

MVP: Engenharia de Dados

Created	Jul 01, 2023	Status	Finished
Last Updated	Set 30, 2023	Author	Maria Vitória Barbosa Valladares



Objetivo

Realizar uma análise abrangente de serviços de streaming, criando um pipeline de dados em uma plataforma de nuvem.

Detalhamento do objetivo:

1. Coleta de Dados:

- Recolher informações de consumo do usuário com base na solicitação de dados do assinante da Netflix e Spotify, o que inclui histórico de consumo, frequência de uso, buscas, e demais dados relevantes;
- Garantir que a coleta de dados esteja em conformidade com as regulamentações de privacidade e proteção de dados aplicáveis.

2. Pipeline de Dados com Google Cloud Platform

- Construir um pipeline completo de extração, transformação e carga dos dados escolhidos;
- Utilizar os serviços da GCP: Cloud Storage; Dataflow; Google BigQuery; Dataplex e Looker Studio.

3. Modelagem Star-Schema

- Implementar um modelo de dados Star-Schema para o Data Warehouse;
- Catalogar todos os dados detalhadamente, determinando seu domínio, categorias e valores máximos e mínimos.

4. Análise de Conteúdo:

- Realizar uma análise quantitativa e qualitativa das bases a fim de garantir a qualidade dos dados;
- Contruir um relatório que contenha as principais informações e insights acerca do perfil do usuário, frequência de uso, preferências e impactos que conteúdos geram em outras plataformas.

Coleta de Dados

Para a obtenção dos dados presentes neste projeto utilizou-se do artigo 15 do Regulamento Geral sobre a Proteção de Dados (RGPD), que garante direitos de informação e de acesso a dados pessoais. Dessa forma, todos os dados utilizados são provenientes do consumo da própria autora do projeto que consente a liberação dessas informações.

Os serviços de streaming escolhidos para este MVP foram o Spotify e Netflix, escolha justificada por ser as ferramentas mais utilizadas pela usuária. Para ambos os serviços a solicitação de uma cópia dos dados pessoais foram feitos direto na plataforma, utilizando a opção "Baixe seus dados", disponível em configurações, após a solicitação os serviços têm até 30 dias para enviar seus dados por e-mail, estes foram enviados no formato CSV e JSON, possuindo as principais informações listadas abaixo.

Spotify

Tipo de dados	O que está incluído
Playlist	<p>Um resumo das playlists criadas ou salvas e todas as músicas salvas, incluindo:</p> <ul style="list-style-type: none">• Nome da playlist.• Data da última modificação da playlist.• Nomes das músicas na playlist.• Nomes dos artistas de cada música.• Nomes de álbuns ou episódios (no caso de podcasts).• Nomes de faixas locais, caso o usuário tenha feito upload de arquivos de áudio armazenados localmente para ouvir no Spotify.• Descrições adicionadas à playlist pelo usuário.• Número de seguidores da playlist.
Histórico de streaming (áudio, vídeo e podcasts)	<p>Lista de itens (ou seja, músicas, vídeos e podcasts) reproduzidos no último ano, incluindo:</p> <ul style="list-style-type: none">• Data e hora, no formato UTC (Tempo Universal Coordenado), do fim do último streaming.• Nome do “criador” de cada streaming (por exemplo, o nome do artista no caso de uma música).• Nomes dos itens reproduzidos (por exemplo, nome da música ou do vídeo).• “msPlayed” - mostra por quantos milésimos de segundos uma faixa foi reproduzida.

Netflix

Tipo de dados	O que está incluído
Um resumo do que foi assistido	
ViewingActivity	<ul style="list-style-type: none">• O nome do perfil usado para assistir.• A data e a hora (UTC) do início da exibição.• A duração da sessão (observação: esse é o tempo total que o título foi assistido, independentemente do aparelho, e não necessariamente de forma consecutiva).• A série ou o filme assistido.• Vídeos que não sejam uma série ou um filme, como trailers ou montagens.
Histórico de buscas	
SearchHistory	<ul style="list-style-type: none">• O nome do perfil usado para solicitar a busca.• O país (com base no endereço IP) de onde a busca foi solicitada.• O tipo de aparelho utilizado para acessar a conta.• A consulta digitada no campo de busca.• A série ou filme resultante da consulta digitada.• As ações realizadas em resposta às buscas (por exemplo, se uma sinopse foi visualizada ou se foi clicado o botão play de uma série ou se uma série ou um filme identificado foi adicionado à “My List” (Minha lista)
Apresenta detalhes das ações executadas durante a navegação	
Clickstream	<ul style="list-style-type: none">• Nome do perfil associado ao clickstream• Tipo do aparelho utilizado no acesso• URL da página acessada na Netflix• Data e hora (UTC) do acesso

A fim de relacionar a base de músicas com a de séries e/ou filmes utilizou-se o dataset Soundtracks of Top 250 IMDb Movies and TV Series disponível no [kaggle](#), o mesmo contém trilhas sonoras usadas nos 250 principais filmes e séries de TV.

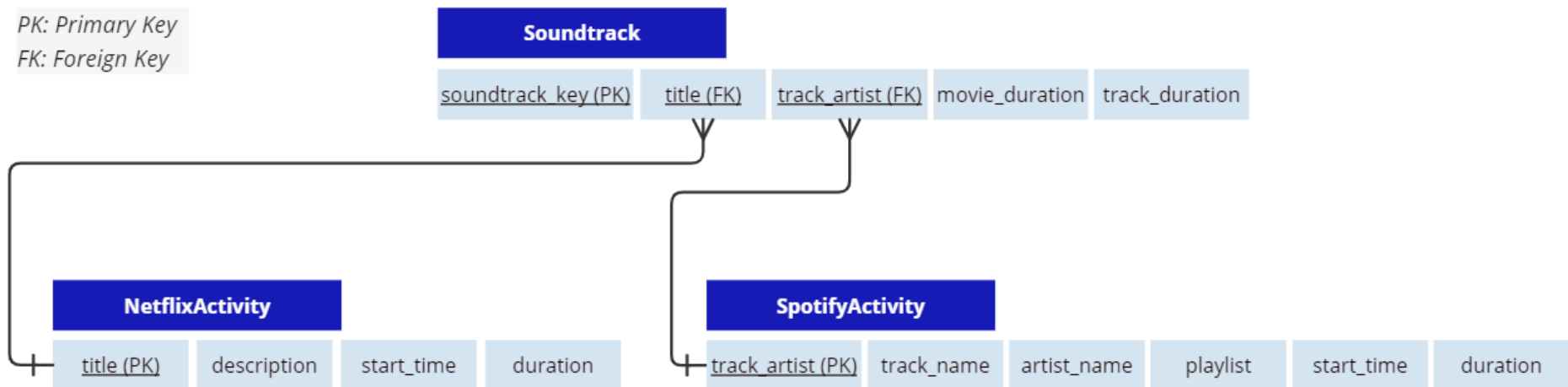
Kaggle

Tipo de dados	O que está incluído
Um resumo do que foi assistido	
Soundtrack	<ul style="list-style-type: none">• Nome do filme ou série de TV• Ano de lançamento• Nome da música que o título foi assistido• Nome do artista que cantou a música

Modelagem

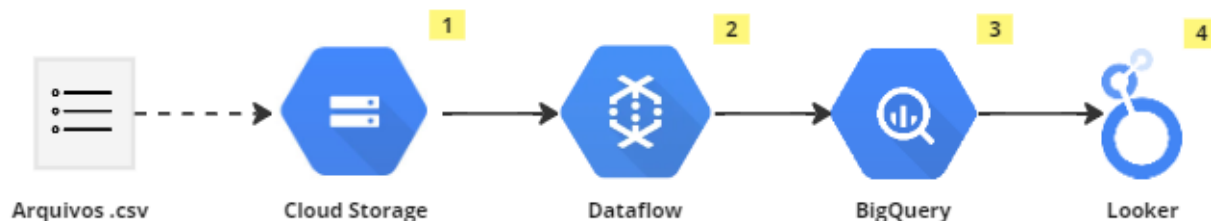
Esquema relacional

Mediante as fontes de dados obtidas, montou-se três tabelas, sendo seus relacionamentos expostos a seguir. A tabela **NetflixActivity** trata-se do arquivo ViewingActivity, enviado pelo streaming Netflix, sendo sua PK o título da série/filme consumido. Já a tabela **SpotifyActivity** refere-se a base Playlist enviada pelo streaming Spotify, optou-se por criar uma coluna composta pela junção track_name e artist_name, isto é música-artista, para ser sua PK. Por fim, a tabela **Soundtrack** é responsável pela relação entre filmes e músicas, para sua PK optou-se pela criação de uma Surrogate Key, soundtrack_key.



Cloud Services

Para esse projeto utilizou-se os serviços de tecnologia de nuvem Google Cloud Platform (GCP).



1. Criação e armazenamento de arquivos .csv em um bucket do Cloud Storage;
2. Orquestração de pipelines de dados, lendo os arquivos no Storage, tratando-os e escrevendo-os no BigQuery;
3. Armazenamento dos dados tratados e padronizados, data warehouse do GCP;
4. Consumo final dos dados, ambiente BI conectado ao BigQuery.

Conceder um papel do IAM

O primeiro passo para utilizar as ferramentas do GCP é conceder um papel no Google Cloud Identity and Access Management (IAM), que consiste em um processo importante para gerenciar as permissões de acesso da plataforma, para isso é necessário:

1. Acessar o Console GCP e navegar até a seção "IAM e administração" no painel.
2. Selecione o projeto no qual deseja conceder um papel e clique em "IAM" para acessar a página de gerenciamento de identidade e acesso.
3. Adicionar um membro ao projeto que você deseja conceder permissões.

A imagem mostra a interface do console de permissões do IAM do Google Cloud para o projeto "mvp-puc".

Permissões do projeto "mvp-puc"
Essas permissões afetam este projeto e todos os recursos dele. [Saiba mais](#)

Visualizar por Principais | Visualizar por Papéis

+ PERMITIR ACESSO | - REMOVER ACESSO

Filtro: Insira o nome ou o valor da propriedade

Papel/Principal	Nome	Herança
<input type="checkbox"/> Editor (1)		
<input type="checkbox"/> Proprietário (1)		
<input checked="" type="checkbox"/> vivalladarez@gmail.com	Vitoria Valladares	

Adicionar participantes
Os principais são usuários, grupos, domínios ou contas de serviço. [Saiba mais sobre os principais no IAM](#)

Novas principais *
[vivalladarez@gmail.com](#)

Atribuir papéis
Os papéis são compostos por conjuntos de permissões e determinam o que o principal pode fazer com o este recurso. [Saiba mais](#)

Papel *
Proprietário

Condição do IAM (opcional) [?](#)
[+ ADICIONAR CONDIÇÃO DO IAM](#)

Acesso total à maioria dos recursos do Google Cloud. Veja a lista de permissões incluídas.

[+ ADICIONAR OUTRO PAPEL](#)

[SALVAR](#) [CANCELAR](#)

Criar um bucket no Data Storage

Para armazenar os datasets obtidos, utilizou-se o Cloud Storage, o qual permite armazenar e recuperar dados de forma segura e escalável. Para utilizar a ferramenta seguiu-se os seguintes passos:

1. Acesso a página do Storage para iniciar a criação de um bucket
2. Nomeação de um nome único para o bucket e a localização de armazenamento dos dados. Esse passo afeta a latência e os custos de transferência de dados.
3. Configuração das opções de controle de acesso, permissões e opções de armazenamento.

← Criar um bucket

Nomeie seu bucket

Escolha um nome definitivo exclusivo. [Diretrizes de nomenclatura](#)

mvp-streamings

Dica: não inclua informações confidenciais

▼ MARCADORES (OPCIONAL)

CONTINUAR

Importante

Preços do local

As taxas de armazenamento variam dependendo da classe de armazenamento dos dados e da localização dos buckets. [Detalhes do preço](#)

Configuração atual: Region / Standard

Item	Custo
us-east1 (Carolina do Sul)	\$0.020 por GB/mês

ESTIMAR SEU CUSTO MENSAL

Escolha onde armazenar seus dados

Essa escolha define a posição geográfica dos dados e afeta o custo, o desempenho e a disponibilidade. Ela não pode ser alterada. [Saiba mais](#)

Tipo de local

- ☐ Multi-region
Disponibilidade mais alta entre áreas maiores
- ☐ Dual-region
Alta disponibilidade e baixa latência em 2 regiões
- ☒ Region
Latência mais baixa em uma única região

us-east1 (Carolina do Sul)

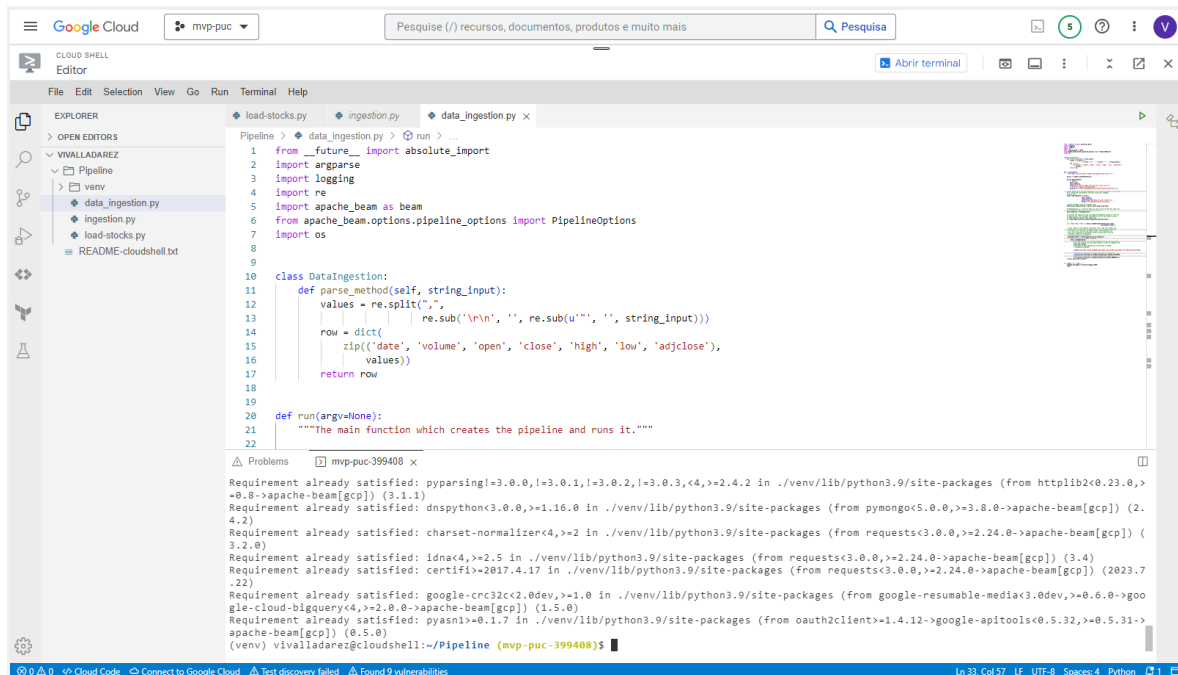
Acesso ao Cloud Shell

Todas as instalações de dependências, criação de projeto e Carga (configuração de fonte e destino dos dados) serão executadas via Cloud Shell, tais passos estão detalhados no README no github do projeto.

1. Configuração do ambiente virtual: a fim de não haver conflitos com versões específicas de pacotes cria-se um ambiente virtual e nele instala-se o apache-beam[gcp]:

```
python3 -m pip install --user virtualenv
virtualenv -p python3 venv
source venv/bin/activate
pip install 'apache-beam[gcp]'
```

2. Script python: após a preparação do ambiente executa-se o código `data_ingestion.py`, este cria um fluxo de trabalho que é executado criando um job no Dataflow.



```
1 from __future__ import absolute_import
2 import argparse
3 import logging
4 import re
5 import apache_beam as beam
6 from apache_beam.options.pipeline_options import PipelineOptions
7 import os
8
9
10 class DataIngestion:
11     def parse_method(self, string_input):
12         values = re.split(",",
13                             re.sub("\r\n", "", re.sub(u" ", "", string_input)))
14         row = dict(
15             zip(('date', 'volume', 'open', 'close', 'high', 'low', 'adjclose'),
16               values))
17         return row
18
19
20 def run(argv=None):
21     """The main function which creates the pipeline and runs it."""
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```

Criar um Conjunto de Dados BigQuery

Para a implementação da carga de dados é necessário que haja um destino para os mesmos, assim é necessário criar um conjunto de dados no BigQuery para armazenar e consultar dados de maneira eficiente, para isso basta:

1. Acessar o Console do GCP e navegar até a página do BigQuery e no painel escolher "Criar conjunto de dados", definindo um nome único.
2. Além disso, é necessário configurar as opções de controle de acesso e permissões especificando também as regiões.

Criar conjunto de dados

ID do projeto
mvp-puc-399408 [MUDAR](#)

Código do conjunto de dados *

Letras, números e sublinhados são permitidos

Tipo de local ?
☒ Região
Especifique uma região para colocar seus conjuntos de dados com outros serviços do Google Cloud.
☐ Multirregional
Permita que o BigQuery selecione uma região dentro de um grupo para atingir limites de cota mais altos.

Região *

Expiração da tabela padrão
☐ Ativar expiração da tabela ?
 Days

Opções avançadas

[CRIAR CONJUNTO DE DADOS](#) [CANCELAR](#)

Executando o Pipeline e Job Dataflow

Antes de executar o script deve-se atentar à habilitação dos serviços do Dataflow. O pipeline executa tarefas de ETL (Extract Transform Load), lendo o arquivo do bucket no Google Storage, transformando os dados para o formato aceito pelo BigQuery e os carregando em uma tabela do conjunto de dados criado. Para iniciar o script basta executar no shell:

```
python data-ingestion.py \
--project=mvp-puc-399408 \
--runner=DataflowRunner \
--staging_location=gs://mvp-streamings/staging \
--temp_location gs://mvp-streamings/temp \
--input 'gs://mvp-streamings/netflix-spotify/*.csv' \
--region=us-east1
```

Problems mvp-puc-399408

```
tempFiles: []
type: TypeValueValuesEnum(JOB_TYPE_BATCH, 1)>
INFO:apache_beam.runners.dataflow.internal.apiclient:Created job with id:
[2023-09-24_08_54_35-12079269339544327129]
INFO:apache_beam.runners.dataflow.internal.apiclient:Submitted job: 2023-0
9-24_08_54_35-12079269339544327129
INFO:apache_beam.runners.dataflow.internal.apiclient:To access the Dataflo
w monitoring console, please navigate to https://console.cloud.google.com/
dataflow/jobs/us-east1/2023-09-24_08_54_35-12079269339544327129?project=mv
p-puc-399408
```

Para acompanhar os resultados da execução é necessário acessar o serviço do Cloud Dataflow e o Job em execução:

GRÁFICO DE JOB

DETALHES DA EXECUÇÃO

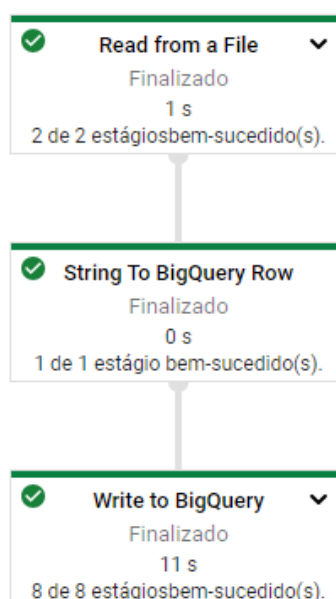
MÉTRICAS DO JOB

CUSTO

RECOMENDAÇÕES (1)

Visualização de etapas do job
Visualização de gráfico ▼

LIMPAR SELEÇÃO



1. Read from a File

O pipeline inicia usando os argumentos fornecidos, que incluem informações como o ID do projeto, o local dos dados (bucket no Google Storage) e a localização para o armazenamento dos arquivos temporários do Dataflow.

2. String to BigQuery Row

Após a leitura dos dados armazenados no Google Storage, este estágio transforma uma linha de arquivo CSV para um objeto de dicionário consumível pelo BigQuery. Além disso esta etapa realiza o tratamento do dataset

- Substitui todas as aspas duplas (") por uma string vazia
- Substitui todas as ocorrências da sequência de escape `\n` por uma string vazia
- divide a string em uma lista de valores com base nas vírgulas como separadores, após a remoção das aspas duplas e das quebras de linha.

3. Write to BigQuery

Por fim, esta etapa cria uma tabela no BigQuery se ela ainda não existir ou exclui todos os dados se ela já existir antes de gravar os dados já tratados.

Resultado BigQuery e Análise exploratória

O resultado final do pipeline pode ser verificado acessando a tabela criada no BigQuery, presente no conjunto de dados streamings, criado anteriormente.

spotifyactivity				spotifyactivity			
ESQUEMA				EXECUTAR			
1 SELECT * FROM `mvp-puc-399408.streamings.spotifyactivity` LIMIT 1000							
Resultados da consulta							
INFORMAÇÕES DO JOB				RESULTADOS			
track_artist				track_name			
artist_name				artist_name			
1 cartapradeus-gdm				carta pra deus			
2 sundaybloody-sunday-remaster...				sunday bloody sunday - remast...			
3 seeunãotecantar-fbc				se eu não te cantar			
4 sômeligar-bk				só me ligar			
5 lugarnamesa-bk				lugar na mesa			
6 ifyouwantlove-nf				if you want love			
7 myband-d12				my band			
8 hope-nf				hope			
9 sundavbloody-sundav-remaster				sundav bloody sundav - remast			

Obs: Este processo foi realizado para as três tabelas: Soundtrack, NetflixActivity e SpotifyActivity.

Finalizado a orquestração dos dados, analisou-se a qualidade dos dados armazenados no BigQuery para garantir a confiabilidade e desempenho na etapa de consumo. Dessa forma, verificou-se a existência de valores nulos, máximos e mínimos. As consultas em SQL a seguir exemplificam tais análises.

```
-- Verificar valores nulos
SELECT COUNT(*) as artist_name_null
FROM `mvp-puc-399408.streamings.spotifyactivity`
WHERE artist_name IS NULL;
```

```
-- Identificar valores máximos e mínimos
SELECT MIN(duration) AS ValorMinimo, MAX(duration) AS ValorMaximo
FROM `mvp-puc-399408.streamings.spotifyactivity`;
```

```
-- Verificar a contagem de valores distintos em uma coluna
SELECT track_name, COUNT(*) as
FROM `mvp-puc-399408.streamings.spotifyactivity`
GROUP BY track_name;
```

```
-- Verificar a contagem de valores distintos em uma coluna
SELECT track_artist, COUNT(*)
FROM `mvp-puc-399408.streamings.spotifyactivity`
GROUP BY track_artist
HAVING COUNT(*) > 1;
```

Todas as consultas retornam com resultados aceitáveis para as 3 bases de dados, isto é, sem valores nulos para as principais colunas (PK) e (FK), sem duplicidade e sem erros.

Data Catalog com Google Dataplex

Tendo todos os dados disponíveis e validados no BigQuery, iniciou-se o processo de construção de um Data Catalog, esse desempenha um papel crucial em qualquer organização no controle e uso eficiente de dados. Assim sendo, utilizou o Google Dataplex, que possibilita catalogar e rastrear metadados de maneira automática e consistente, para isso executou-se os seguintes passos:

1. No BigQuery catalogar e adicionar todos os campos informativos das tabela
2. Ativação e configuração do serviço do Dataplex no GCP
3. Criação de um Glossário de dados e sincronizados com as bases do BigQuery
4. Determinação de Tags e Termos de Negócios

Estes procedimentos proporcionaram uma documentação relativa ao domínio, categorias e descrição de dados e que também podem ser vistas no próprio BigQuery.

netflixactivity

CONSULTA

COMPARTILHAR

COPIAR

SNAPSHOT

EXCLUIR

EXPORTAR

ESQUEMA

DETALHES

PREVIEW

LINHAGEM

PERFIL DE DADOS

QUALIDADE DOS DADOS

Filtro

Insira o nome ou o valor da propriedade

<input type="checkbox"/>	Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas	Descrição
<input type="checkbox"/>	title	STRING	NULLABLE					Título do filme ou série
<input type="checkbox"/>	description	STRING	NULLABLE					Descrição do título (Temporada, sequência do filme, etc)
<input type="checkbox"/>	start_time	DATE	NULLABLE					Primeira execução do título (YYYY-MM-DD)
<input type="checkbox"/>	duration	INTEGER	NULLABLE					Tempo de duração da título em cartaz em (ms)

Google Cloud

mvp-puc

data

Dataplex

Glossários

PRÉ-VISUALIZAÇÃO

Explorar

Pesquisa

Ítem catálogo

Modelos de tag

Grupos de entradas

Glossários

Ítem lakes

Gerenciar

Seguro

Processar

Filtro

Search

streamings

NetflixActivity

description

duration

start_time

title

Soundtrack

movie_duration

track_duration

SpotifyActivity

artist_name

playlist

track_artist

track_name

title

PÁGINA INICIAL

criar categoria

add term

title

mvp-puc-399408 > us-east1 > streamings > NetflixActivity

Term -Última modificação: 25 de set. de 2023 17:06:47

Administrador:

Título do filme ou série, do tipo string.

editar descrição

Entradas relacionadas

Filtro

Insira o nome ou o valor da propriedade

Nome da entrada

netflixactivity

As informações adicionadas no Data Catalog referem-se às tabelas NetflixActivity, SpotifyActivity e Soundtrack, buscou-se disponibilizar tanto o conceito descritivo das tabela e atributos como seus metadados.

Por fim, como o print de todo o catálogo iria gerar diversas imagens, o que aumentaria o tamanho do relatório, resumiu-se todas as informações cadastradas no Data Catalog na página a seguir.

Soundtrack

Coluna	Informações
title	Refere-se ao nome do filme/série reproduzido, pk da tabela netflixactivity <ul style="list-style-type: none">• VARCHAR• 400 caracteres
track_artist	Refere-se a pk da tabela spotifyactivity <ul style="list-style-type: none">• VARCHAR• 400 caracteres
movie_duration	Refere-se ao tempo de reprodução do filme/série <ul style="list-style-type: none">• Inteiro• milissegundos
track_duration	Refere-se ao tempo de reprodução da música <ul style="list-style-type: none">• Inteiro• milissegundos

NetflixActivity

Coluna	Informações
title	Refere-se ao nome do filme/série reproduzido, pk da tabela <ul style="list-style-type: none">• VARCHAR• 400 caracteres
description	Descrição do título do filme ou série reproduzida <ul style="list-style-type: none">• String• Valor máximo: 300 caracteres• Valor mínimo: 2 caracteres
start_time	Refere-se a data da reprodução do título <ul style="list-style-type: none">• Timestamp• yyyy-MM-dd HH:mm:ss
duration	Refere-se ao tempo de reprodução do filme/série <ul style="list-style-type: none">• Inteiro• milissegundos

SpotifyActivity

Coluna	Informações
track_artist	Refere-se a pk da tabela, sendo uma junção das colunas track_name e artist_name <ul style="list-style-type: none">• VARCHAR• 400 caracteres
track_name	Refere-se ao nome da faixa reproduzida <ul style="list-style-type: none">• VARCHAR• 400 caracteres
artist_name	Refere-se ao nome do artista da faixa <ul style="list-style-type: none">• VARCHAR• 400 caracteres
playlist	Playlist onde a música foi adicionada <ul style="list-style-type: none">• VARCHAR• 400 caracteres
start_time	Refere-se a data da reprodução da track_name <ul style="list-style-type: none">• Timestamp• yyyy-MM-dd HH:mm:ss
duration	Refere-se ao tempo de reprodução do filme/série <ul style="list-style-type: none">• Inteiro• milissegundos

Looker Studio

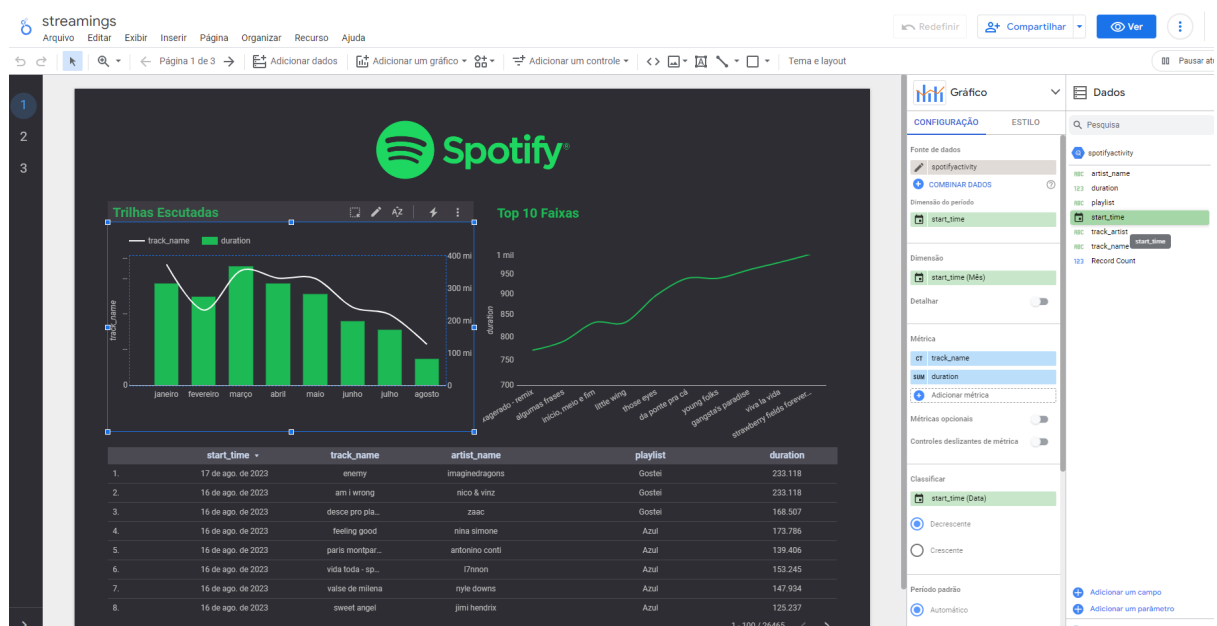
Tendo todos os dados disponíveis e catalogados no nosso data warehouse, utilizou-se o Looker para realizar o consumo, analisando todas as informações disponíveis no dataset e, assim, concluir o objetivo deste trabalho. Para isso, é necessário os seguintes passos:

1. No BigQuery utiliza-se a opção “Explorar com Looker Studio”
2. Conecta-se ao relatório novas fontes de dados em “Data Sources”
3. Cria-se medidas e visualizações

The screenshot shows the Looker Studio interface. On the left is the 'Explorer' panel with a search bar and a list of resources under 'mvp-puc-399408', including 'streamings' with sub-items 'netflixactivity', 'soundtrack', and 'spotifyactivity'. The main area shows a query editor with a query: `SELECT * FROM `mvp-puc-399408.streaming.spotifyactivity` LIMIT 1000`. Below the query is a 'Resultados da consulta' table with columns 'track_artist', 'track_name', and 'artist_name'. The table contains three rows of data.

Linha	track_artist	track_name	artist_name
1	blaaablaa--	blaa blas blaa	-
2	blaaablaa--	blaa blas blaa	-
3	blaaablaa--	blaa blas blaa	-

Escolheu-se construir a análise no formato relatório a fim de extraí-lo e anexá-lo a este documento, os resultados obtidos nessa construção serão explicitados na próxima sessão.



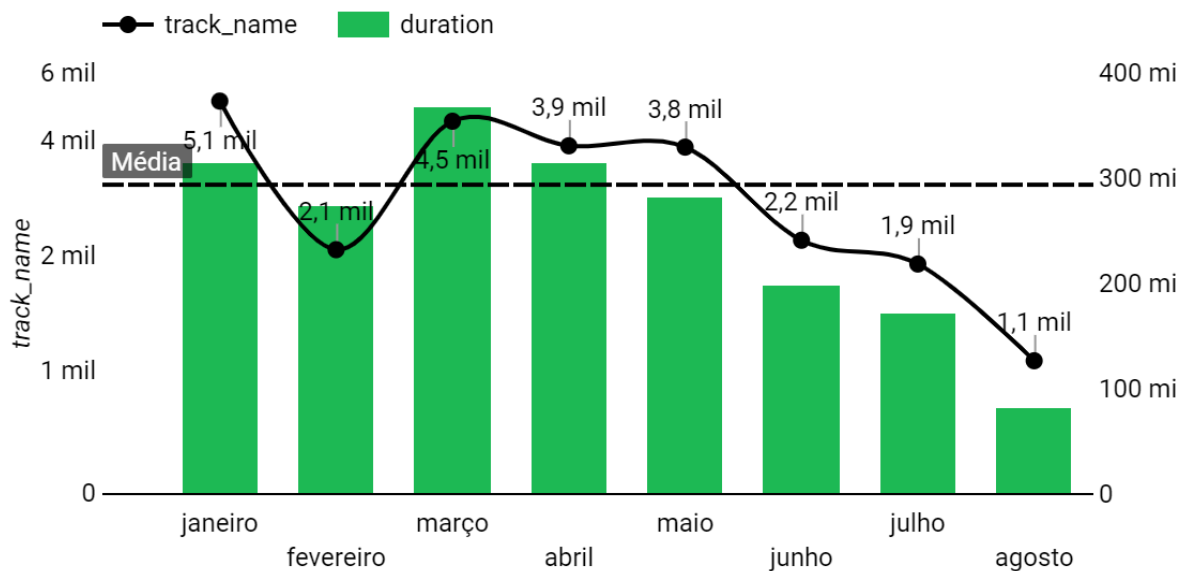
Resultados e Conclusão

Resultados

1. Spotify

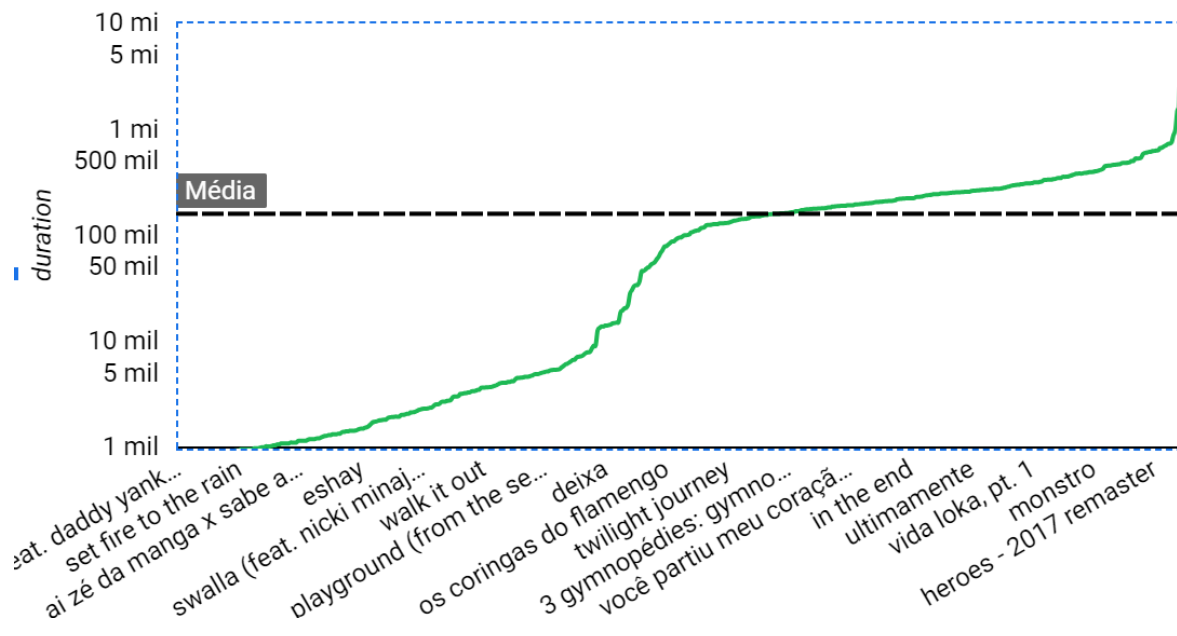
a) Tempo de consumo quantidade de tracks tocadas

Ao analisar o streaming Spotify para o ano de 2023 verificou-se a média de 3 mil músicas reproduzidas por mês, este número inclui reproduções de diferentes playlists e buscas dentro do aplicativo, Dentre os meses analisados destaca-se o mês de janeiro com o consumo de 5,1 mil tracks tocadas. Por outro lado, ao analisar o tempo de consumo, destaca-se o mês de março com 367.983.526 milissegundos de reprodução, isto é, 102 horas.



b) Top Tracks do mês

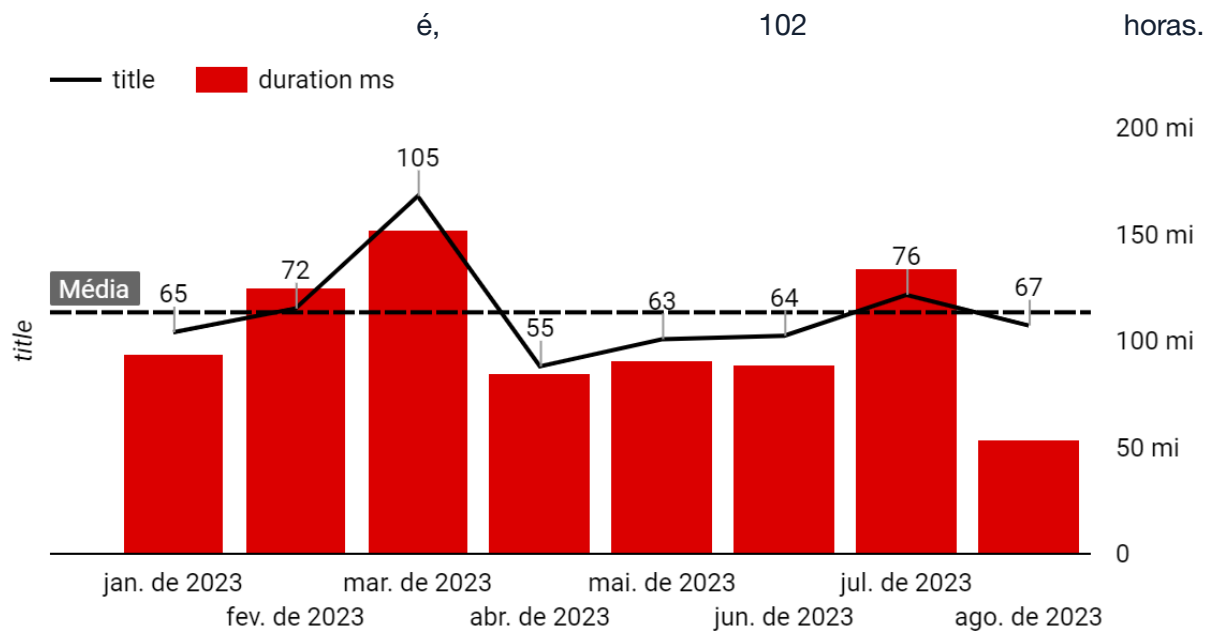
Por fim, dentre as mais tocadas no mês, Heroes, canção do David Bowie, lidera o Top Faixas, com 25 minutos de reprodução, de uma média global de 2,6 minutos.



2. Netflix

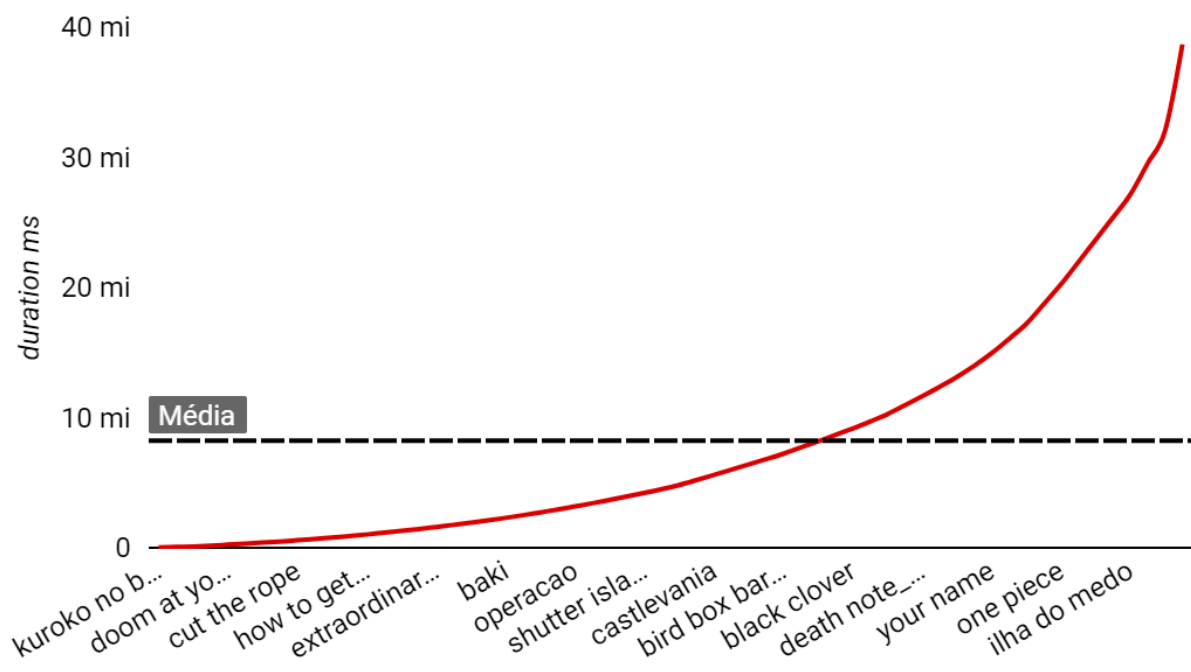
c) Tempo de consumo quantidade de títulos assistidos

Ao analisar o streaming Netflix para o ano de 2023 verificou-se a média de 70 títulos assistidos por mês, este número inclui diferentes episódios de uma séries, trailers e amostras rápidas de títulos em cartaz. Dentre os meses analisados destaca-se o mês de março com 105 títulos consumidos. Por outro lado, ao analisar o tempo de consumo, obtém-se uma média de 28 horas de consumo mensal.



d) Top titulos do mês

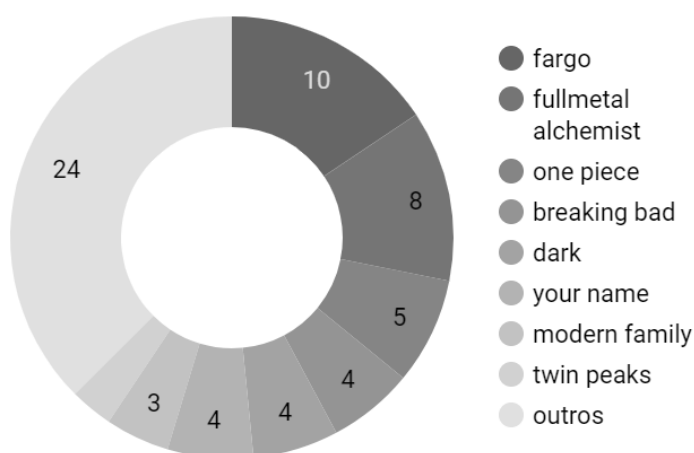
Por fim, dentre os títulos mais assistidos no mês, Ilha do medo, estrelando por Leonardo DiCaprio lidera com 7 horas de reprodução, de uma média global de 2,2 horas.



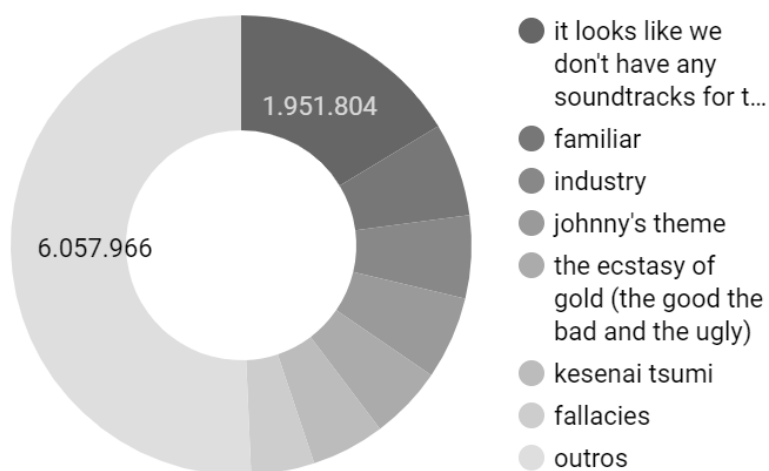
3. Interação entre plataformas

Como última etapa de análise, buscou-se identificar a influência entre plataformas, a partir do dataset Soundtrack, observou-se a relação entre títulos assistidos e faixas musicais executadas pertencentes a aquele conteúdo.

Assim sendo, observou-se a influência de títulos como fargo, dark e modern family na descoberta de novas músicas, o gráfico abaixo demonstra este resultado, tendo destaque para a série Fargo, que a partir do seu consumo influenciou a descoberta de 10 faixas no Spotify.



Tais faixas, geram um consumo no streaming de música de mais de 3 horas durante o ano, com hits como johnny's theme, industry e outros.



obs: Os relatórios finais gerados no Looker Studio estão anexados no final deste relatório.

Conclusão

Este trabalho teve como objetivo principal analisar dados de streamings, criando um pipeline de dados em uma plataforma de nuvem como meio de aprendizagem para o módulo de Engenharia de Dados da Pós-graduação de Data Science & Analytics da PUC-Rio.

Com base no exposto, inicialmente, optou-se pelo Google Cloud Platform como a ferramenta de nuvem a ser utilizada, o que se revelou uma escolha acertada devido à grande variedade de ferramentas oferecidas e à abundante documentação disponível para cada serviço. Vale destacar como ponto positivo o recurso Qwiklabs, um laboratório prático oferecido pela Google para implementações básicas das ferramentas em nuvem, o qual permitiu práticas iniciais na plataforma sem custos adicionais.

Posteriormente, a obtenção das bases utilizadas neste trabalho apresentaram uma dificuldade inicial devido ao tempo de respostas das plataformas de streaming, as quais demoraram cerca de 7 a 30 dias para envio dos dados. Além disso, é importante ressaltar que a base vínculo (filmes e trilha sonora) apresenta apenas 250 títulos, o que influencia na carência de informações que conectam os dados entre Netflix e Spotify.

Por outro lado, a implementação do pipeline de dados se mostrou eficiente, intuitiva e de baixo custo, gerando um aprendizado significativo de diversos serviços do Google Cloud como: Cloud Storage; Dataflow; Google BigQuery; Dataplex e Looker Studio.

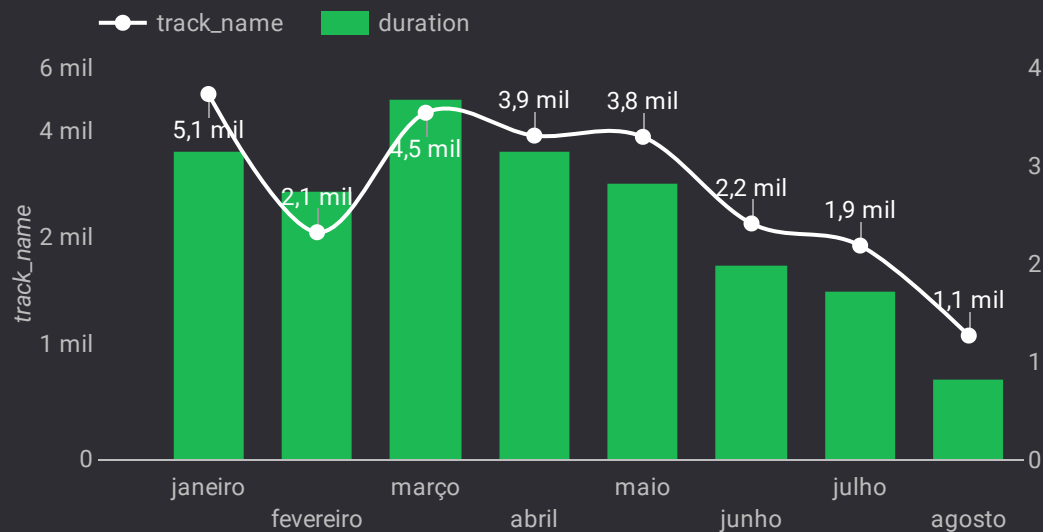
No consumo final, os insights obtidos responderam às principais perguntas do negócio: tempo de consumo, tipo de consumo e interação entre plataformas. Contudo, é importante ressaltar que algumas informações não foram incluídas no projeto devido ao prazo de entrega. Adições como gênero, horários de consumo, principais buscas e abertura de segmentação dos dados trariam mais riqueza à análise.

Mediante o exposto, pode-se afirmar que o objetivo do trabalho foi atingido. Em suma, como possíveis próximos passos do projeto podemos identificar:

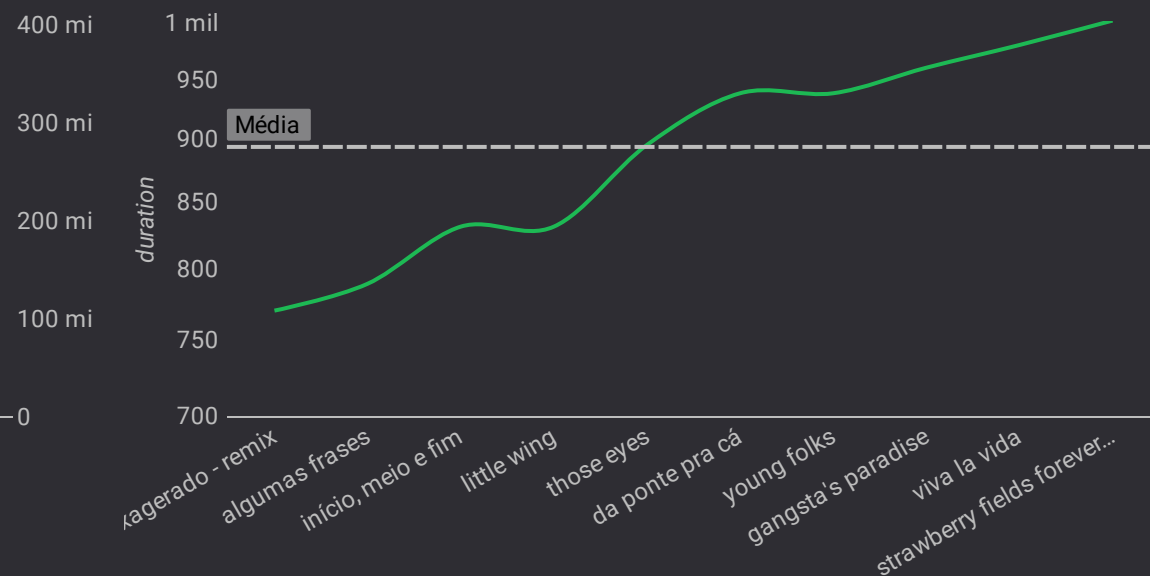
- O incremento de mais plataformas de streaming a base de dados
- Criar/Enriquecer um dataset de relacionamento entre as plataformas de streaming
- Categorizar o cliente de acordo com o tipo dos seus dados
- Obter o dado de maneira mais rápida e otimizada, este consiste no passo com maior dificuldade devido a necessidade de disponibilização por parte das empresas de streaming



Trilhas Escutadas



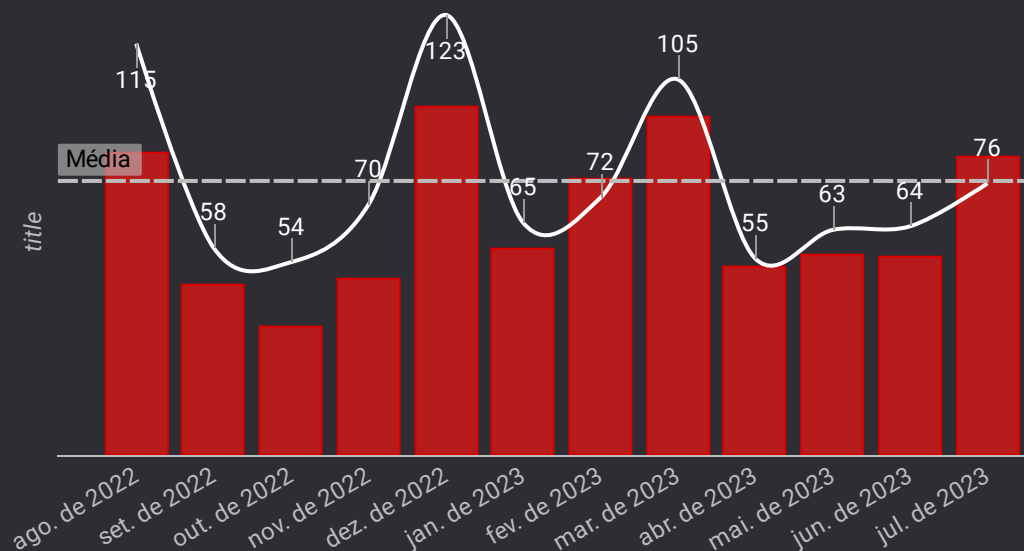
Top 10 Faixas



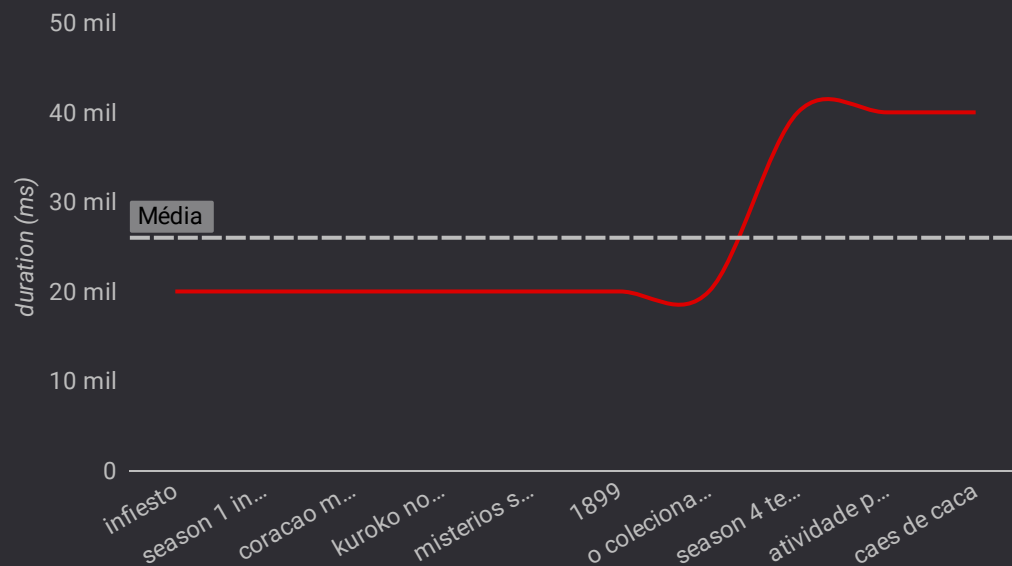
	start_time ▾	track_name	artist_name	playlist	duration
1.	17 de ago. de 2023	enemy	imaginedragons	Gostei	233.118
2.	16 de ago. de 2023	am i wrong	nico & vinz	Gostei	233.118
3.	16 de ago. de 2023	desce pro pla...	zaac	Gostei	168.507
4.	16 de ago. de 2023	feeling good	nina simone	Azul	173.786
5.	16 de ago. de 2023	paris montpa...	antonino conti	Azul	139.406
6.	16 de ago. de 2023	vida toda - sp...	l7nnon	Azul	153.245
7.	16 de ago. de 2023	valse de milena	nyle downs	Azul	147.934
8.	16 de ago. de 2023	sweet angel	jimi hendrix	Azul	125.237

NETFLIX

Títulos Assistidos



Top 10 Títulos



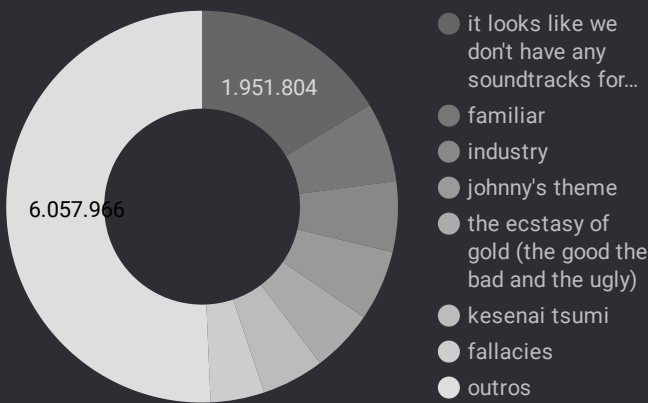
	start_time ▾	title	description	duration
1.	15 de ago. de 2023	temporada 3 (clipe)	Temporada 3 (Clipe): Sintonia	7000
2.	15 de ago. de 2023	temporada 1 (trailer)	Temporada 1 (Trailer): Record of Ragnarok	4000
3.	15 de ago. de 2023	temporada 2 (clipe)	Temporada 2 (Clipe): Baki Hanma	9000
4.	15 de ago. de 2023	temporada 1 (teaser)	Temporada 1 (Teaser): Next in Fashion	13000
5.	14 de ago. de 2023	temporada 1 (clipe)	Temporada 1 (Clipe): Que Chegue a Voce: Kimi ni Todoke	9000
6.	14 de ago. de 2023	death note	DeATH NOTE: Death Note: Renewal (episodio 26)	16000
7.	14 de ago. de 2023	miniserie (clipe)	Miniserie (Clipe): Um Conto de Fadas Perfeito	4000
8.	14 de ago. de 2023	temporada 2 (trailer)	Temporada 2 (Trailer): Vikings: Valhalla	4000



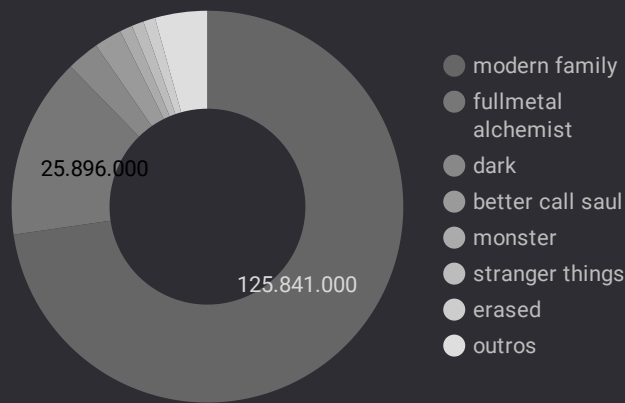
Quantidade de Faixas descobertas por Título



Consumo por Faixas Descobertas



Consumo dos Títulos



	title ▾	track_name	movie_duration	track_duration	movie_duration	track_duration
1.	your name	zen zen zense	48000	3553	48.000	3.553
2.	your name	yumetourou (dream lantern)	48000	353	48.000	353
3.	your name	sparkle	48000	35355	48.000	35.355
4.	your name	nandemonaiya	48000	230930	48.000	230.930
5.	vinland saga	it looks like we don't have a...	144000	244	144.000	244
6.	vikings	it looks like we don't have a...	925000	3344	925.000	3.344
7.	twin peaks	twin peaks theme	8000	12212	8.000	12.212
8.	twin peaks	it looks like we don't have a...	8000	243242	8.000	243.242