

Sprint: Engenharia de Dados

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



Objetivo

Objetivo: Realizar uma análise abrangente de serviços de streaming, utilizando dados de consumo, a fim de obter insights acerca do perfil do usuário, frequência de uso, preferências e impactos que conteúdos geram em outras plataformas.

Detalhamento do objetivo:

1. Coleta de Dados:

- a. 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;
- b. 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

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

3. Modelagem Star-Schema

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

4. Análise de Conteúdo:

- a. Realizar uma análise quantitativa e qualitativa das bases a fim de garantir a qualidade dos dados;
- b. Contruir um dashboard que contenha as principais informações e insights.

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 nas configurações, após a solicitação os serviços têm até 30 dias para enviar seus dados por e-mail, estes serão no formato CSV ou JSON e possuem 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.

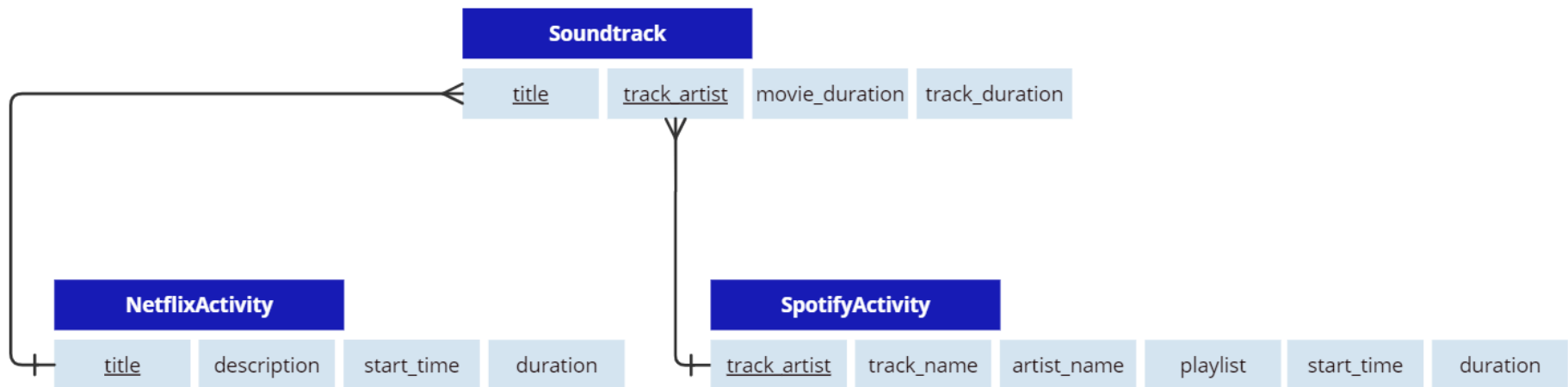
Kaggle

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

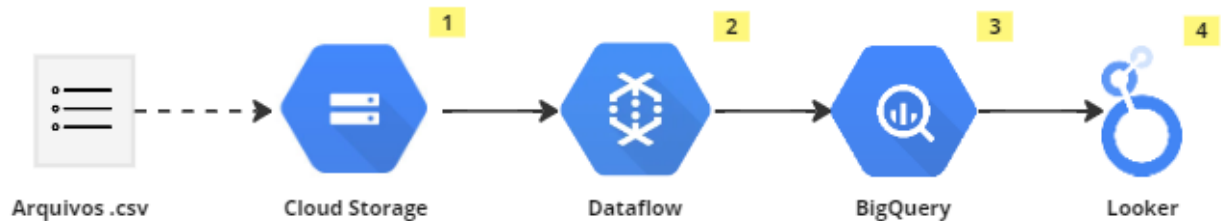
Esquema relacional

obs: Adicionar chave primária Soundtrack



Cloud Services

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



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 aos recursos da plataforma.

1. Acesse o Console do Google Cloud Platform (GCP) e navegue 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. Na página IAM, clique em "Adicionar" para adicionar um membro ao projeto e insira o endereço de e-mail, nome de usuário ou grupo que você deseja conceder permissões.

A imagem mostra a interface do console de permissões do IAM do Google Cloud Platform 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

O Cloud Storage é um serviço de armazenamento de objetos na nuvem que permite armazenar e recuperar dados de forma segura e escalável e oferece opções de armazenamento de baixo custo e fácil integração com outras soluções do Google Cloud.

1. Acesse o GCP e busque a página "Storage", clique em "Create bucket".
2. Escolha de um nome único o bucket e a localização de armazenamento dos dados do bucket. Isso afetará a latência e os custos de transferência de dados.
3. Configure as opções de controle de acesso, permissões e opções de armazenamento e clique em criar.

← Criar um bucket



Nomeie seu bucket


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

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)

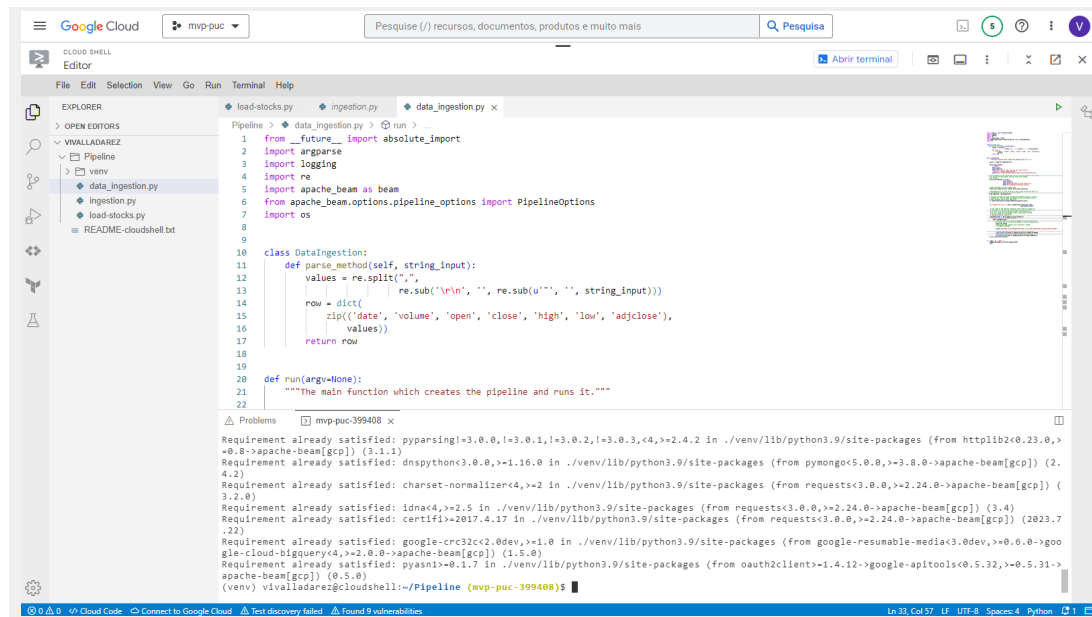
Acesso ao Cloud Shell

Todas as instalações de dependências, criação de projeto e Carga (configuração de fonte e destino dos dados) será executada via Cloud Shell, tais passos estão detalhados no README.md 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. Após a preparação do ambiente executa-se o código `data_ingestion.py`, este cria um fluxo de trabalho que é executado usando o *DataflowRunner*.



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 unico.
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 *

streamings

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 *

us-east1 (Carolina do Sul)

Expiração da tabela padrão

☐ Ativar expiração da tabela ?

Idade máxima padrão da tabela

Days

Opções avançadas

▼

CRIAR CONJUNTO DE DADOS

CANCELAR

Executando o Pipeline e Job Dataflow

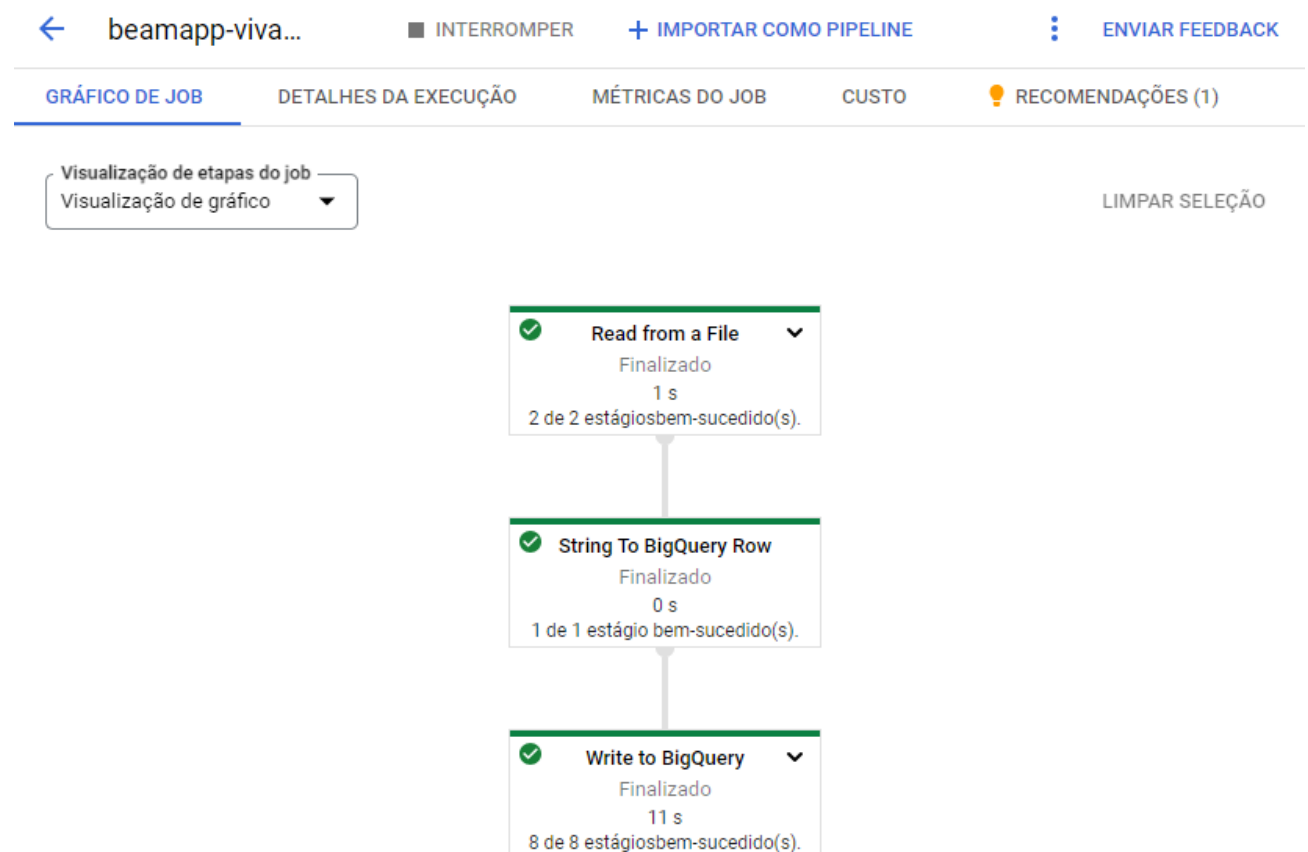
Antes de executar o script deve-se atentar à habilitação dos serviços do Dataflow. O pipeline executa tarefas de ETL (Extract Transform Load), lendo o arquivo do bucket no Google Storage, transformando os dados para o formato aceito pelo BigQuery e os carregando em uma tabela do conjunto de dados criado.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



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 \r\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			
JSON				DETALHES DA EXECUÇÃO			
GRÁFICO							
Nome do campo	Tipo	Modo	track_artist	track_name	artist_name		
track_artist	STRING	NULLABLE	cartapradeus-gdm	carta pra deus	gdm		
track_name	STRING	NULLABLE	sundaybloody-sunday-remaster...	sunday bloody sunday - remast...	u2		
artist_name	STRING	NULLABLE	seeunãotecantar-fbc	se eu não te cantar	fbc		
playlist	STRING	NULLABLE	sómeligar-bk	só me ligar	bk		
start_time	DATE	NULLABLE	lugarnamesa-bk	lugar na mesa	bk		
duration	INTEGER	NULLABLE	ifyouwantlove-nf	if you want love	nf		
			myband-d12	my band	d12		
			hope-nf	hope	nf		
			sundavbloody-sundav-remaster	sundav bloody sundav - remast	u2		

Data Catalog com Google Dataplex

Tendo todos os dados disponíveis 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, para isso utilizou o Google Dataplex, que possibilita catalogar e rastrear metadados de maneira automática e consistente, para isso executou-se os seguintes passos:

- 1. Ativação e configuração do serviço do Dataplex no GCP
- 2. Criação de um Glossário de dados e sincronizados com as bases do BigQuery
- 3. Determinação de Tags e Termos de Negócios

Google Cloud

mvp-puc

data

Dataplex

Glossários PRÉ-VISUALIZAÇÃO

title

PÁGINA INICIAL

criar categoria

add term

Explorar

Pesquisa

Criar catálogo

Modelos de tag

Grupos de entradas

Glossários

Criar 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

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

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

netflixactivity

CONSULTA

COMPARTILHAR

COPIAR

SNAPSHOT

EXCLUIR

EXPORTAR

ESQUEMA

DETALHES

PREVIEW

LINHAGEM

PERFIL DE DADOS

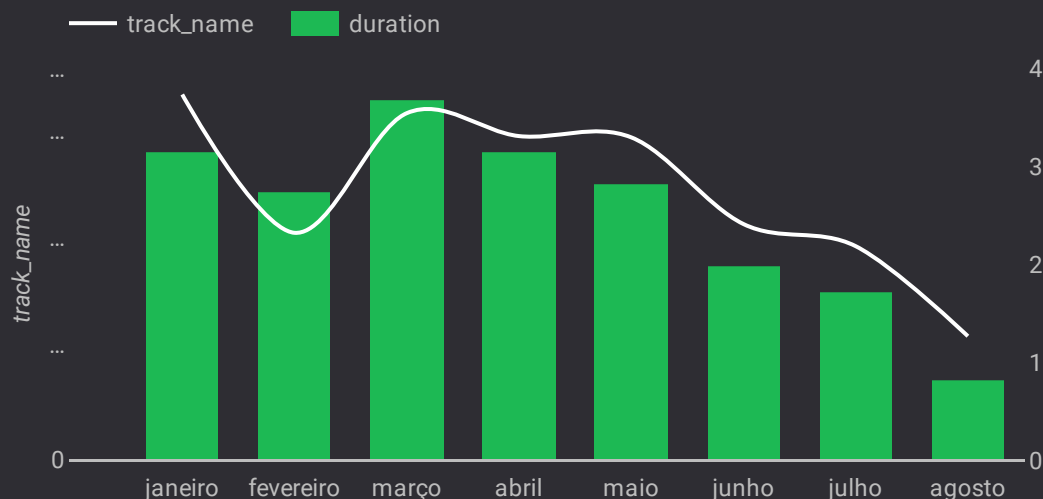
QUALIDADE DOS DADOS

Filtro Insira o nome ou o valor da propriedade

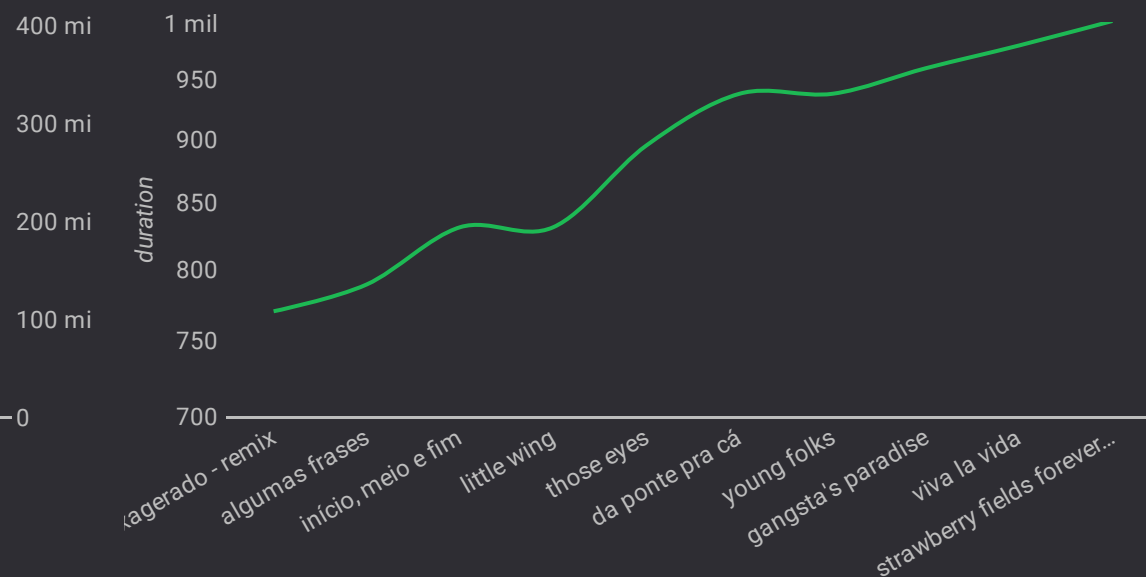
<input type="checkbox"/>	Nome do campo	Tipo	Modo	Chave	Compilação	Valor padrão	Tags de políticas	Descrição
<input type="checkbox"/>	title	STRING	NULLABLE					Título do filme ou série
<input type="checkbox"/>	description	STRING	NULLABLE					Descrição do título (Temporada, sequência do filme, etc)
<input 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)



Trilhas Escutadas



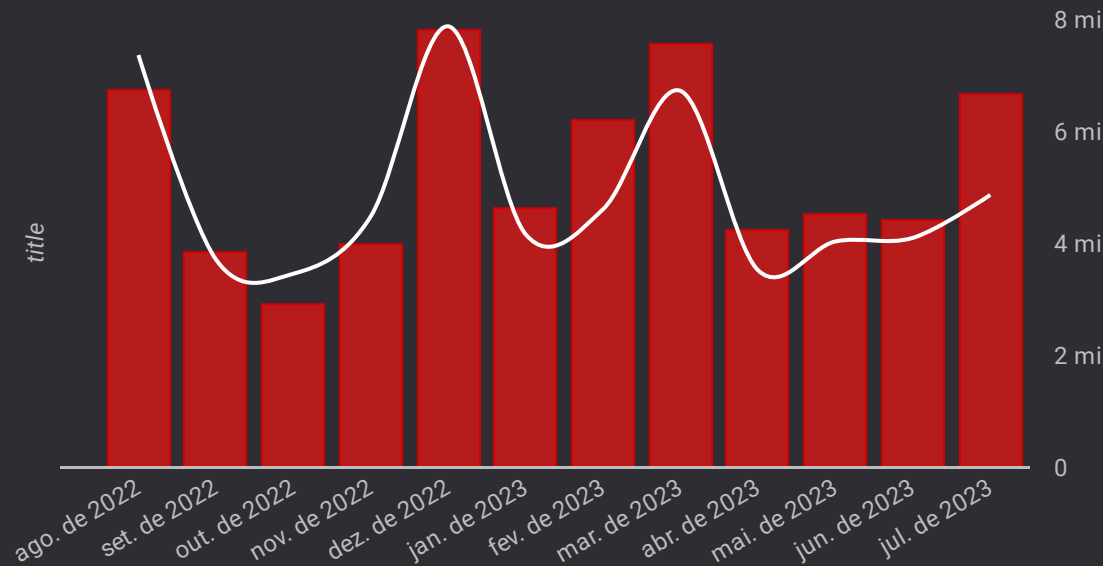
Top 10 Faixas



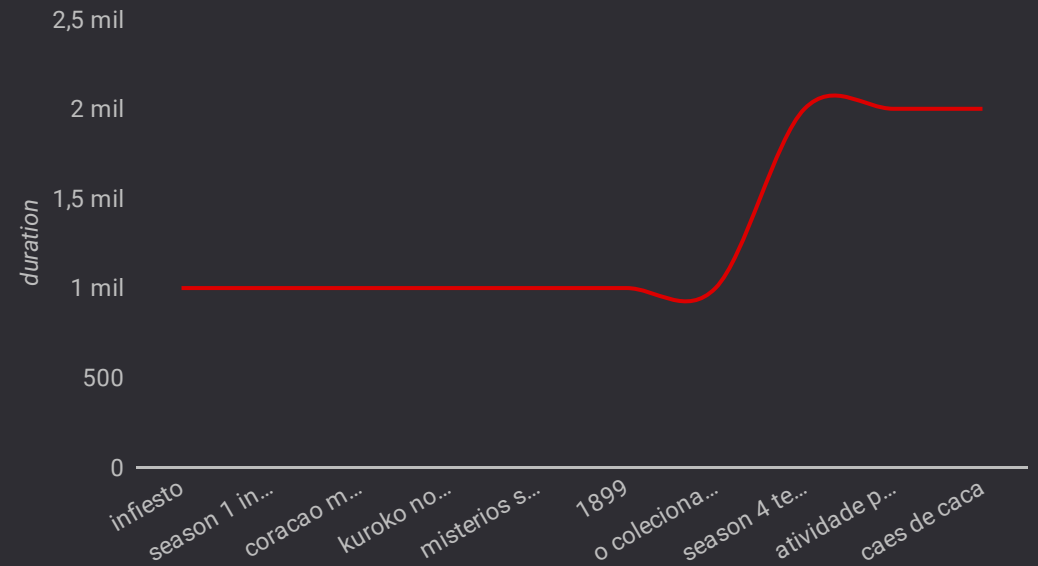
	start_time ▾	track_name	artist_name	playlist	duration
1.	17 de ago. de 2023	enemy	imaginedragons	Gostei	233.118
2.	16 de ago. de 2023	am i wrong	nico & vinz	Gostei	233.118
3.	16 de ago. de 2023	desce pro pla...	zaac	Gostei	168.507
4.	16 de ago. de 2023	feeling good	nina simone	Azul	173.786
5.	16 de ago. 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



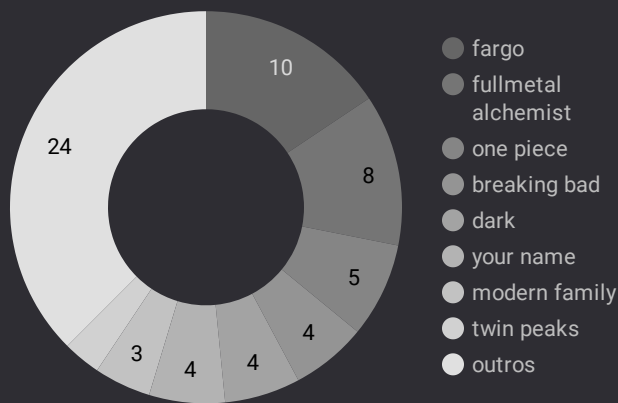
Top 10 Títulos



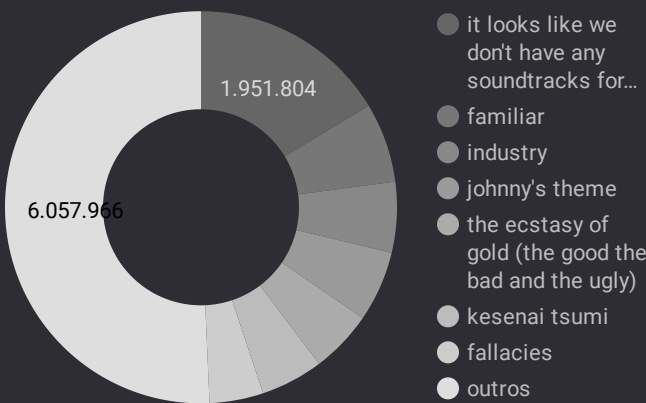
	start_time ▾	title	description	duration
1.	15 de ago. de 2023	temporada 3 (clipe)	Temporada 3 (Clipe): Sintonia	7000
2.	15 de ago. de 2023	temporada 1 (trailer)	Temporada 1 (Trailer): Record of Ragnarok	4000
3.	15 de ago. de 2023	temporada 2 (clipe)	Temporada 2 (Clipe): Baki Hanma	9000
4.	15 de ago. 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	minissérie (clipe)	Minissérie (Clipe): Um Conto de Fadas Perfeito	4000
8.	14 de ago. de 2023	temporada 2 (trailer)	Temporada 2 (Trailer): Vikings: Valhalla	4000



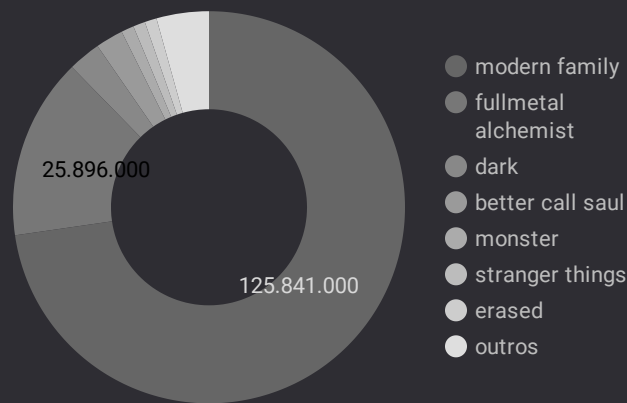
Quantidade de Faixas descobertas por Título



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