.

[15] Disponível em: < https://github.com/Davi98/streaming_crawler>. Acesso em: 30/09/2024.
[16] Disponível em: < https://www.metabase.com>. Acesso em: 30/09/2024.

Figura 15: Gráficos de médias gerais de avaliações dos filmes nos serviços de *streaming* IMDb, JustWatch, TMDB, Rotten Tomatoes (audiência e crítica), Filmow e Letterboxd.

Na Figura 15, é possivel vizualizar médias gerais de avaliações dos filmes, obtidas a partir das fontes IMDb, JustWatch, TMDB, Rotten Tomatoes (audiência e crítica), Filmow e Letterboxd. Observou-se que, no IMDb, as médias variam entre 5,9 e 6,48, com destaque para o Star Plus (6,48) , que apresenta a maior média, seguido de perto pelo Disney Plus (6,47). No JustWatch, o Star Plus se sobressai com uma média de 7,31, seguido pelo Globoplay com 7,16. No TMDb, as avaliações são mais uniformes, com o Paramount Plus alcançando a maior média, de 6,66.

Quando se analisam as notas de audiência no Rotten Tomatoes, a Disney Plus lidera com uma média de 7,25, demonstrando uma forte aceitação entre os espectadores. Nas avaliações da crítica no mesmo site, a Disney Plus novamente se destaca com uma média de 7,27, indicando uma alta qualidade percebida tanto pelo público quanto pela crítica. No Filmemow, as médias variam entre 5,91 e 6,64, com a Disney Plus novamente apresentando a maior média de 6,64. No Letterboxd, observa-se uma distribuição mais balanceada das avaliações, com o Star Plus alcançando a maior média de 6,37.

Essas análises indicam que algumas plataformas, como Disney Plus e Star Plus, conseguem manter uma performance consistente em relação a filmes em termos de

recepção tanto do público quanto da crítica. Esse desempenho pode ser atribuído à curadoria cuidadosa e à diversidade dos conteúdos oferecidos, evidenciando estratégias eficazes de satisfação do usuário.

Os resultados obtidos oferecem uma perspectiva sobre as estratégias adotadas pelos serviços de *streaming* para atender às demandas do mercado. A utilização da Metabase para visualização dos dados foi fundamental para identificar padrões e comparações relevantes, oferecendo uma base sólida para futuras investigações e decisões estratégicas no setor de *streaming* de filmes.

A tabela apresentada na Figura 16 destaca os 25 melhores filmes de acordo com as avaliações do IMDb e sua disponibilidade nos principais serviços de *streaming* no Brasil. Esta análise é crucial para entender como as plataformas se posicionam em termos de conteúdo de alta qualidade, frequentemente referenciado por críticos e público.

Top 25 pelo IMDB

Titulo Original	Nota Imdb	Netflix	Max	Globoplay	Disneyplus	Primevideo	Paramountplus	Starplus
Max Steel: The Dawn of Morphos	9.7	false	false	false	false	true	false	false
The Shawshank Redemption	9.3	false	true	false	false	false	false	false
The Godfather	9.2	true	false	false	false	false	true	true
Alexander Babu: Alex in Wonderland	9.2	false	false	false	false	true	false	false
Soy Luna: El último concierto	9.1	false	false	false	true	false	false	false
Zakir Khan: Tathastu	9.1	false	false	false	false	true	false	false
The Lord of the Rings: The Return of the King	9	false	true	false	false	true	false	false
Schindler's List	9	true	true	true	false	false	false	false
The Godfather Part II	9	false	false	true	false	false	true	true
Ressignificar	9	false	false	false	false	true	false	false

Figura 16: Tabela com os 25 melhores filmes segundo o IMDb e sua disponibilidade nos serviços de *streaming*.

A tabela da Figura 16 mostra que títulos como *The Shawshank Redemption* (no Brasil, *Um Sonho de Liberdade*) e *The Godfather* (no Brasil, *O Poderoso Chefão*) estão disponíveis em várias plataformas, como Netflix e Prime Video, refletindo uma estratégia de ampliação de alcance por meio de filmes aclamados. Por outro lado, filmes como *Max Steel: The Dawn of Morphos* (no Brasil, *Max Steel: Todo Poderoso Morphos*) e *Zakir Khan: Tathastu* possuem uma distribuição mais limitada, estando em apenas uma única plataforma e, destacando a exclusividade de certos conteúdos em plataformas específicas. A presença de filmes altamente avaliados em múltiplas plataformas pode indicar um esforço contínuo para atrair e manter assinantes através de um catálogo diversificado e de qualidade.

Para a análise das médias gerais de avaliações, foi utilizada uma série de consultas SQL combinando ‘UNION ALL’ para unificar os os resultados oriundos de tabelas das diferentes plataformas. A operação ‘AVG(nota_especifica)’ foi aplicada para calcular a média das notas dos filmes específicas de cada plataforma, enquanto a cláusula ‘WHERE plataforma = TRUE’ filtrava os registros para cada plataforma. Esse método permitiu apresentar a média geral de todos os filmes de cada plataforma lado a lado e ordená-las do menor para a maior com o comando ‘ORDER BY’, facilitando a visualização comparativa.

```sql
SELECT 'Prime Video' AS plataforma, AVG(nota_justwatch) AS media
FROM `views.Justwatch`
WHERE primevideo = TRUE
UNION ALL
SELECT 'Disney+' AS plataforma, AVG(nota_justwatch) AS media
FROM `views.Justwatch`
WHERE disneyplus = TRUE
UNION ALL
SELECT 'Globo Play' AS plataforma, AVG(nota_justwatch) AS media
FROM `views.Justwatch`
WHERE globoplay = TRUE
UNION ALL
SELECT 'HBO Max' AS plataforma, AVG(nota_justwatch) AS media
FROM `views.Justwatch`
WHERE hbomax = TRUE
UNION ALL
SELECT 'Netflix' AS plataforma, AVG(nota_justwatch) AS media
FROM `views.Justwatch`
WHERE netflix = TRUE
UNION ALL
SELECT 'Paramount+' AS plataforma, AVG(nota_justwatch) AS media
FROM `views.Justwatch`
WHERE paramountplus = TRUE
UNION ALL
SELECT 'Star+' AS plataforma, AVG(nota_justwatch) AS media
FROM `views.Justwatch`
WHERE starplus = TRUE
ORDER BY media ASC;
```

Figura 17: Consulta SQL que ilustra a lógica de combinação das médias das notas das plataformas.

Na Figura 17, a consulta SQL utiliza a função ‘AVG’ para calcular a média das notas no JustWatch dos filmes disponíveis em várias plataformas de *streaming*. A combinação dos resultados é feita através de ‘UNION ALL’, permitindo que as médias sejam visualizadas lado a lado. A ordenação final com ‘ORDER BY media ASC’ organiza os resultados do menor para o maior, melhorando a clareza visual dos dados como pode ser visto na Figura 15.

Todos os gráficos de barras apresentados na Figura 15 seguem essa lógica, alterando apenas a nota específica que é alvo da operação de ‘AVG’. O uso de ‘UNION ALL’ é crucial para unificar as médias das diversas plataformas, facilitando a análise comparativa. Adicionalmente, na criação da tabela dos top 25 filmes, o comando ‘LIMIT’ foi utilizado para restringir os resultados às 25 linhas mais bem avaliadas, ordenadas de forma descendente pela coluna ‘nota_imdb’ para destacar os filmes com as melhores avaliações.

Para ilustrar o processo de obtenção dos dados, utilizamos o exemplo da plataforma Netflix. As consultas SQL realizadas seguem um padrão similar para outras plataformas, com variações na cláusula WHERE para adaptar o filtro de acordo com a plataforma ou o critério de análise específico. A presença de um filme em uma plataforma é indicada por colunas booleanas; se a coluna correspondente ao serviço é verdadeira (‘TRUE’), o filme está disponível na plataforma, caso contrário, não está. Por exemplo, a consulta SQL na Figura 18 conta o número de filmes disponíveis na Netflix:

```sql
SELECT
  COUNT(*) AS `count`
FROM
  `views.Justwatch`
WHERE
  `views.Justwatch`.`netflix` = TRUE
```

Figura 18: Consulta SQL do número total de filmes disponíveis na Netflix.

A consulta da Figura 18 conta o número de registros na tabela Justwatch em que a coluna ‘netflix’ possui o valor TRUE, indicando que o filme está disponível na Netflix. Consultas similares foram realizadas para as demais plataformas, substituindo o filtro pela respectiva coluna de cada serviço. Esse dado é visualizado na Figura 19 que expressa o número total de filmes na Netflix, indicando a abrangência do catálogo da plataforma:

Figura 19: Número total de filmes no Netflix, filmes exclusivos e média de duração dos filmes.

A fim de fornecer uma visão mais detalhada e complementar às análises textuais, os gráficos completos para a Netflix e outras plataformas de *streaming* foram incluídos na seção Apêndice I ao final deste estudo. Esse apêndice permite uma análise visual mais profunda dos dados apresentados, facilitando a compreensão das diferenças e similaridades entre os catálogos das diversas plataformas de *streaming* avaliadas.

Para determinar o número de filmes exclusivos, foi utilizada uma consulta que verifica se os filmes estão presentes apenas na Netflix e não em outras plataformas. A exclusividade dos filmes foi determinada pela consulta SQL expressa na Figura 20:

```
SELECT
  COUNT(*) AS `count`
FROM
  `views.Justwatch`
WHERE
  (`views.Justwatch`.`netflix` = TRUE)

   AND (`views.Justwatch`.`starplus` = FALSE)
  AND (`views.Justwatch`.`globoplay` = FALSE)
  AND (`views.Justwatch`.`hbomax` = FALSE)
  AND (`views.Justwatch`.`paramountplus` = FALSE)
  AND (`views.Justwatch`.`primevideo` = FALSE)
  AND (`views.Justwatch`.`disneyplus` = FALSE)
```

Figura 20: Consulta SQL do número de filmes exclusivos na Netflix.

A consulta da Figura 20 conta quantos registros na tabela Justwatch possuem a coluna ‘netflix’ como verdadeira (‘TRUE’), e todas as outras colunas de plataformas de *streaming*, como ‘starplus’, ‘globoplay’, ‘hbomax’, ‘paramountplus’, ‘primevideo’ e ‘disneyplus’, marcadas como falsas (‘FALSE’). Dessa forma, é possível conferir os filmes disponíveis exclusivamente na Netflix ao se filtrarem os registros em que a coluna ‘netflix’ é verdadeira e todas as outras colunas referentes às plataformas são falsas.

Para calcular a média da duração dos filmes, usamos a consulta SQL exibida na Figura 21:

```
SELECT
  AVG(`views.Justwatch`.`tempo_filme`) AS `avg`
FROM
  `views.Justwatch`
WHERE
  `views.Justwatch`.`netflix` = TRUE
```

Figura 21: Consulta SQL da média de duração dos filmes na Netflix.

A consulta na Figura 21 utiliza a função ‘AVG’ para calcular a média do tempo de duração dos filmes (‘tempo_filme’) disponíveis na Netflix, considerando apenas os registros em que a coluna ‘netflix’ é verdadeira.

A análise das médias das notas dos filmes segue o mesmo princípio, com consultas SQL que aplicam a função ‘AVG’ às colunas de avaliação, como ‘nota_imdb’, ‘nota_justwatch’, entre outras. Por exemplo, para calcular a média das notas no IMDb para filmes disponíveis na Netflix, a consulta que pode ser vista na Figura 22 é estruturada da seguinte forma:

```sql
SELECT
  AVG(`views.Justwatch`.`nota_imdb`) AS `avg`
FROM
  `views.Justwatch`
WHERE
  `views.Justwatch`.`netflix` = TRUE
```

Figura 22: Consulta SQL da média das notas no IMDb para filmes na Netflix.

Outro exemplo é a média das notas no JustWatch, obtida com a seguinte consulta representada na Figura 23:

```sql
SELECT
  AVG(`views.Justwatch`.`nota_justwatch`) AS `avg`
FROM
  `views.Justwatch`
WHERE
  `views.Justwatch`.`netflix` = TRUE
```

Figura 23: Consulta SQL da média da nota no JustWatch para filmes na Netflix.

As consultas retratadas nas Figuras 22 e 23 consideram apenas os registros em que a coluna ‘netflix’ é verdadeira (‘TRUE’) e calculam as médias das notas atribuídas pelos usuários. A mesma lógica é aplicada para as demais plataformas e métricas, adaptando o nome da coluna na cláusula ‘WHERE’ conforme necessário. Isso permite calcular o número total de filmes, a média de duração, e as diferentes médias de avaliações (IMDb, TMDb, Letterboxd, entre outras) para cada serviço de *streaming* analisado.

A Figura 24 ilustra as médias das notas de diferentes fontes de avaliação (IMDb, JustWatch, Rotten Tomatoes, Filmow, Letterboxd), permitindo uma comparação visual clara das avaliações de usuários e críticos. Na Figura 24 é possivel ver todas as médias lado a lado. É importante lembrar que as médias são calculadas de acordo com a quantidade de filmes presentes na vizualização, ou seja, não são todos os filmes que possuem uma nota do Rottentomatoes visto que esses dados vem de um *inner join* de uma outra vizualização.

Figura 24: Médias das notas do Rotten Tomatoes (audiência e crítica), Filmow, JustWatch, IMDb, TMDB e Letterboxd dos filmes da Netflix.

Além das notas, a análise dos sentimentos das críticas e das audiências, obtidos da fontes do Rotten Tomatoes, é crucial para entender a percepção qualitativa dos filmes. Os gráficos de pizza representandos na Figura 25, mostram a proporção de sentimentos positivos e negativos, são gerados a partir de consultas pelo valor dos sentimentos para um dado serviço de streaming. Por exemplo, o gráfico à direita da Figura 25 mostra que 58% dos filmes possuem audiência postiva da Netflix, segundo o Rotten Tomatoes, contra 48% que possuem um sentimento negativo.

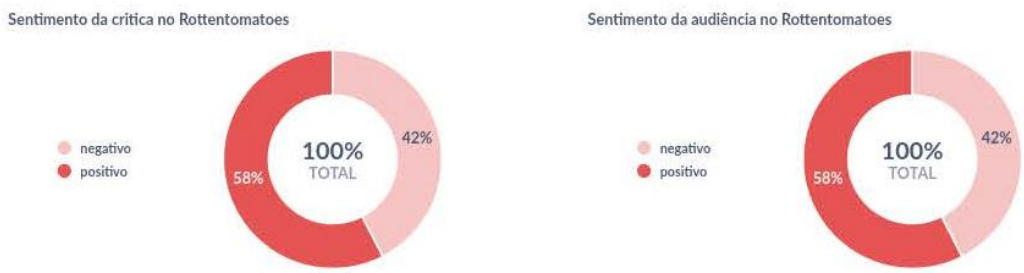


Figura 25: Gráficos de pizza dos sentimentos da audiência e da crítica no Rotten Tomatoes para filmes na Netflix.

A consulta para obtenção dos sentimentos da audiência no Rotten Tomatoes é mostrada na Figura 26.

```
SELECT
  `views.Justwatch_Rottentomatoes`.`sentimento_audiencia` AS `sentimento_audiencia`,
  COUNT(*) AS `count`
FROM
  `views.Justwatch_Rottentomatoes`
WHERE
  (`views.Justwatch_Rottentomatoes`.`netflix` = TRUE)

   AND (
    `views.Justwatch_Rottentomatoes`.`sentimento_audiencia` IS NOT NULL
  )
  AND (
    (
      `views.Justwatch_Rottentomatoes`.`sentimento_audiencia` <> ''
    )

    OR (
      `views.Justwatch_Rottentomatoes`.`sentimento_audiencia` IS NULL
    )
  )
GROUP BY
  `sentimento_audiencia`
ORDER BY
  `sentimento_audiencia` ASC
```

Figura 26: Consulta SQL do sentimento da audiência do Rotten Tomatoes para filmes na Netflix.

A consulta para obtenção dos sentimentos da audiência no Rotten Tomatoes que é mostrada na Figura 26 conta o número de linhas na coluna ‘sentimento_audiencia’ da tabela ‘Justwatch_Rottentomatoes’, filtrando apenas os registro pertencentes à Netflix,

isto é, cuja coluna ‘netflix’ possui valor verdadeiro (‘TRUE’). Assim, os dados são agrupados pelos sentimentos atribuídos pelos espectadores, que podem corresponder a ‘positivo’ ou ‘negativo’. Uma consulta semelhante é utilizada para sentimentos da crítica, conforme ilustrado pela Figura 27, permitindo uma análise abrangente da recepção dos filmes oferecidos pela Netflix:

```
SELECT
  `views.Justwatch_Rottentomatoes`.`sentimento_critico` AS `sentimento_critico`,
  COUNT(*) AS `count`
FROM
  `views.Justwatch_Rottentomatoes`
WHERE
  (`views.Justwatch_Rottentomatoes`.`netflix` = TRUE)

   AND (
    `views.Justwatch_Rottentomatoes`.`sentimento_critico` IS NOT NULL
  )
  AND (
    (
      `views.Justwatch_Rottentomatoes`.`sentimento_critico` <> ''
    )

    OR (
      `views.Justwatch_Rottentomatoes`.`sentimento_critico` IS NULL
    )
  )
GROUP BY
  `sentimento_critico`
ORDER BY
  `sentimento_critico` ASC
```

Figura 27: Consulta SQL do sentimento da crítica do Rotten Tomatoes para filmes na Netflix.

Para entender como as classificações indicativas dos filmes são distribuídas na plataforma Netflix, utilizamos uma consulta SQL que agrupa os filmes de acordo com suas classificações, conforme mostrado na Figura 28:

```
SELECT
  `views.Justwatch`.`classificacao_indicativa` AS `classificacao_indicativa`,
  COUNT(*) AS `count`
FROM
  `views.Justwatch`
WHERE
  (`views.Justwatch`.`netflix` = TRUE)

   AND (
    `views.Justwatch`.`classificacao_indicativa` IS NOT NULL
  )
  AND (
    (`views.Justwatch`.`classificacao_indicativa` <> '')

    OR (
      `views.Justwatch`.`classificacao_indicativa` IS NULL
    )
  )
GROUP BY
  `classificacao_indicativa`
ORDER BY
  `classificacao_indicativa` ASC
```

Figura 28: Consulta SQL da distribuição da classificação indicativa para filmes na Netflix.

A sequência de comandos SQL da Figura 28 agrupa os registros por classificação indicativa, podendo ser, 0(Livre para todos os públicos), 10, 12, 14, 16 e 18 anos, facilitando assim a visualização da distribuição das classificações. A visualização desses dados através de um gráfico de barras (histograma) nos permite identificar facilmente quais classificações são mais comuns entre os títulos disponíveis, conforme é possível ver na Figura 29. Nessa, percebe-se que a classificação mais frequente corresponde à faixa etária de 16 anos, que possui 645 ocorrências, seguida pela faixa etária de 14 anos, que possui 591 ocorrências.:

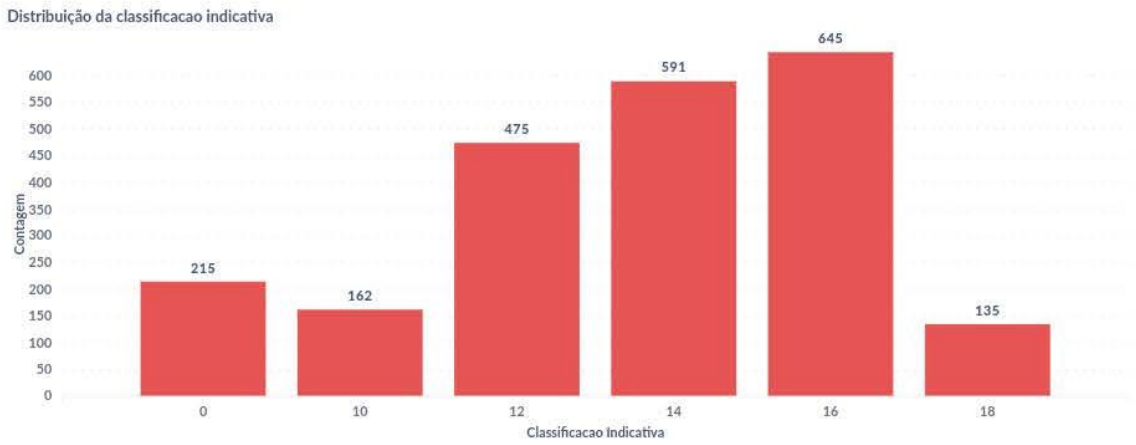


Figura 29: Gráfico de barras ilustrando a distribuição da classificação indicativa dos filmes na Netflix.

Além disso, a análise da variedade de gêneros disponíveis na Netflix, realizada através de uma consulta SQL representando na Figura 30 que conta e agrupa os filmes de acordo com seus gêneros, permite visualizar quais tipos de filmes são mais predominantes na plataforma (Figura 30):

```sql
SELECT
  `views.Justwatch`.`genero_filme` AS `genero_filme`,
  COUNT(*) AS `count`
FROM
  `views.Justwatch`
WHERE
  (`views.Justwatch`.`netflix` = TRUE)

   AND (`views.Justwatch`.`genero_filme` IS NOT NULL)
  AND (
    (`views.Justwatch`.`genero_filme` <> '')

    OR (`views.Justwatch`.`genero_filme` IS NULL)
  )
GROUP BY
  `genero_filme`
ORDER BY
  `count` ASC,
  `genero_filme` ASC
```

Figura 30: Consulta SQL da distribuição de gênero dos filmes na Netflix.

Analogamente ao caso anterior, a sequência de comandos SQL retratada na Figura 30 conta o número de registros na tabela ‘Justwatch’ em que a coluna ‘netflix’ é verdadeira (‘TRUE’), de modo a filtrar pelos filmes oferecidos por esse serviço de streaming. Dessa maneira, são agrupados os registros obtidos pelo valor da coluna ‘genero_filme’, desconsiderando-se os registros que possuem o valor dessa coluna como nulo ou uma string vazia. Os resultados são então ordenados primeiro pelo número de registros, obtido pelo comando count, em ordem ascendente e, em caso de empate no número de registros, então pelo valor de ‘genero_filme’ em ordem alfabética. É possível então ter uma visão clara da diversidade do catálogo, conforme expresso pela Figura 31. Nessa, pode-se constatar que o genêro mais frequente é a comédia, com 1.043 registros, seguido pelo drama, com 942 registros. Por outro lado, o gênero menos frequente corresponde ao western, representado por apenas 10 filmes:

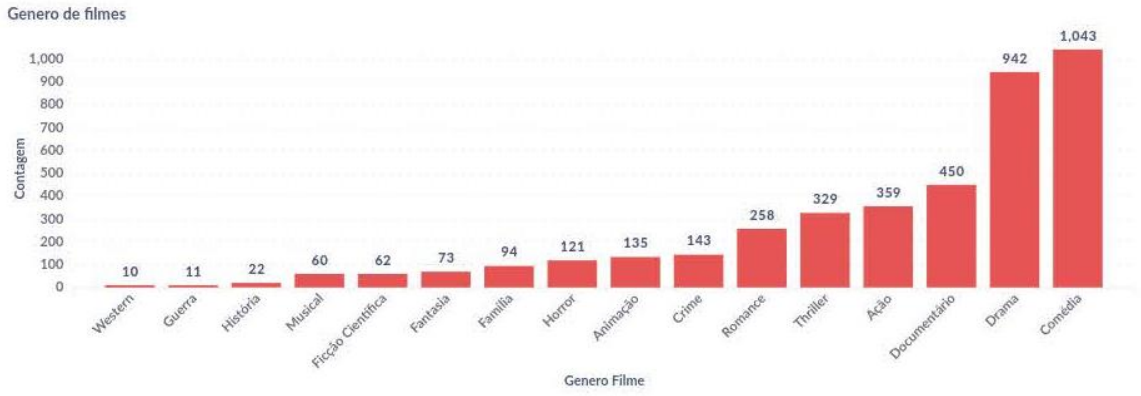


Figura 31: Gráfico ilustrando a diversidade de gêneros de filmes no catálogo da Netflix.

Para entender melhor como diferentes gêneros de filmes são avaliados na Netflix, calculou-se a média das notas do IMDb para cada genêro, conforme Figura 32.

```
SELECT
  `views.Justwatch`.`genero_filme` AS `genero_filme`,
  AVG(`views.Justwatch`.`nota_imdb`) AS `avg`
FROM
  `views.Justwatch`
WHERE
  (`views.Justwatch`.`netflix` = TRUE)

   AND (`views.Justwatch`.`genero_filme` IS NOT NULL)
  AND (
    (`views.Justwatch`.`genero_filme` <> '')

    OR (`views.Justwatch`.`genero_filme` IS NULL)
  )
GROUP BY
  `genero_filme`
ORDER BY
  `avg` ASC,
  `genero_filme` ASC
```

Figura 32: Consulta SQL da média da nota do IMDb por gênero para filmes na Netflix.

A consulta contida na Figura 32 calcula a média por genêro das notas dos filmes da Netflix a partir do IMDb, ajudando a identificar quais gêneros tendem a receber melhores avaliações dos usuários. Para tal, a consulta agrupa os resultados pelo valor da coluna ‘genero_filme’, descartando aqueles para os quais o valor de ‘genero_filme’ é nulo ou é uma string vazia. Os registros são então ordenados pela média (‘AVG’) das notas IMDb em ordem ascendente e, em caso de empate na média, pelo valor de ‘genero_filme’ em ordem alfabética. Os resultados da consulta são mostrados na Figura 33, em que é possível verificar que o gênero que atinge a maior nota corresponde ao de guerra, com uma média de 7.14, seguido pelo documentário, que possui média de 6.98:

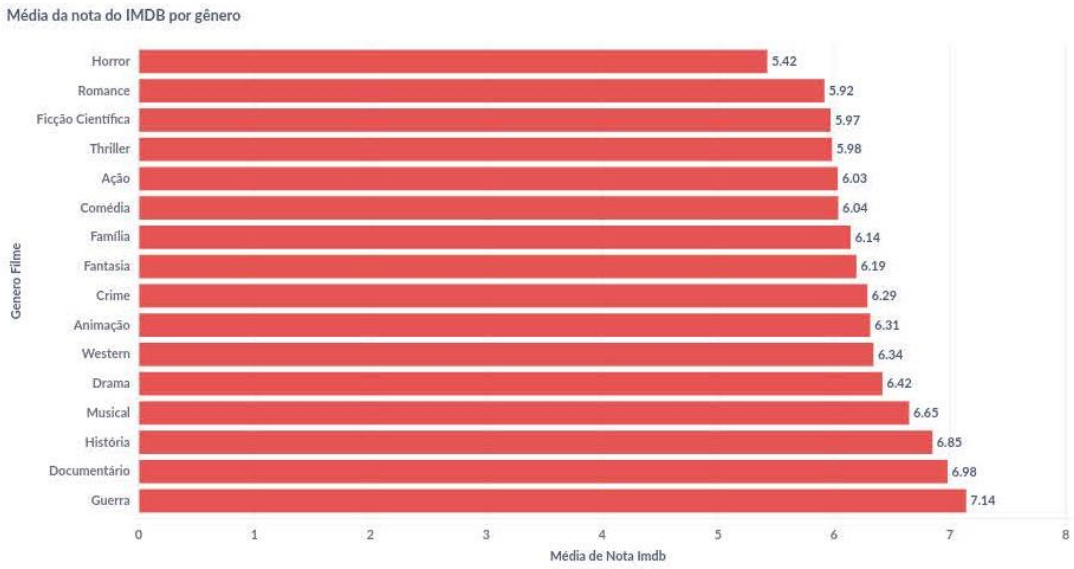


Figura 33: Gráfico ilustrando a média da nota do IMDb por gênero dos filmes na Netflix.

Já a análise da distribuição dos filmes por país de produção na Netflix fornece *insights* sobre a diversidade geográfica do conteúdo disponível na plataforma. Utilizando um gráfico de mapa-múndi, é possível visualizar a contagem de filmes por país de maneira clara e intuitiva, tal como mostrado na Figura 34, em que os países que mais produzem obras estão associados a cores mais intensas e os que menos produzem, a cores mais fracas. O país que mais produz são os Estados Unidos, seguido da Índia:

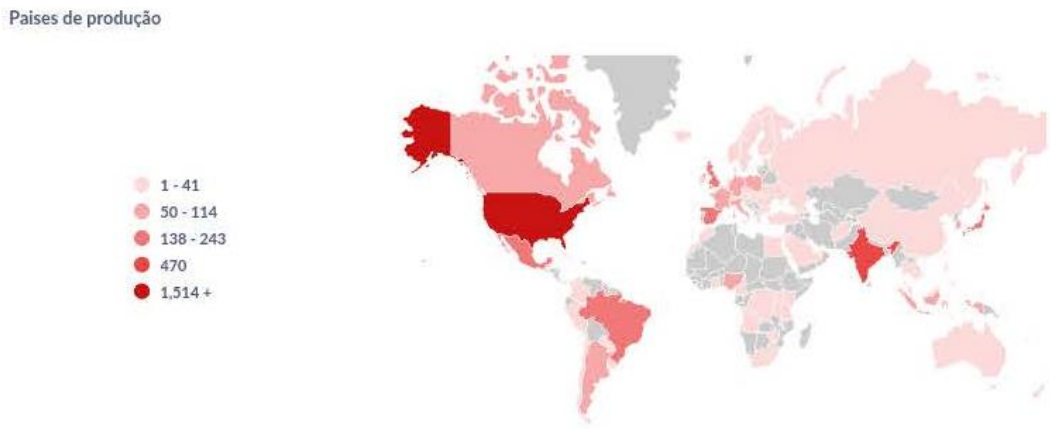


Figura 34: Gráfico de mapa-múndi ilustrando a distribuição por país de produção para filmes na Netflix..

A consulta que organiza os filmes por país de produção, destacando a origem geográfica dos títulos está descrita na Figura 35:

```
SELECT
  `views.Justwatch`.`pais_producao` AS `pais_producao`,
  COUNT(*) AS `count`
FROM
  `views.Justwatch`
WHERE
  (`views.Justwatch`.`netflix` = TRUE)

   AND (`views.Justwatch`.`pais_producao` IS NOT NULL)
  AND (
    (`views.Justwatch`.`pais_producao` <> '')

    OR (`views.Justwatch`.`pais_producao` IS NULL)
  )
GROUP BY
  `pais_producao`
ORDER BY
  `pais_producao` ASC
```

Figura 35: Consulta SQL da distribuição por país de produção para filmes na Netflix.

A consulta na Figura 35 conta o número de registros na tabela ‘Justwatch’ em que a coluna ‘netflix’ é verdadeira (‘TRUE’), e agrupa essa contagem a coluna “pais_producao”. Dessa forma é possivel contar o número de filmes por pais de produção.

Por sua vez, a fim de se analisar a distribuição dos filmes disponíveis na Netflix por ano de lançamento, utilizou-se uma consulta SQL que agrupa os filmes de acordo com o ano em que foram lançados por meio da coluna ‘ano_lancamento’ (Figura 36):

```sql
SELECT
  `views.Justwatch`.`ano_lancamento` AS `ano_lancamento`,
  COUNT(*) AS `count`
FROM
  `views.Justwatch`
WHERE
  (`views.Justwatch`.`netflix` = TRUE)

   AND (`views.Justwatch`.`ano_lancamento` IS NOT NULL)
  AND (
    (`views.Justwatch`.`ano_lancamento` <> '')

    OR (`views.Justwatch`.`ano_lancamento` IS NULL)
  )
GROUP BY
  `ano_lancamento`
ORDER BY
  `ano_lancamento` ASC
```

Figura 36: Consulta SQL da distribuição por ano de lançamento para filmes na Netflix.

A consulta da Figura 35 realiza o filtro onde a coluna “netflix” é verdadeira(TRUE) , conta o número de registros e agrupo pela coluna “ano_lancamento”. Dessa forma, é possivel consultar o número de filmes por cada ano de lancamento prsentes na plataforma.

Essa análise ajuda a entender a evolução do catálogo da plataforma ao longo do tempo, mostrando como a oferta de filmes cresceu e diversificou. A Figura 37 exibe tal evolução do catálogo da Netflix por meio de uma série temporal, em que é possível identificar que o ano de 2022 é o que possui mais filmes no catalago da netflix, totalizando 563 filmes.

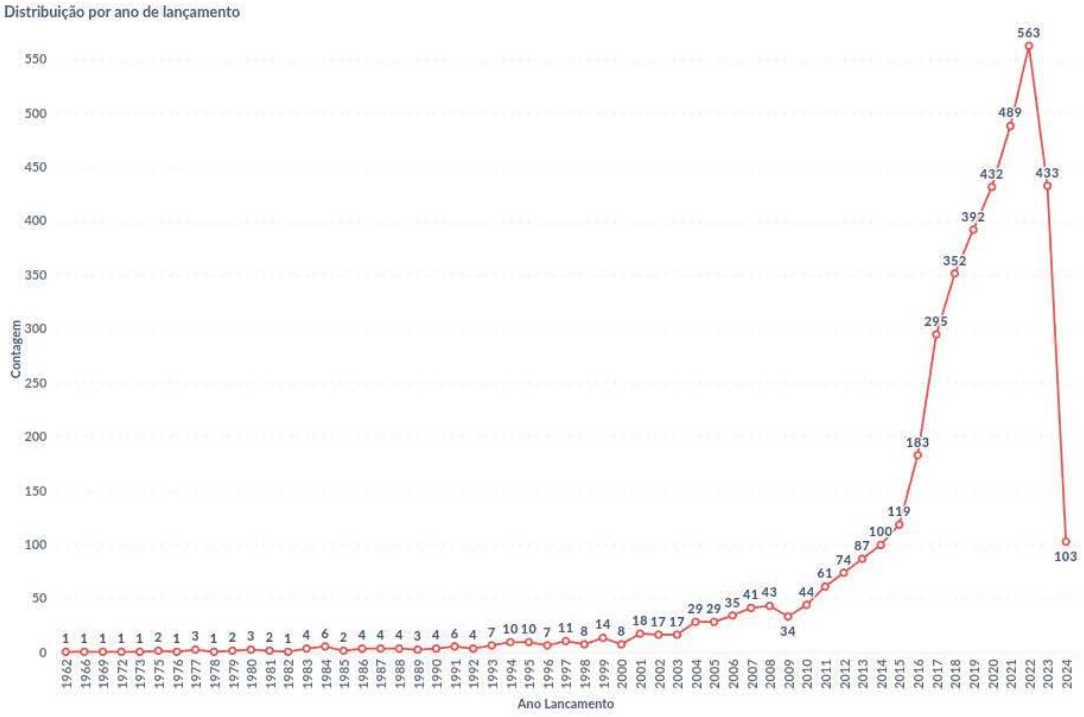


Figura 37: Gráfico ilustrando a distribuição de filmes na Netflix por ano de lançamento.

Por fim, para identificar os filmes mais bem avaliados na Netflix de acordo com o IMDb, foram selecionados os 10 títulos com as maiores notas, tal como ilustrado pela Figura 38:

```sql
SELECT
  `views.Justwatch`.`classificacao_indicativa` AS `classificacao_indicativa`,
  `views.Justwatch`.`titulo_original` AS `titulo_original`,
  `views.Justwatch`.`nota_imdb` AS `nota_imdb`,
  `views.Justwatch`.`pais_producao` AS `pais_producao`,
  `views.Justwatch`.`genero_filme` AS `genero_filme`
FROM
  `views.Justwatch`
WHERE
  `views.Justwatch`.`netflix` = TRUE
ORDER BY
  `nota_imdb` DESC
LIMIT
  10
```

Figura 38: Consulta SQL do Top 10 IMDb para filmes na Netflix.

A consulta 38 seleciona algumas colunas da tabela Justwatch e filtra os filmes da Netflix deixando a coluna netflix como verdadeira. Além disso, a query ordena os filmes pela “nota_imdb” de forma decrescente e limita aparecer apenas 10 registros, assim, os 10 filmes com as melhores notas no IMDb apareceram no resultado da consulta como pode ser visto na Figura 39.

A consulta da Figura 38 destaca os filmes de maior destaque na Netflix, segundo a avaliação dos usuários do IMDb, oferecendo uma visão clara dos títulos mais valorizados pelos espectadores. Assim, o filme com a maior nota na plataforma corresponde a "Lista de Schindler", com uma nota de 9.2:

Top 10 IMDB

Titulo Original	Nota Imdb	Pais Producao	Genero Filme
The Godfather	9.2	US	Drama
Schindler's List	9	US	Drama
12th Fail	9	IN	Drama
Pulp Fiction	8.9	US	Crime
David Attenborough: A Life on Our Planet	8.9	GB	Drama
Fight Club	8.8	DE	Drama
Forrest Gump	8.8	US	Comédia
Bo Burnham: Inside	8.7	US	Comédia
Ingoma	8.7	ZA	Drama
Chhota Bheem Maha Shaitaan Ka Mahayudh	8.7	IN	Família

Figura 39: Gráfico ilustrando o Top 10 IMDb para filmes na Netflix

As análises do catálogo da Netflix, que englobaram desde a quantidade total de filmes e exclusividade de títulos até a avaliação por diferentes fontes e a diversidade de gêneros, revelaram uma ampla variedade de conteúdo. Essa diversidade é reflexo das estratégias de curadoria da plataforma, evidenciadas pela presença de títulos exclusivos e bem avaliados.

A metodologia de análise aplicada, que combinou técnicas de ETL com visualizações detalhadas em gráficos, foi eficaz na exploração dos dados de *streaming*. Segundo Wang *et al*. (2023), "o sucesso da tomada de decisão baseada em dados depende da qualidade dos dados, que se refere ao grau de dados utilizáveis" (Wang *et al*., 2023, p. 1161).[17] Isso reforça a importância de garantir a integridade e precisão dos dados utilizados nas análises, especialmente quando se trata de decisões estratégicas sobre curadoria de conteúdo e personalização dos serviços de *streaming* que são determinantes para esse tipo de negócio.

Os gráficos completos, disponíveis na seção Apêndice I, permitem uma análise mais aprofundada e facilitam a comparação entre os catálogos das diversas plataformas de *streaming* avaliadas. Embora o foco deste capítulo tenha sido a Netflix, o mesmo rigor

[17] "The success of data-driven decision-making depends on data quality, which refers to the degree of usable data."

metodológico foi aplicado às análises dos catálogos de outras plataformas de *streaming*, cujos resultados detalhados também podem ser encontrados no Apêndice I. Esta pesquisa, além de esclarecer a composição do catálogo da Netflix, estabelece uma base sólida para futuras pesquisas no campo dos serviços de *streaming*, contribuindo para a compreensão das dinâmicas e tendências deste mercado em constante evolução.

7. Conclusão

7.1. Considerações finais

A presente pesquisa explorou os catálogos de filmes dos principais serviços de *streaming* no Brasil, utilizando técnicas de *Web Scraping* e ETL para revelar a amplitude e a diversidade desses acervos. A análise mostrou que plataformas como a Netflix oferecem uma vasta gama de títulos, não apenas em termos de quantidade, mas também em diversidade de gêneros, classificações indicativas e origens geográficas. Essa variedade reflete o esforço das plataformas em atender a um público diverso, buscando oferecer conteúdo que ressoe com diferentes gostos e preferências.

Este estudo esclarece as práticas de curadoria e a oferta de conteúdo das plataformas de *streaming* no Brasil, além de fornecer um modelo metodológico útil para outras pesquisas na área de mídia digital. A transparência na análise de dados e a acessibilidade das informações são fundamentais para entender o impacto cultural e econômico dessas plataformas na sociedade atual.

Os dados coletados e analisados pelo projeto agora serão democratizados, permitindo que qualquer usuário tenha visibilidade completa de todo o catálogo de filmes dos principais serviços de streaming sem a necessidade de assinar essas plataformas. Essa abertura facilita o acesso às informações sobre títulos disponíveis, suas avaliações e outras características, promovendo transparência e permitindo que os usuários tomem decisões informadas sobre os serviços que desejam utilizar.

Os dados sobre as avaliações dos usuários e da crítica especializada revelam uma preocupação constante das plataformas em manter um catálogo de alta qualidade. Filmes bem avaliados, presentes em várias plataformas, indicam uma estratégia clara de expansão de público e fidelização de assinantes, demonstrando a importância de um conteúdo atraente e variado.

A metodologia utilizada, que combinou a extração dos dados usando a biblioteca Scrapy com ferramentas de visualização como o Metabase, provou ser eficaz para uma análise detalhada e comparativa entre os serviços. A inclusão de gráficos no Apêndice I deste trabalho permite uma análise mais visual e direta, facilitando a identificação de padrões e tendências. Isso tem o potencial de auxiliar consumidores e profissionais do setor, assim como contribuir para uma compreensão mais profunda das estratégias de conteúdo das plataformas.

Spilker e Colbjørnsen (2020) argumentam que o *streaming* é um fenômeno em constante evolução, com práticas que vão além da simples transmissão de conteúdo. Eles destacam que "as tecnologias de *streaming* não são apenas sistemas tecnológicos fechados: os mercados ainda são imaturos, e as soluções ainda estão abertas para evoluir" (Spilker; Colbjørnsen, 2020, p. 1216).[18] Isso significa que as plataformas de *streaming* precisam adaptar constantemente seus modelos de negócio e estratégias de curadoria para atender a uma audiência diversificada e exigente (Spilker; Colbjørnsen, 2020, p. 1222).[19] Essa observação é corroborada pelas análises realizadas neste estudo, que revelam um esforço contínuo das plataformas em expandir e diversificar seus catálogos, respondendo às demandas do mercado e às preferências dos usuários.

É importante notar que o estudo enfrentou algumas limitações, como a dependência de fontes externas de dados e possíveis variações nos diferentes critérios de avaliação dos portais de crítica. Esses fatores sugerem que as conclusões aqui apresentadas devem ser vistas como uma parte de um quadro maior e em constante evolução. O monitoramento contínuo e a atualização dos dados são essenciais para acompanhar as mudanças no mercado de *streaming*.

7.2. Trabalhos futuros

Para futuras pesquisas, seria valioso aprofundar a investigação sobre como as estratégias de curadoria de conteúdo impactam a satisfação e o comportamento do público. Estudos que incluam abordagens qualitativas podem complementar os dados quantitativos e oferecer uma visão mais completa do mercado. Além disso, a expansão para analisar outros formatos de conteúdo, como séries e documentários, poderia enriquecer ainda mais a compreensão sobre o papel das plataformas de *streaming*.

Além disso, seria interessante contornar algumas limitações como a melhora na coleta de dados de algumas fontes e conseguir dados também de fontes que não foram contempladas no projeto.

[18] "Part of the rationale for this approach is that streaming technologies are not closed off as technological systems: Markets are still immature, and solutions are yet open to evolve."
[19] "Streaming is a field where the strategies and practices of the parties involved are rapidly evolving and shifting."

8. Referências bibliográficas

CARDOSO, Bruno. *Apple TV+ fecha 1º trimestre com 3% do mercado no Brasil.* MacMagazine, 19 abril 2023. Disponível em: <https://macmagazine.com.br/post/2023/04/19/apple-tv-fecha-1o-trimestre-com-3-do-mercado-no-brasil/>. Acesso em: 27/02/2024.

CASTELLS, Manuel. *Communication Power*. Oxford University Press, 2009.

DAVI98. *Streaming Crawler* [repositório de código]. Disponível em: < https://github.com/Davi98/streaming_crawler>. Acesso em: 30/09/2024.

JUSTWATCH. *JustWatch Brasil.* Disponível em: https://www.justwatch.com/br. Acesso em: 19/07/2023.

KIMBALL, Ralph; ROSS, Margy. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley, 2013.

MACMAGAZINE. *JustWatch Streaming Charts.* Disponível em: <https://macmagazine.com.br/wp-content/uploads/2023/04/19-justwatch-q1-2023.png>. Acesso em: 28/03/2024.

MATOS, Thaís. *Streaming ganha ainda mais força, após eventos serem cancelados devido ao coronavírus*. G1, 17 mar. 2020. Disponível em: <https://g1.globo.com/pop-arte/noticia/2020/03/17/*streaming*-ganha-ainda-mais-forca-nos-eua-apos-eventos-serem-cancelados-devido-ao-coronavirus.ghtml>. Acesso em: 05/04/2024.

MITCHELL, Ryan. *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly Media, 2018.

NERY, Carmen. *Em 2022, streaming estava presente em 43,4% dos domicílios com TV.* Agência de Notícias do IBGE, 09 nov. 2023. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/38306-em-2022-*streaming*-estava-presente-em-43-4-dos-domicilios-com-tv>. Acesso em: 27/02/2024.

SCRAPY.ORG. *Scrapy.* Disponível em: https://doc.scrapy.org/en/latest/_images/scrapy_architecture_02.png. Acesso em: 26/02/2024.

SPILKER, H. S.; Colbjørnsen, T. The dimensions of streaming: toward a typology of an evolving concept. *Media, Culture & Society*, v. 42, n. 7–8, p. 1210–1225, 2020.

Disponível em: <https://doi.org/10.1177/0163443720904587>. Acesso em: 24/06/2024.

WANG, J.; LIU, Y.; LI, P. *et al*. Overview of data quality: examining the dimensions, antecedents, and impacts of data quality. *Journal of Knowledge Economy*, v. 15, p. 1159–1178, 2024. Disponível em: <https://doi.org/10.1007/s13132-022-01096-6>. Acesso em: 24/06/2024.

.

9. Apêndice I

9.1. Gráfico Análise Geral

Figura 40: Gráfico Analise Geral

9.2. Gráfico Amazon Prime

Figura 41: Gráfico Amazon Prime

9.3. Gráfico Disney Plus

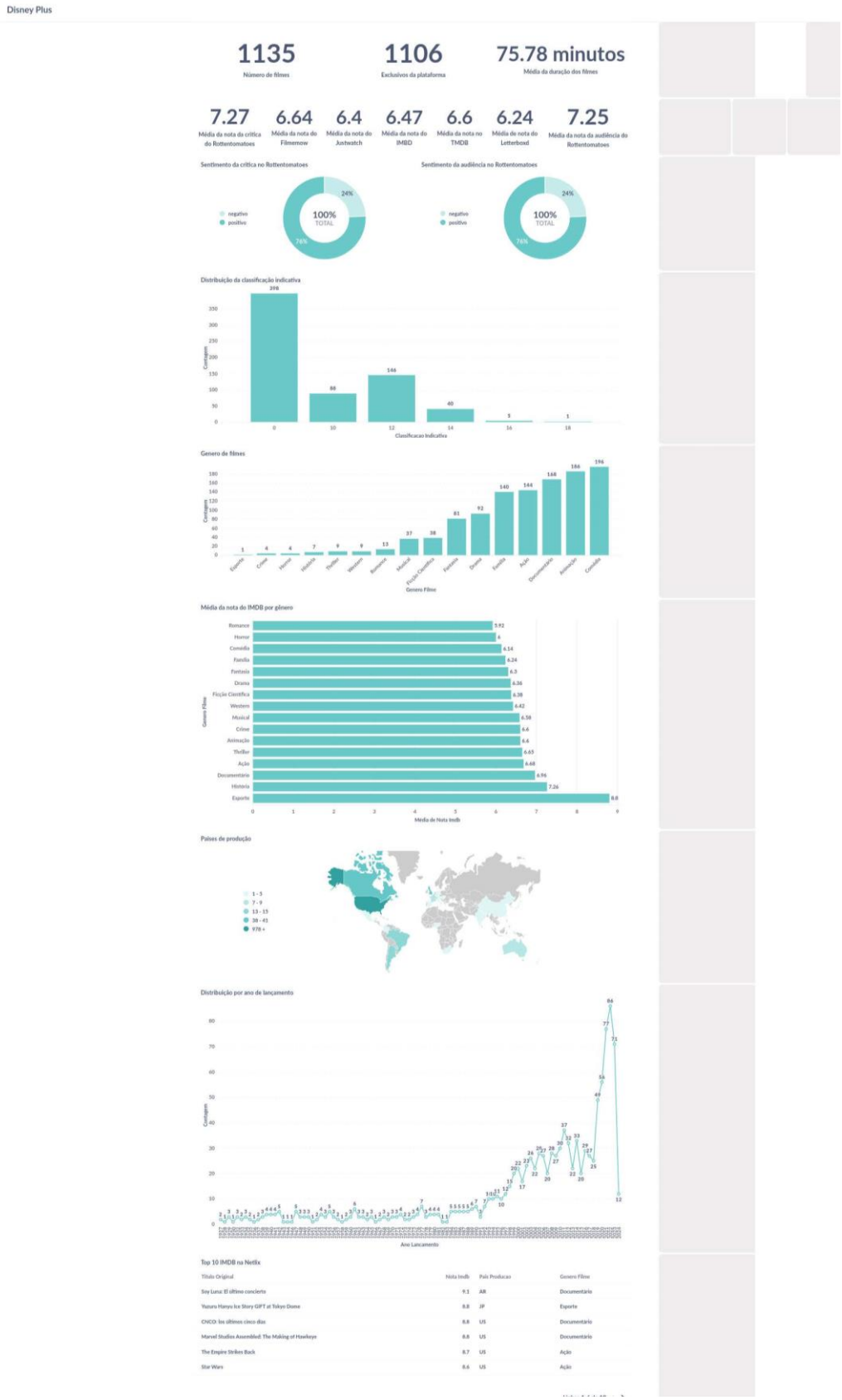


Figura 42: Gráfico Disney Plus

9.4. Gráfico Globoplay

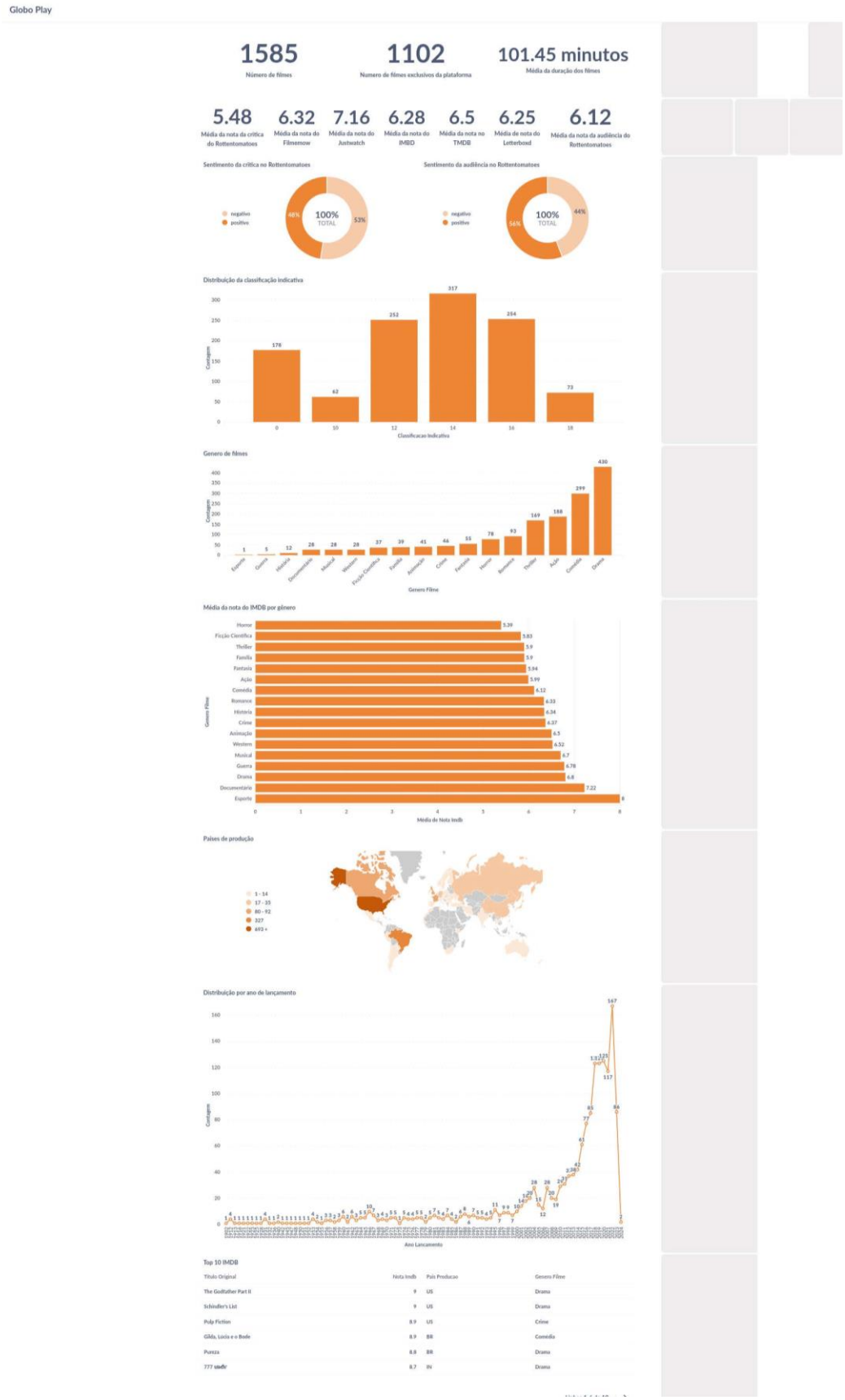


Figura 43: Gráfico Globoplay

9.5. Gráfico HBO Max

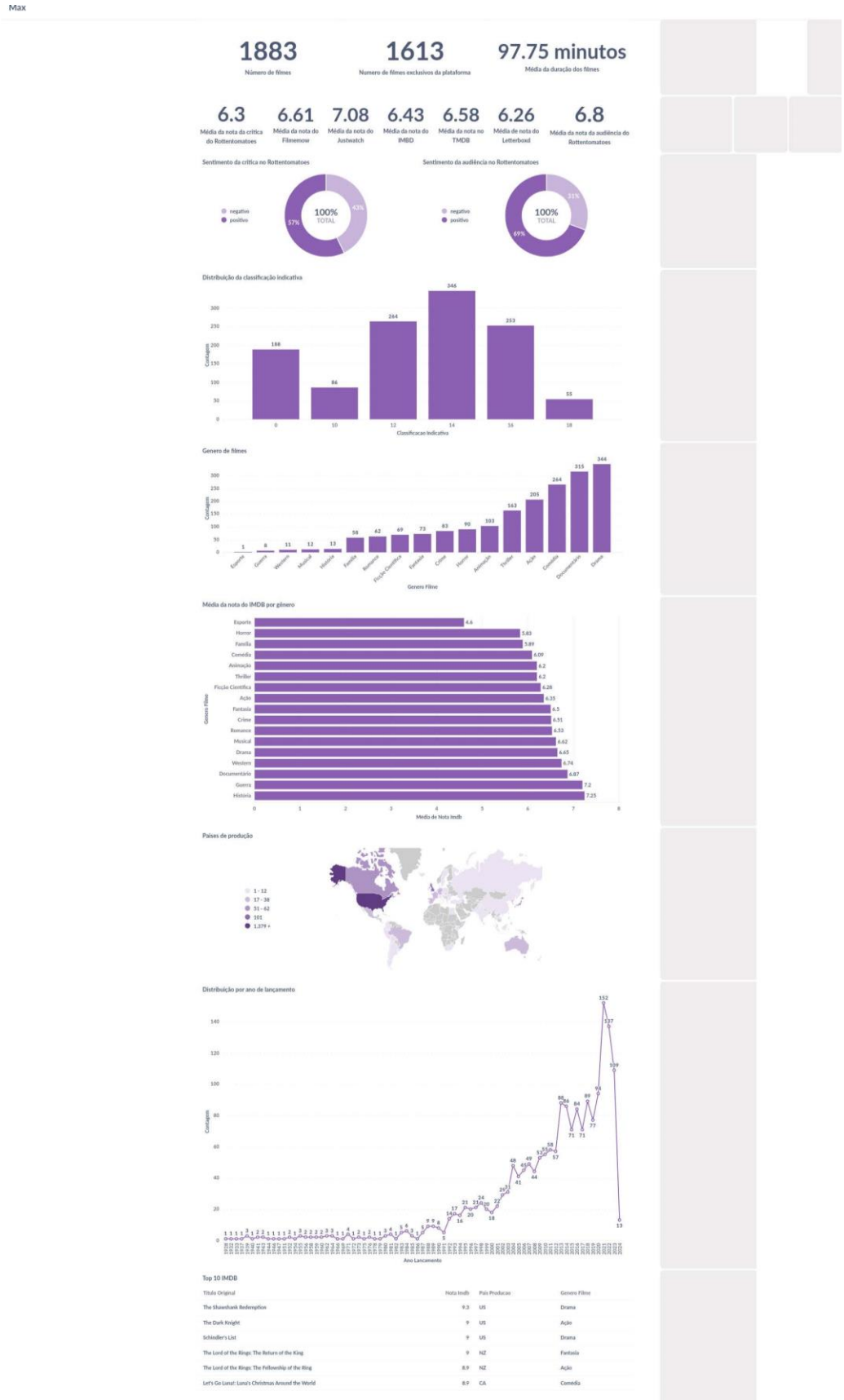


Figura 44: Gráfico HBO Max

9.6. Gráfico Netflix

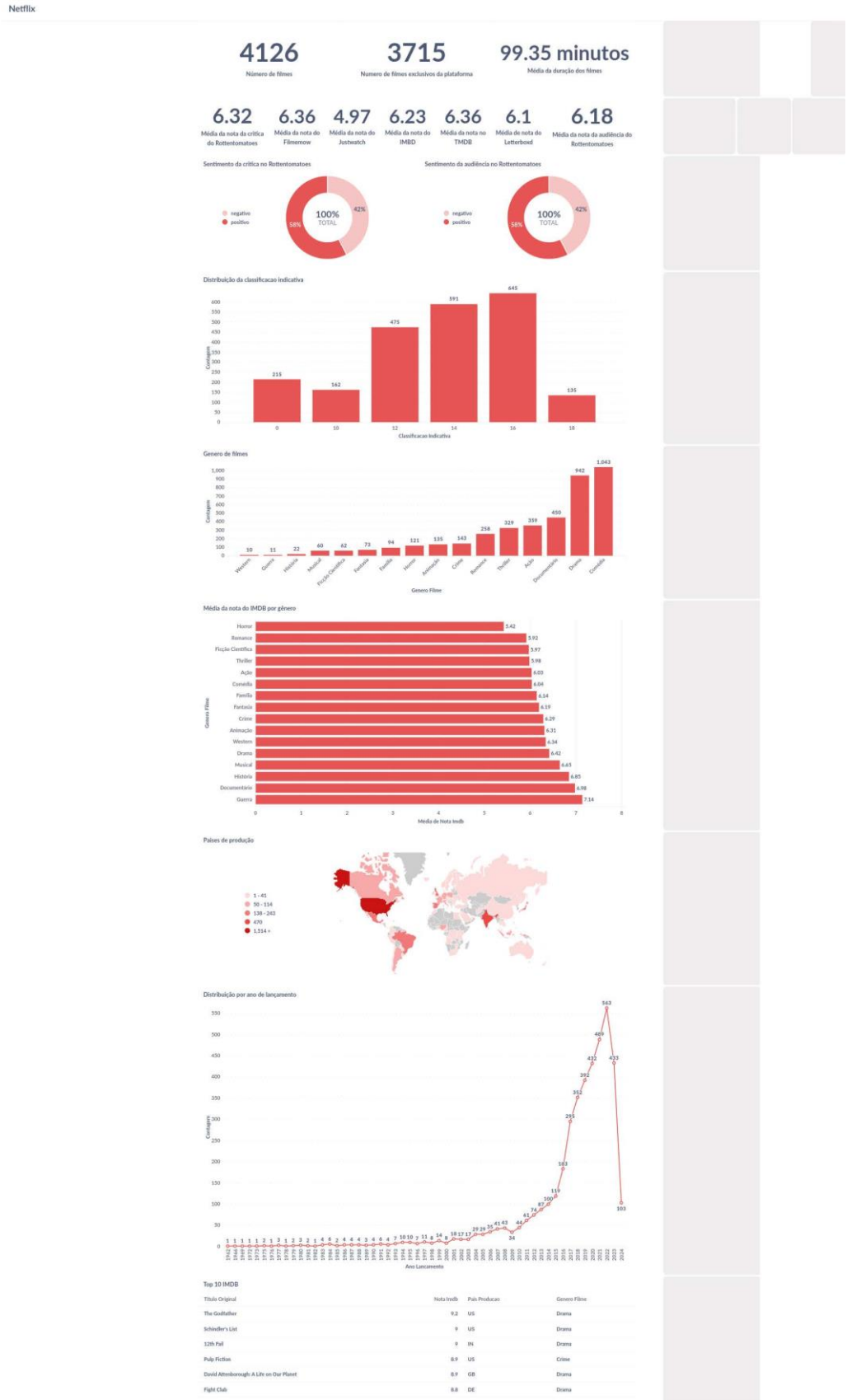


Figura 45: Gráfico Netflix

9.7. Gráfico Paramount Plus

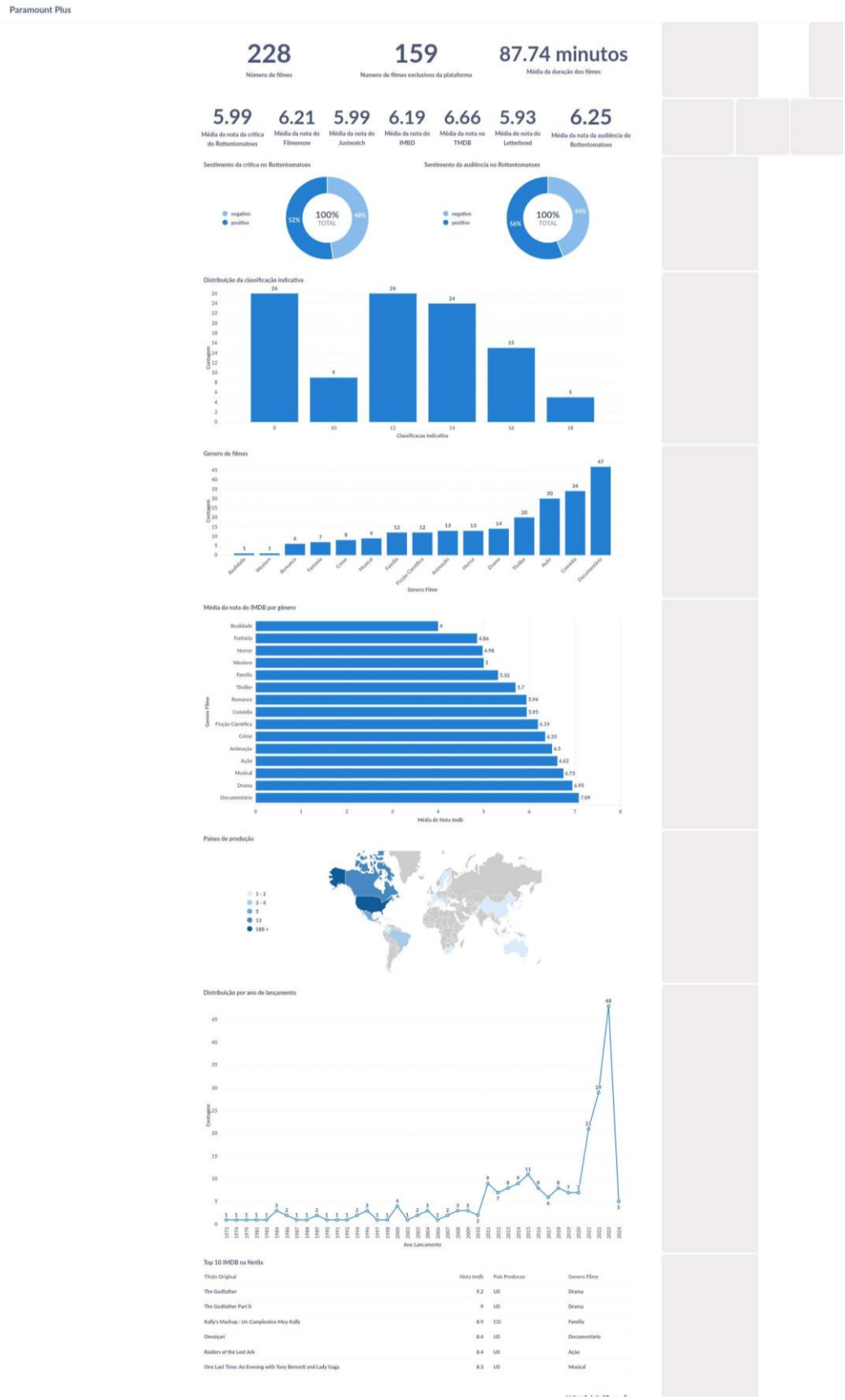


Figura 46: Gráfico Paramount Plus

9.8. Gráfico Star Plus

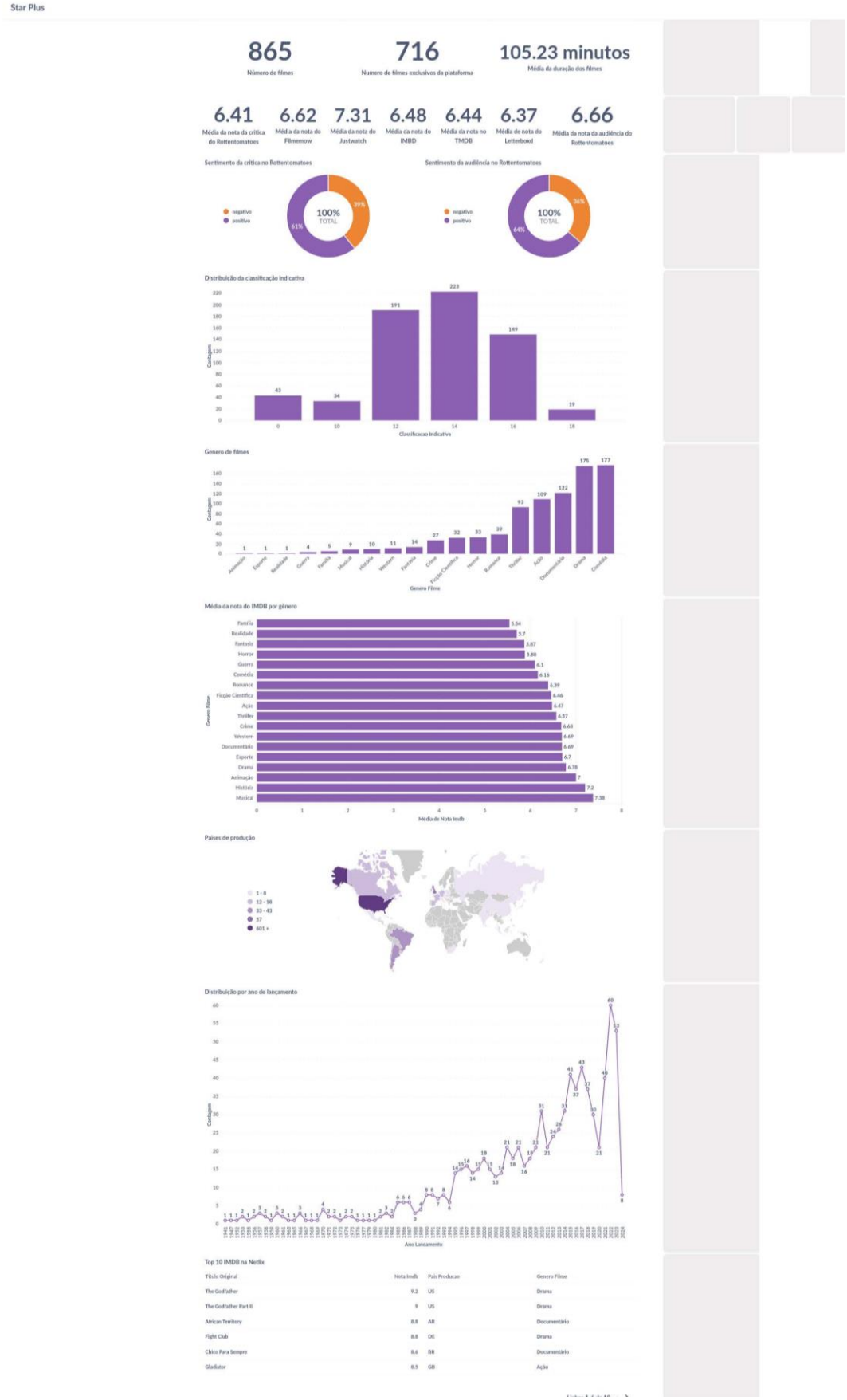


Figura 47: Gráfico Star Plus