**PROJETO FINAL**

**REDES SOCIAIS:**

**Analisando dados do**

**Youtube**

**e**

**LinkedIn**

**Elaborado por:**

André Victor Moreira Costa

Eduardo Mathias

Thiago Regis

Victor de Oliveira Gonçalves

# **ÍNDICE**

# **REQUISITOS OBRIGATÓRIOS.....................................................................03**

# **OBJETIVOS....................................................................................................04**

# **MOTIVAÇÃO DO PROJETO..........................................................................04**

# **FERRAMENTAS.............................................................................................05**

# **FLUXOGRAMA..............................................................................................06**

# **KAGLE............................................................................................................07**

# **GOOGLE CLOUD STORAGE (Data Lake) ...............................................07**

# **ROTEIRO YOUTUBE.....................................................................................09**

# **ROTEIRO LINKEDIN......................................................................................32**

# **CONTATOs....................................................................................................43**

1. **REQUISITOS OBRIGATÓRIOS**

* Obrigatoriamente os datasets devem ter formatos diferentes (CSV / Json / Parquet / SQL / NoSQL) e 1 deles obrigatoriamente tem que ser em CSV.
* Operações com Pandas (limpezas, transformações e normalizações)
* Operações usando PySpark com a descrição de cada uma das operações.
* Operações utilizando o SparkSQL com a descrição de cada umas das operações.
* Os datasets utilizados podem ser em língua estrangeira, mas devem ao final terem seus dados/colunas exibidos na língua PT-BR.
* os datasets devem ser salvos e operados em armazenamento cloud, obrigatoriamente dentro da plataforma GCP (não pode ser usado Google drive ou armazenamento alheio ao google).
* os dados tratados devem ser armazenados também em GCP, mas obrigatoriamente em um data Lake (GSTorage), DW (BigQuery) ou em ambos.
* Deve ser feito análises dentro do Big Query utilizando a linguagem padrão SQL com a descrição das consultas feitas.
* Deve ser criado no Data Studio um Dashboard simples para exibição gráfica dos dados tratados trazendo insights importantes.
* E deve ser demonstrado em um workflow simples (gráfico) as etapas de ETL.

1. **OBJETIVO**

A proposta deste trabalho consiste em elaborar uma ETL (Extract, Transform and Load) para bancos de dados de redes sociais.

Deverá conter nessa ETL a base do banco de dados, o repositório onde será armazenado, como os dados serão tratados e manipulados, seu destino em um Data Lake/ Data Warehouse e por fim um dashboard com insights retirados dessa base no Google Data Studio.

1. **MOTIVAÇÃO DO PROJETO**

Possibilitar que potenciais clientes possam entender de forma visual quais são as tendências e categorias que geram alto engajamento nas redes sociais propostas, a fim de através dessas visualizações possam tomar decisões de como agir de acordo com o cenário observando diversos cenários.

* Youtube:

Com os dados desse Dataset, o objetivo é realizar uma leitura dos canais e vídeos mais acessados, comentados e curtidos, traçando uma tendência do público ao consumo de cada tipo de vídeo separado por categoria e país. Um dos principais objetivos é analisar os vídeos mais acessados por categoria e país, para analisar qual tipo onde e qual tipo de conteúdo é mais consumido. Com essa análise também foi possível verificar qual categoria e qual país mais interage nessa rede social, fazendo uma relação de Curtida por Visualização, Descurtida por Visualização e Comentário por Visualização.

O intervalo de tempo deste Dataset é de 12/08/2020 à 17/11/2021

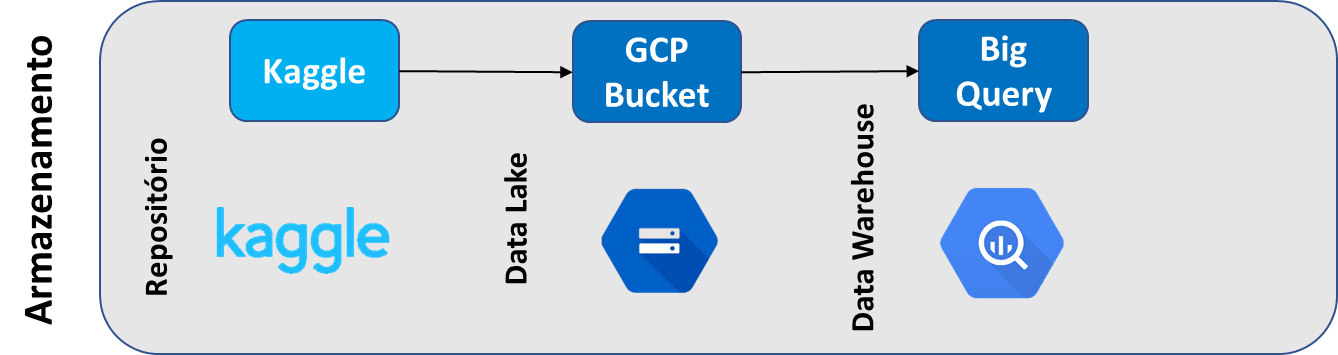
* LinkedIn:

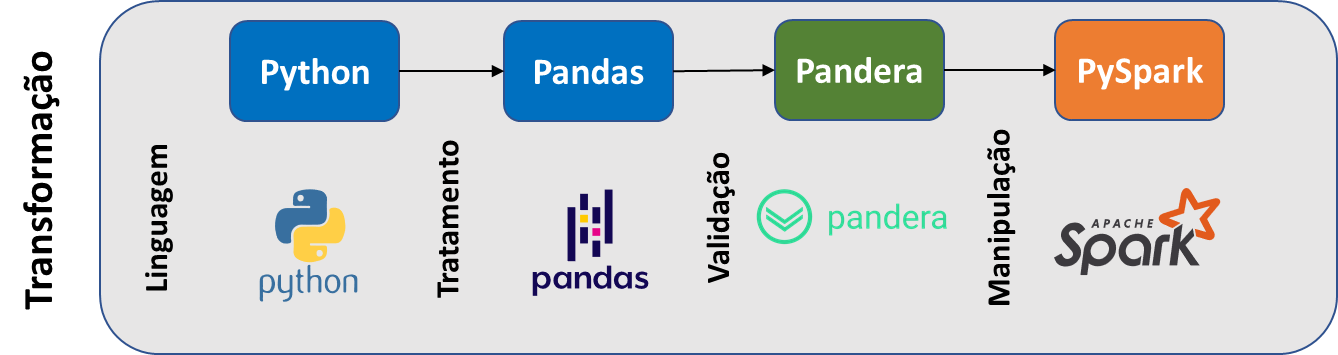
Já com os dados do Dataset do LinkedIn, o objetivo é realizar uma análise de proporção referente à etnia, nacionalidade e gênero, assim como eles se apresentam na foto de perfil. Um conjunto de fotos de cada usuário foi analisada para identificar as suas emoções, como felicidade, raiva, medo, tristeza entre outros. O objetivo foi traçar um paralelo com a quantidade de seguidores e promoções que possuem

Esse Dataset foi atualizado 2 anos atrás, em 2019.

1. **FERRAMENTAS**

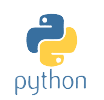
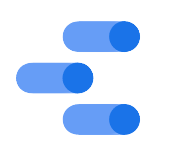


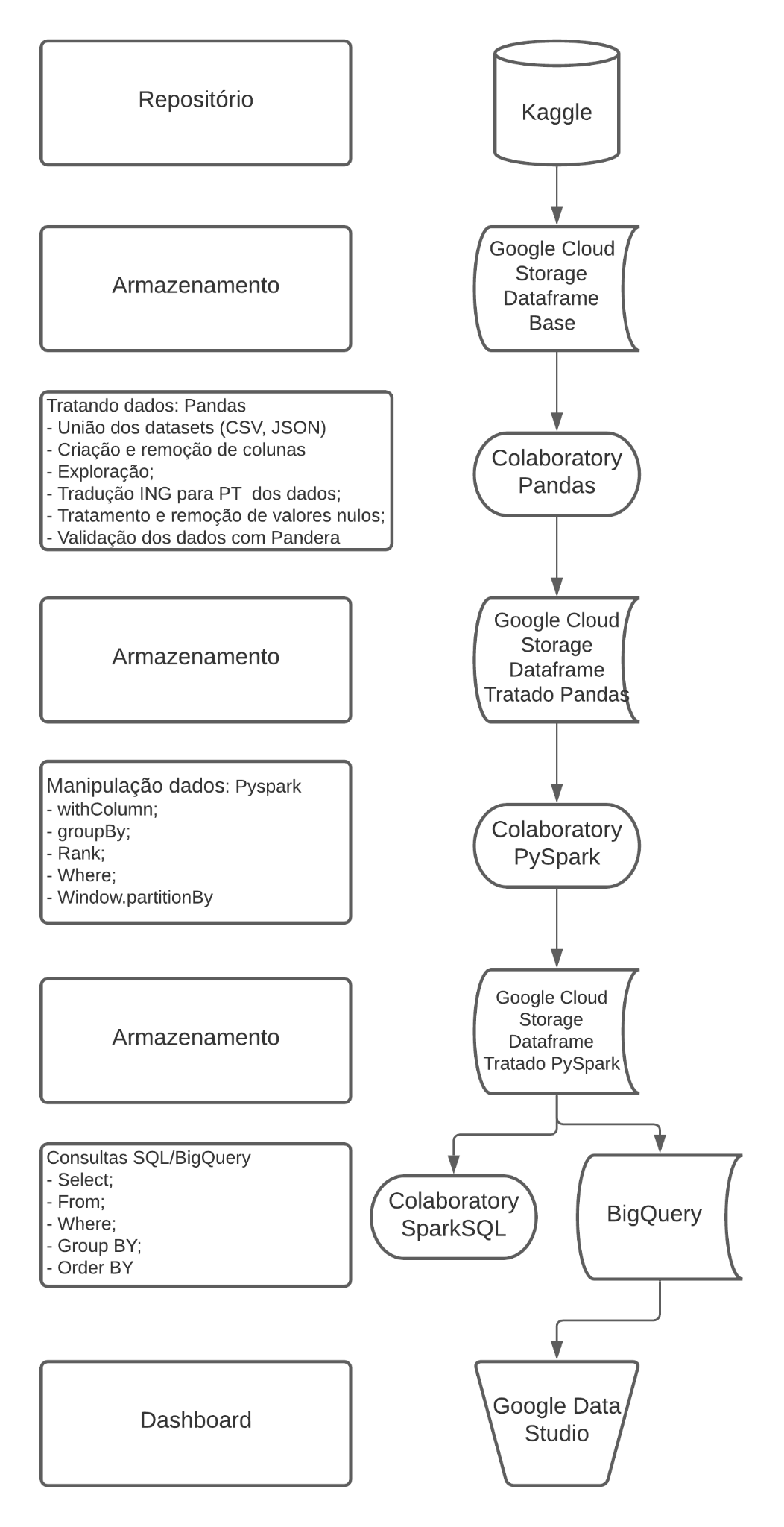
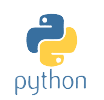
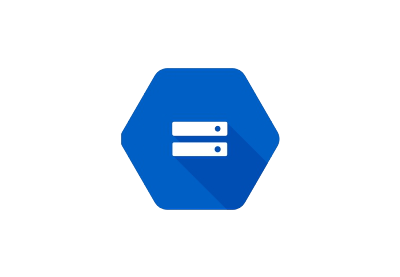
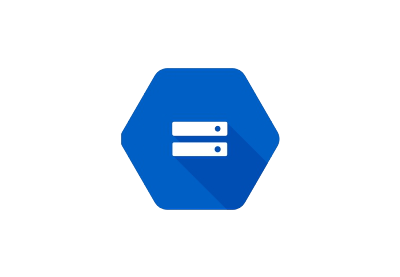
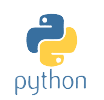
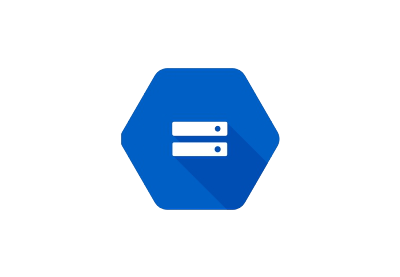






1. **FLUXOGRAMA**





1. **KAGGLE**

Kaggle é uma comunidade online voltada principalmente para cientistas de dados e praticantes de machine learning, onde os usuários podem procurar e publicar datasets de diversos assuntos. Os datasets públicas vem de fontes de usuários individuais, empresas de diversos portes e institutos de pesquisa, muitas estimulando os usuários a resolver cases da forma mais eficiente, estimulando com prêmios monetários. Devido esse ambiente propiciar bases de dados de alta qualidade, foram escolhidos dois datasets de redes sociais distintas (YouTube e LinkedIn) de acordo com o tema deste trabalho.





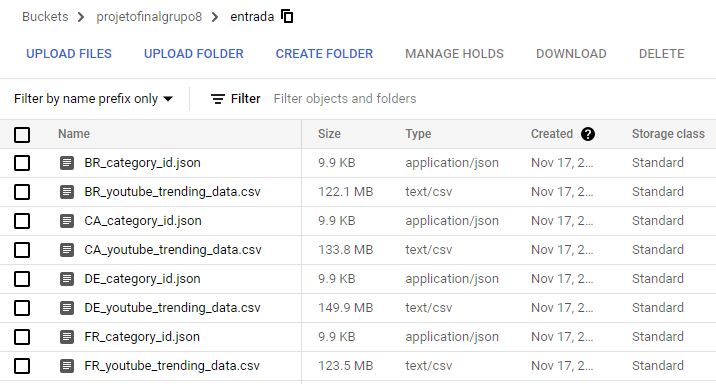
1. **GOOGLE CLOUD STORAGE (Data Lake)**

Todos os arquivos utilizados nos notebooks são armazenados no Bucket do Google Cloud Storage, para isso foram definidas, duas pastas, uma de “entrada”, onde ficam os dados retirados do Kaggle (a base de dados original) e uma pasta de “saída”, onde são armazenados os datasets resultantes das transformações feitas nos notebooks etapa por etapa.

Na pasta de “entrada”, estão os arquivos das duas bases de dados escolhidas:

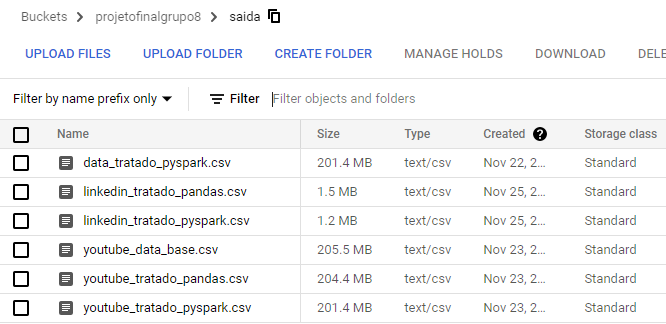
* [YouTube Trending Video Dataset (updated daily)](https://www.kaggle.com/rsrishav/youtube-trending-video-dataset?select=BR_youtube_trending_data.csv)
* [LinkedIn Profile Data](https://www.kaggle.com/omashish/linkedin-profile-data)

**Pasta de Entrada (Bucket)**



A pasta de “saída” armazena os arquivos tratados, transformados e manipulados nas etapas do projeto, cada um com um nome de acordo com a etapa que foi concluída e sendo usados cada um em um respectivo notebook do projeto, no BigQuery e também como base para o Dashboard no Google Data Studio.

**Pasta de Saída (Bucket)**



1. **ROTEIRO YOUTUBE**

[Base de dados youtube](https://www.kaggle.com/rsrishav/youtube-trending-video-dataset?select=BR_youtube_trending_data.csv)

A premissa de escolher esse dataset foi cumprir o requisito de usar arquivos de formatos diferentes (CSV e JSON) e ser um dataset interessante, onde fosse possível retirar insights.

Esse banco de dados é composto por 11 “CSV’s” de países diferentes com dados dos vídeos que estiveram em destaque no período de 12/08/2020 até 17/11/2021 e também arquivos “JSON” complementares, que trabalham como dicionários das categorias dos vídeos.

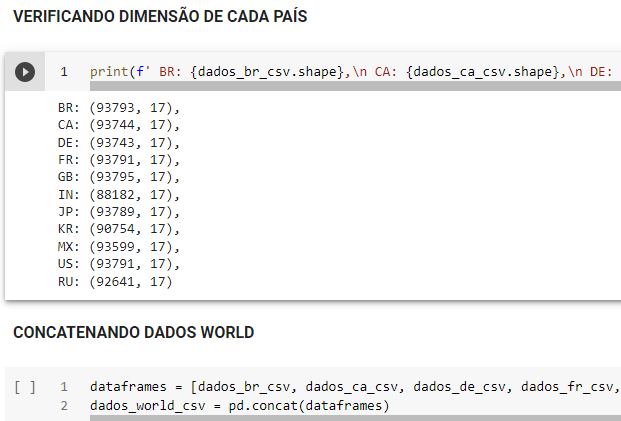




**1º notebook: 001\_SOCIAL\_MEDIA\_COMBINACAO\_DE\_DATASETS\_Nivel\_Pandas**

Neste notebook foram feitas as primeiras explorações dos dados, combinando os datasets de cada país em um só, criando uma coluna nova chamada “country” para que fosse possível identificar de qual país pertence o vídeo que entrou em destaque, usando o comando “concat” da biblioteca Pandas, assim gerando um dataset “mundo”.

O comando “concat” é uma forma de união dos datasets como se fosse um “empilhamento”, desde que as colunas sejam de mesmo nome e tipo de dados.

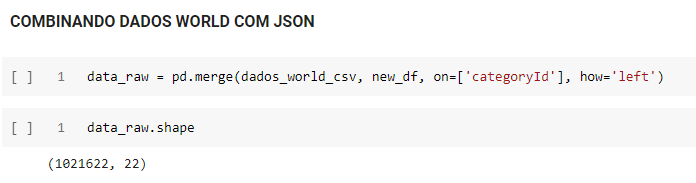


Decidimos manter as siglas dos países que compõem os dados, são eles:

* BR = Brasil
* CA = Canadá
* DE = Deutschland - Alemanha
* FR = França
* GB = Grã-Bretanha
* IN = Índia
* JP = Japão
* KR = República da Coréia
* MX = México
* US = Estados Unidos da América
* RU = Rússia

Em seguida foi necessário “descompactar” o arquivo “JSON” para que fosse possível trazer o nome das categorias para o dataset “mundo”, através de um LOOP FOR, assim, gerando um novo dataset “base” com pouco mais de 1 milhão de linhas e 22 colunas.





Optamos por remover 8 colunas que estão descritas dentro do notebook, por considerarmos que não haveriam informações interessantes dessas colunas, reduzindo para 14.

Por fim, exportamos essa primeira etapa do nosso banco de dados para o Google Cloud Storage com o nome “youtube\_data\_base.csv”.

Link do notebook: <https://colab.research.google.com/drive/1WnvhPZPWyKgbGIbuDqBOdxzxD0CZp-x0?usp=sharing>

**2º notebook: 002\_SOCIAL\_MEDIA\_TRATAMENTO\_DE DADOS\_Nivel\_Pandas**

Dando sequência ao projeto, foi carregado do Google Cloud Storage o “CSV” feito no final do 1º notebook.

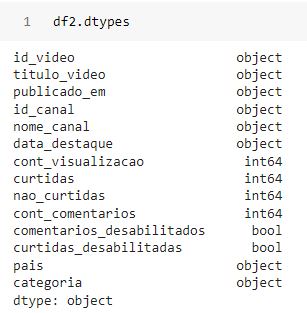
Nesta etapa, o foco principal é o tratamento dos dados, para que possam ser usados de forma mais “limpa” em etapas posteriores.

O primeiro passo foi traduzir os termos em inglês nos nomes das colunas e dentro das colunas de categoria:

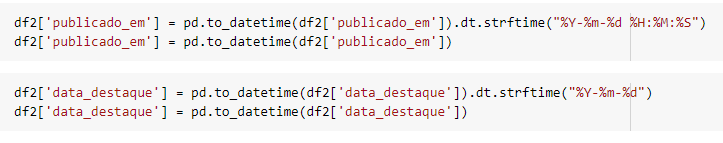


Em seguida foi verificado o tipo dos dados que estavam em cada coluna, através disso, pode-se notar que os dados dos campos relacionados à data, estavam em forma e texto, devido a isso, eles foram convertidos com seus devidos comandos como visto abaixo:

**Antes**  **Depois**

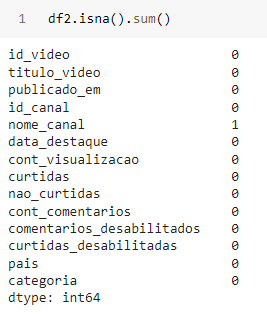
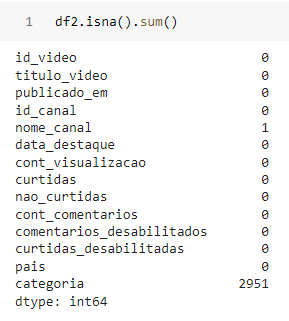


**Mudança de String para Datetime**

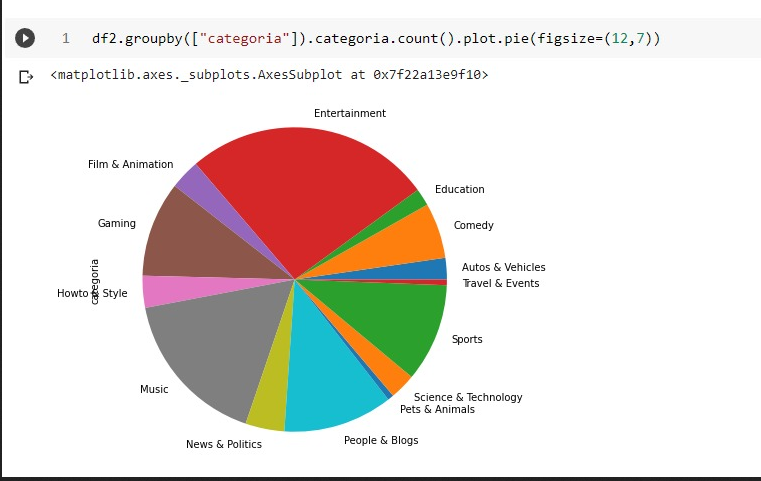


Logo foi a etapa de lidar com os dados nulos. Na coluna de categoria foram identificados quase 3000 linhas onde a categoria estava como dado nulo, foi decidido substituir para “outros” a fim de manter as informações.

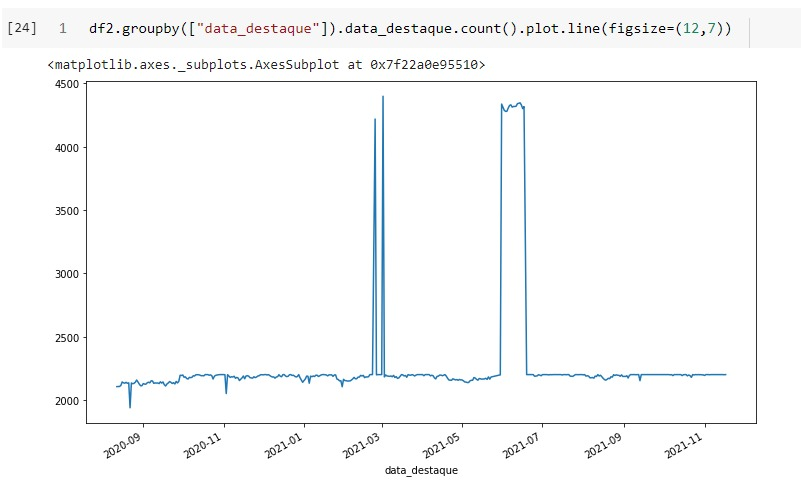
**Antes** **Depois**



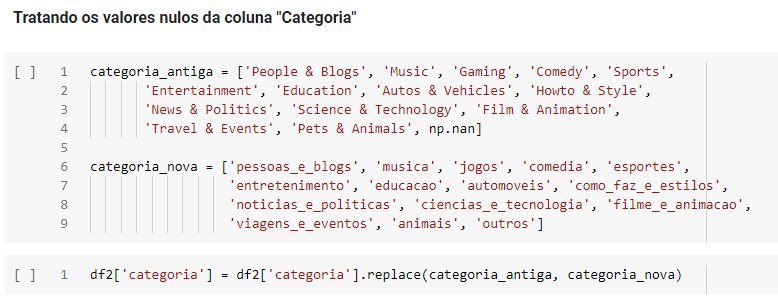
**Plotagem de Categoria do vídeo**

****

**Plotagem da data que vídeo entrou em destaque**



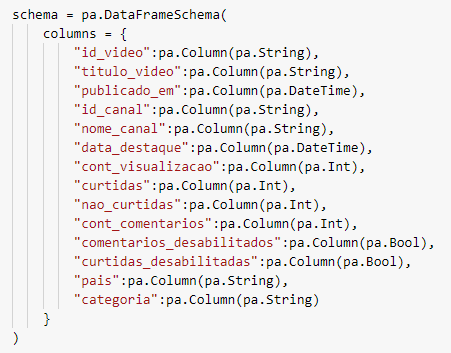
Usamos a plotagem para analisar possíveis desvios de dados em algumas colunas.



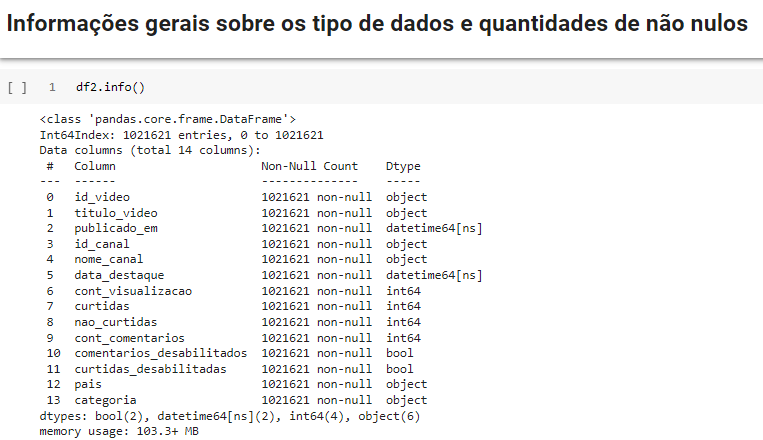
Ainda neste tópico, foi observado que apenas uma linha no dataset inteiro estava com a informação nula no coluna “nome\_canal”, devido a isso neste caso foi decidido remover essa linha como mostrado abaixo.



Seguindo a linha de raciocínio, onde os termos em inglês já foram traduzidos, o tipo dos dados foi estabelecidos e os dados nulos tratados, a fim de garantir a integridade do tipos dos dados por coluna, foi feita uma verificação dos dados com a biblioteca “Pandera” com sucesso no resultado.



Após todo esse tratamento os dados se apresentam da seguinte forma:



Finalizando este 2º notebook salvando o arquivo como “youtube\_tratado\_pandas.csv” no Google Cloud Storage para dar sequência a próxima etapa.

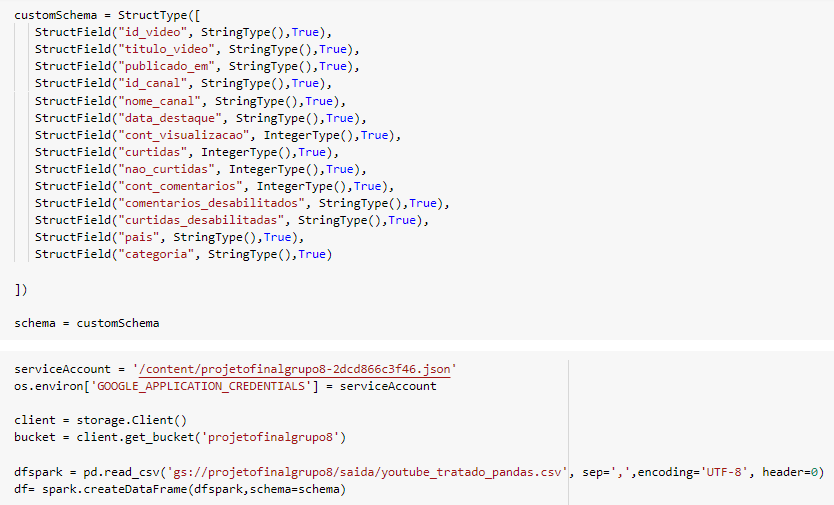


Link notebook: <https://colab.research.google.com/drive/1C8isdhWYarRfel_dIwKj-aMtJZgAGK4S?usp=sharing>

**3º notebook: 003\_SOCIAL\_MEDIA\_PARAMETROS\_Nivel\_Pyspark**

Iniciando a etapa de manipulação no PySpark, foi construído o “schema”, onde já é feita a verificação dos tipos dos dados, caso o dado seja incompatível com o tipo definido no “schema” a coluna vem com dados nulos, assim valendo como uma verificação da qualidade dos dados.

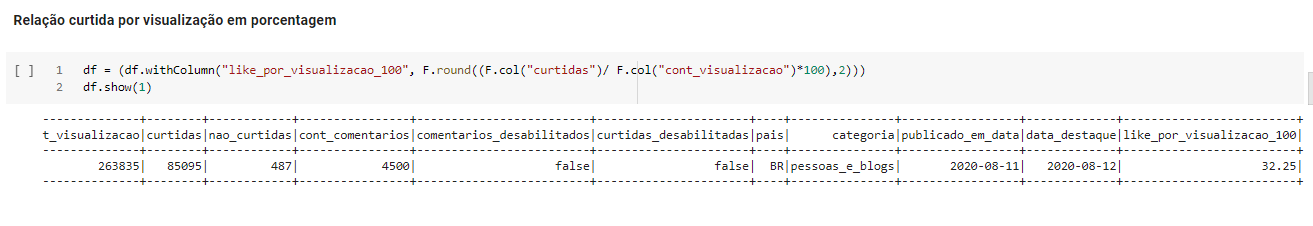
Para que fosse possível carregar o arquivo que foi tratado na etapa de Pandas, foi necessário primeiramente carregar o dataset com a biblioteca Pandas e logo na linha seguinte utilizar a criação de dataframe do PySpark para que fosse possível interpretar o dataset nesse formato como na figura abaixo:



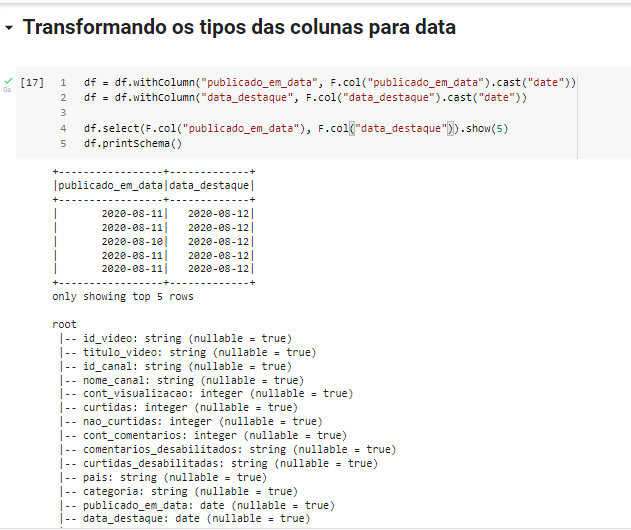
Na sequência foram criadas dentro desse dataset as colunas de publicação do vídeo e a data que entrou em destaque para o formato apenas do dia em que isso ocorre e posteriormente removidas as que estavam no formato antigo, como mostrado abaixo:



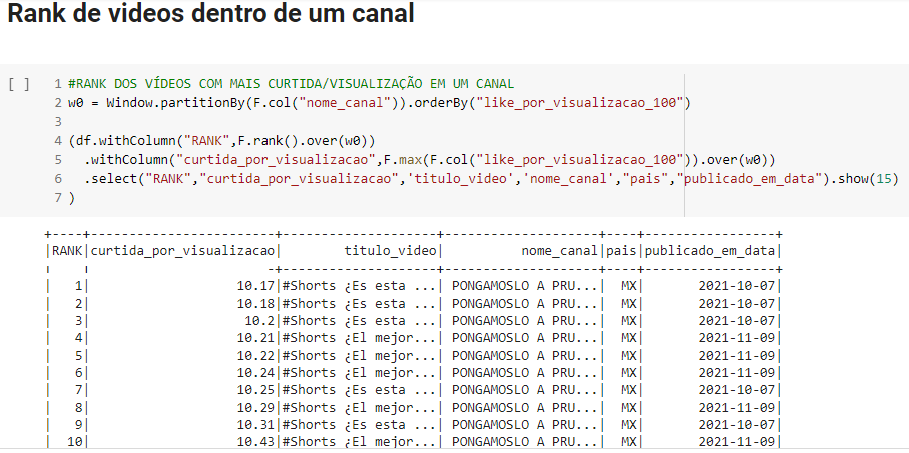
Aqui algumas novas colunas são criadas com o comando “.withColumn”.

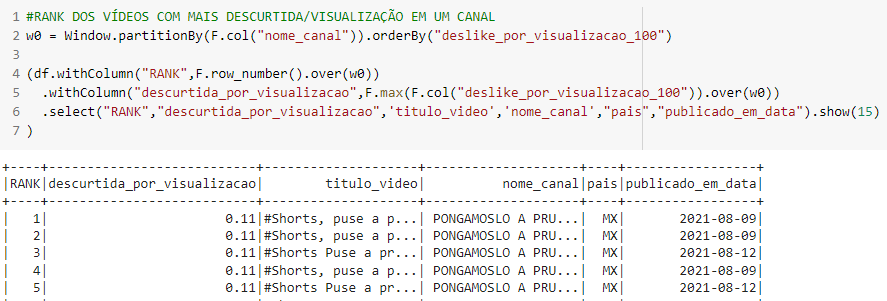


Como esse é um novo notebook, usando PySpark, também foi feita a transformação das colunas de data para o tipo “date” usando o comando “.cast()”

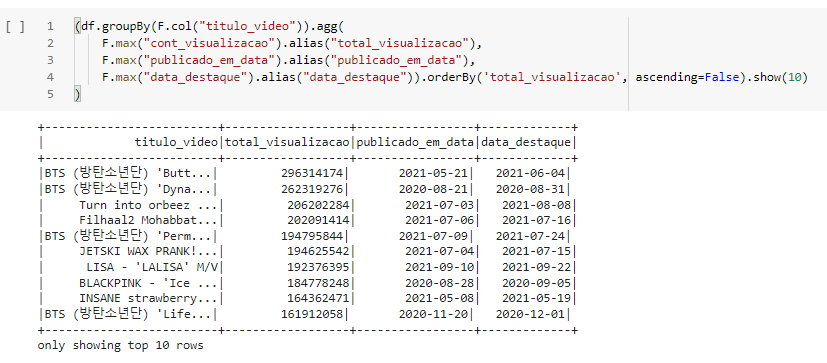


A partir daqui são mostrados insights usando a combinação de códigos para chegar em um resultado, segue abaixo as consultas:





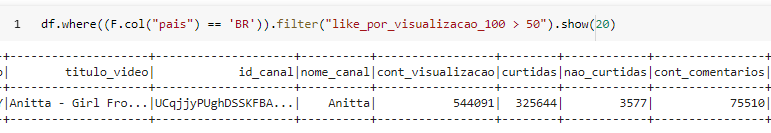
**Vídeos com mais visualizações**



**Canais com maior soma de visualizações**



**Vídeos com mais de 50% de curtidas em relação às visualizações**



Finalizando a etapa de PySpark, chega a hora de salvar o dataset manipulado no Google Cloud Storage.

O arquivo foi salvo com o nome “youtube\_tratado\_pyspark.csv”.

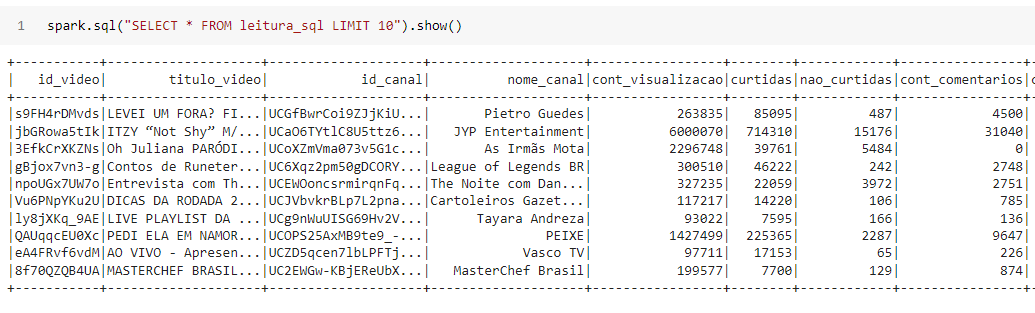


Link notebook: <https://colab.research.google.com/drive/1nitGJWjnnipgUPr-m8hntc1VHwyXvaN1?usp=sharing>

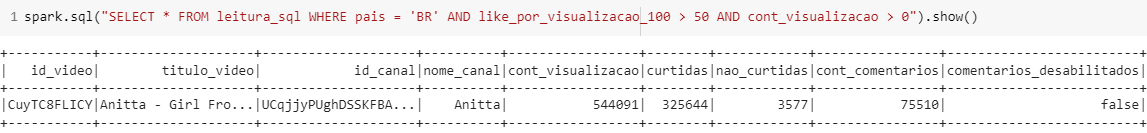
**4º Notebook: 004\_SOCIAL\_MEDIA\_CONSULTAS\_Nivel\_SparkSQL**

Neste notebook foram feitas as consultas SQL através do SparkSQL, estas mesmas consultas estão sendo usadas no BigQuery (Data Warehouse) do projeto logo após este tópico, segue abaixo algumas amostras logo abaixo de todas as consultas.

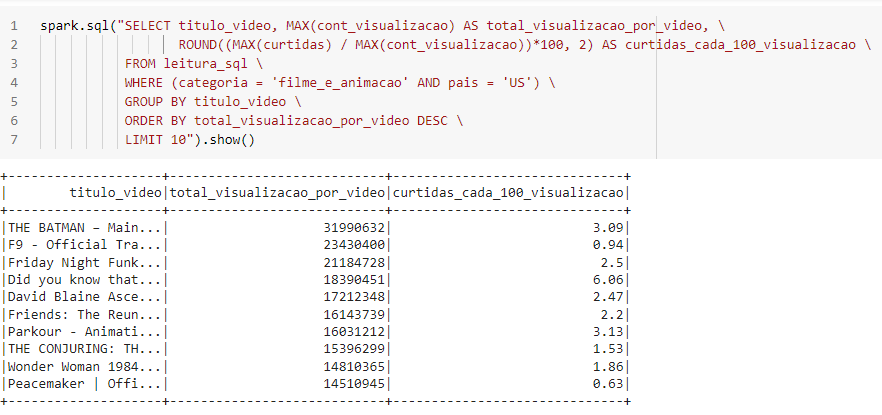
**00.Consulta geral da tabela:**



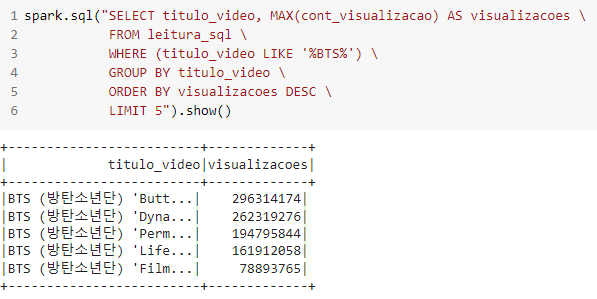
**01.CONSULTA VIDEOS COM MAIS DE 50% DE CURTIDAS EM RELAÇÃO ÀS VISUALIZAÇÕES**



**02.Consulta vídeos da categoria filmes e animações nos EUA com sua visualizações e percentual de curtidas por visualização**



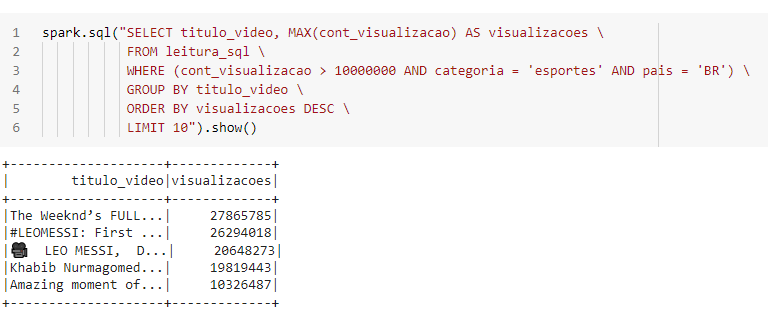
**03.CONSULTA VÍDEOS QUE TENHAM O NOME DO BTS COM MAIS VISUALIZAÇÕES**



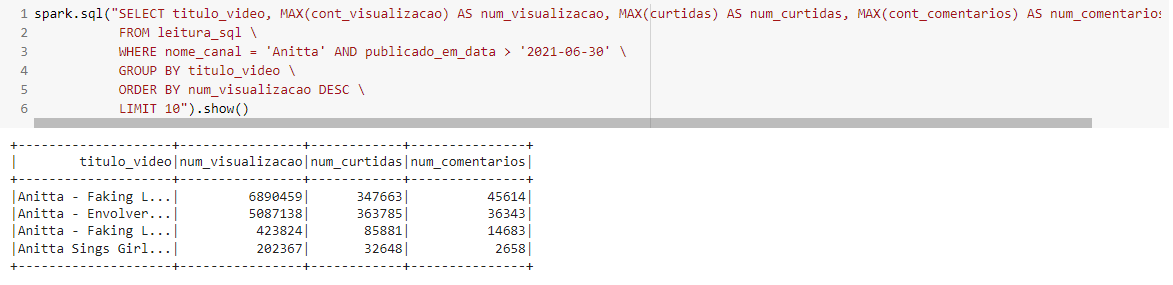
**04.CONSULTA VÍDEOS QUE TENHAM O NOME DO NEYMAR COM MAIS VISUALIZAÇÕES**

****

**05.Consulta de vídeos com mais de 10 milhões de visualizações da categoria esportes do youtube Brasil**



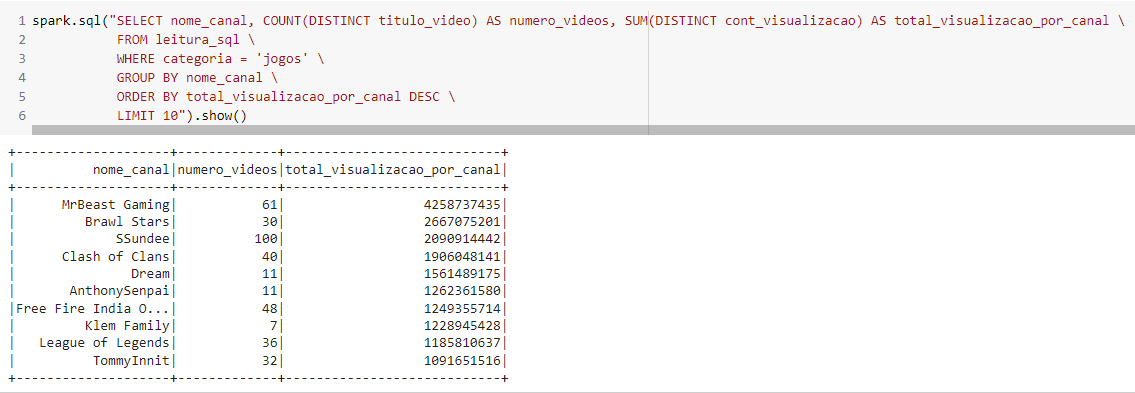
**06.CONSULTA VIDEOS ANITTA COM NÚMERO DE VISUALIZAÇÕES, CURTIDAS E COMENTÁRIOS PUBLICADOS DEPOIS DE 2021-06-30**



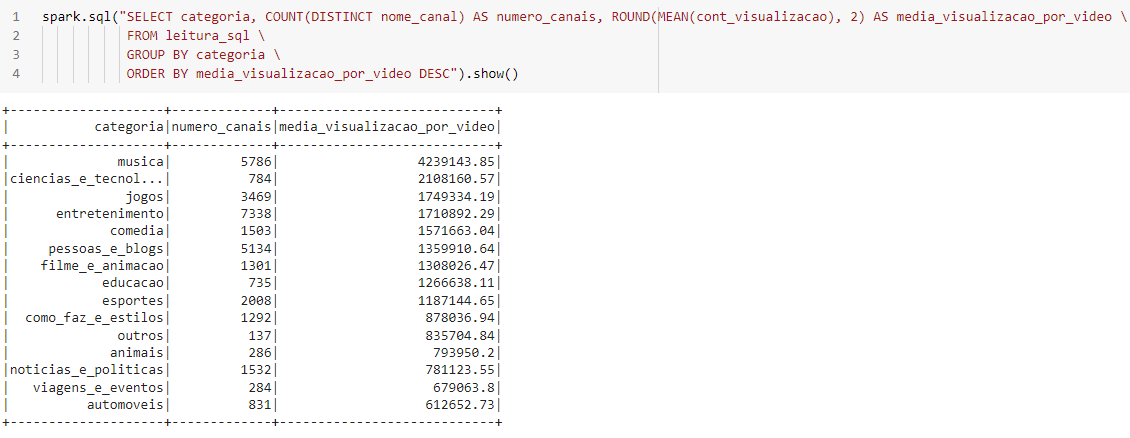
**07.CONSULTA VIDEOS ANITTA COM NÚMERO DE VISUALIZAÇÕES, CURTIDAS E COMENTÁRIOS**



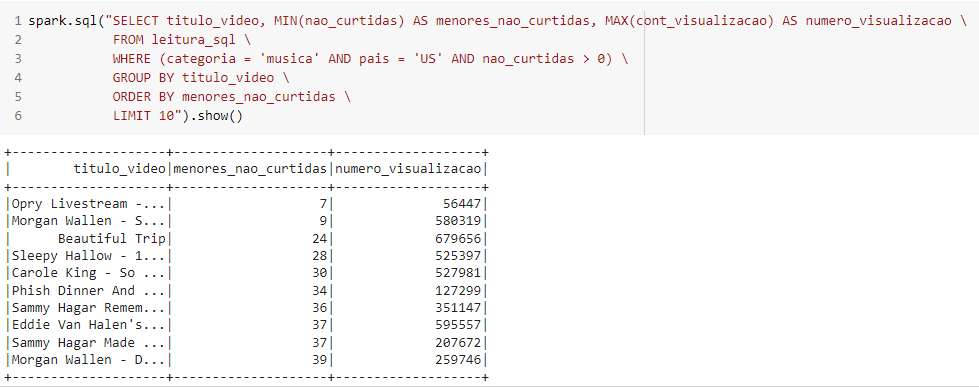
**08.CONSULTA DE TOTAL DE VISUALIZAÇÕES E NUMERO DE VIDEOS DOS CANAIS DA CATEGORIA DE JOGOS**



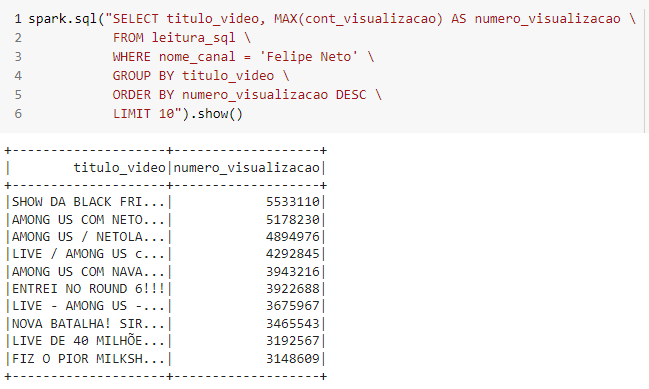
**09.Consulta número de canais e média de visualizações por vídeo da categoria**



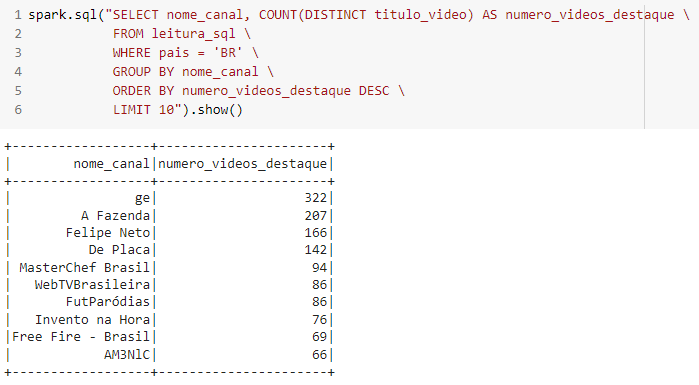
**10.CONSULTA VIDEOS COM MENOR NÚMERO DE DESCURTIDAS MAIORES QUE ZERO NA CATEGORIA MÚSICA NOS EUA, MOSTRANDO NÚMERO DE VISUALIZAÇÕES**



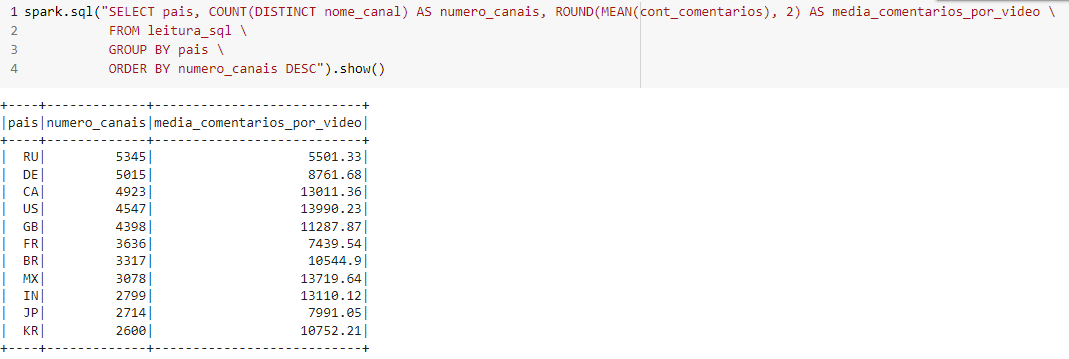
**11.CONSULTA VIDEOS DO CANAL DO FELIPE NETO COM MAIOR NÚMERO DE VISUALIZAÇÕES**



**12.CONSULTA CANAIS BRASILEIROS COM MAIOR NÚMERO DE VÍDEOS PUBLICADOS**



**13.CONSULTA NÚMERO DE CANAIS POR PAÍS E SUA MÉDIA DE COMENTÁRIOS POR VÍDEO**



**14.CONSULTA VIDEOS COM MAIS VISUALIZAÇÕES**



**Link do notebook:** <https://colab.research.google.com/drive/10s6seXWQckPZvza7XJozd80KutQ7jvt9?usp=sharing>

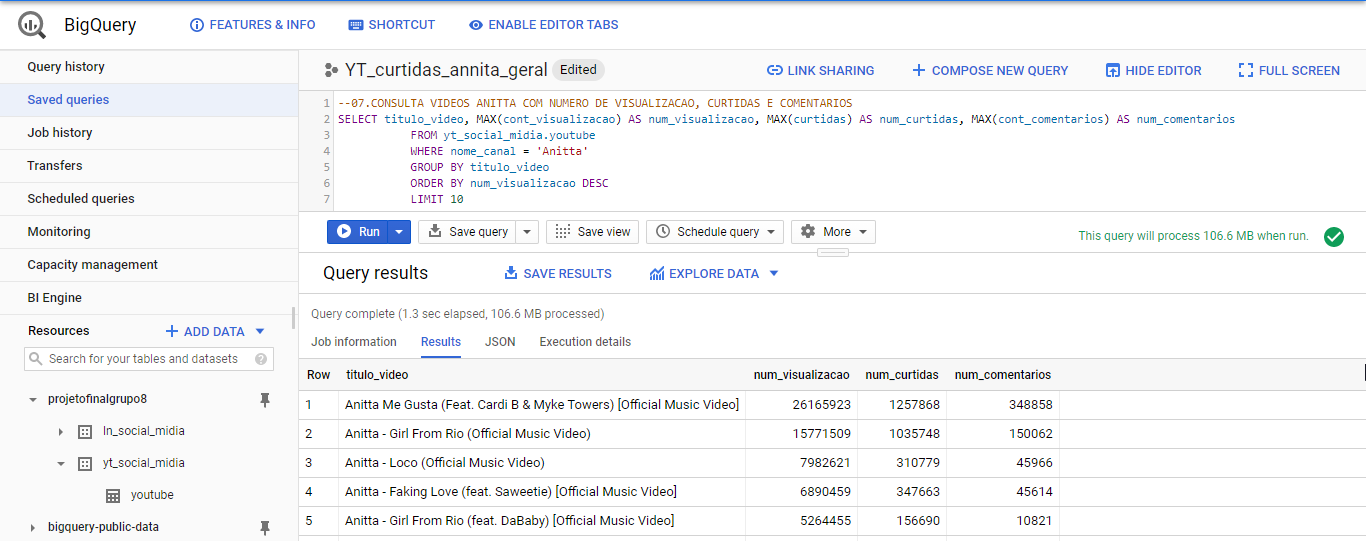
**BIGQUERY (DATA WAREHOUSE) - YOUTUBE**

Data Warehouse é onde os dados estão dispostos de forma mais organizada e prontos para consulta no Google Cloud.

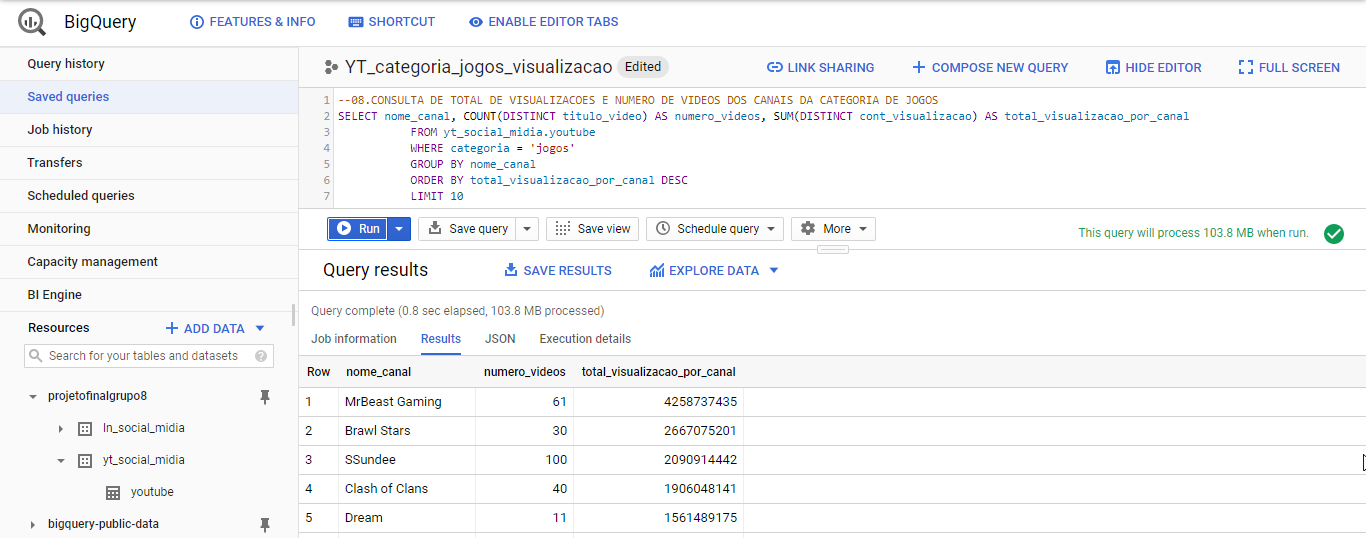
Os datasets usados nessa etapa podem vir via upload externo, mas nesse caso, foi migrado diretamente do Google Cloud Storage, onde todas as etapas dos bancos de dados usados, estão armazenados, e nesse ponto sendo usado para consultar através da linguagem SQL dentro do BigQuery.

Logo abaixo será mostrado algumas amostras, de como são as consultas salvas nesse Data Warehouse. Elas estão seguindo a mesma lógica das consultas do sparkSQL.

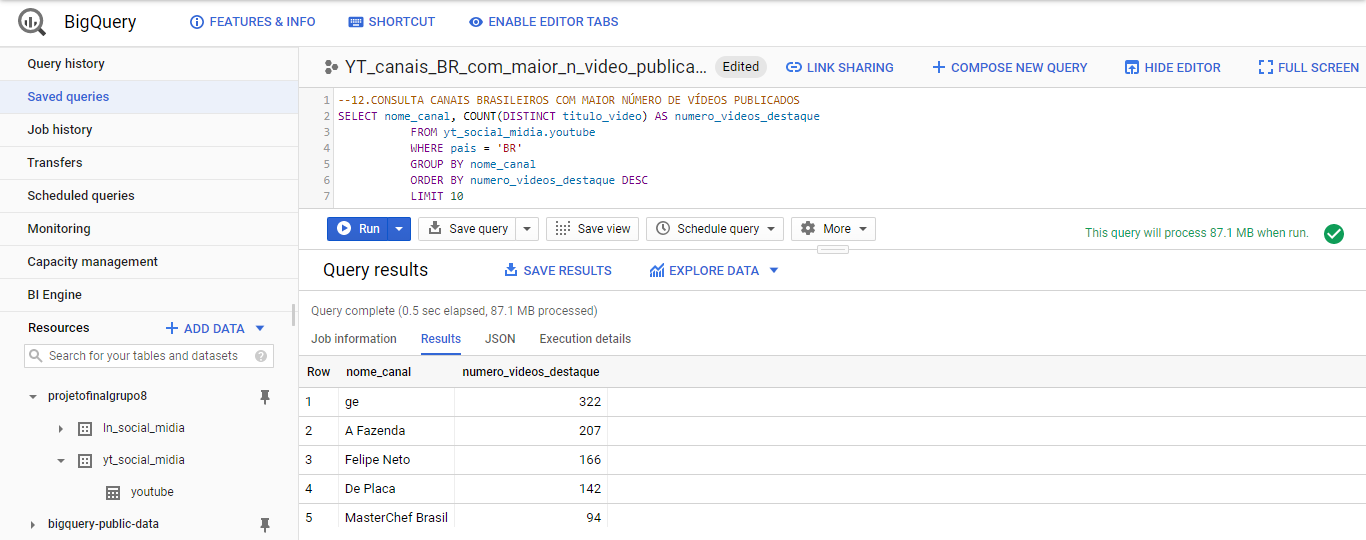
**07.Consulta videos Anitta com número de visualizações, curtidas e comentários**



**08.Consulta de total de visualizações e número de vídeos dos canais da categoria de jogos**



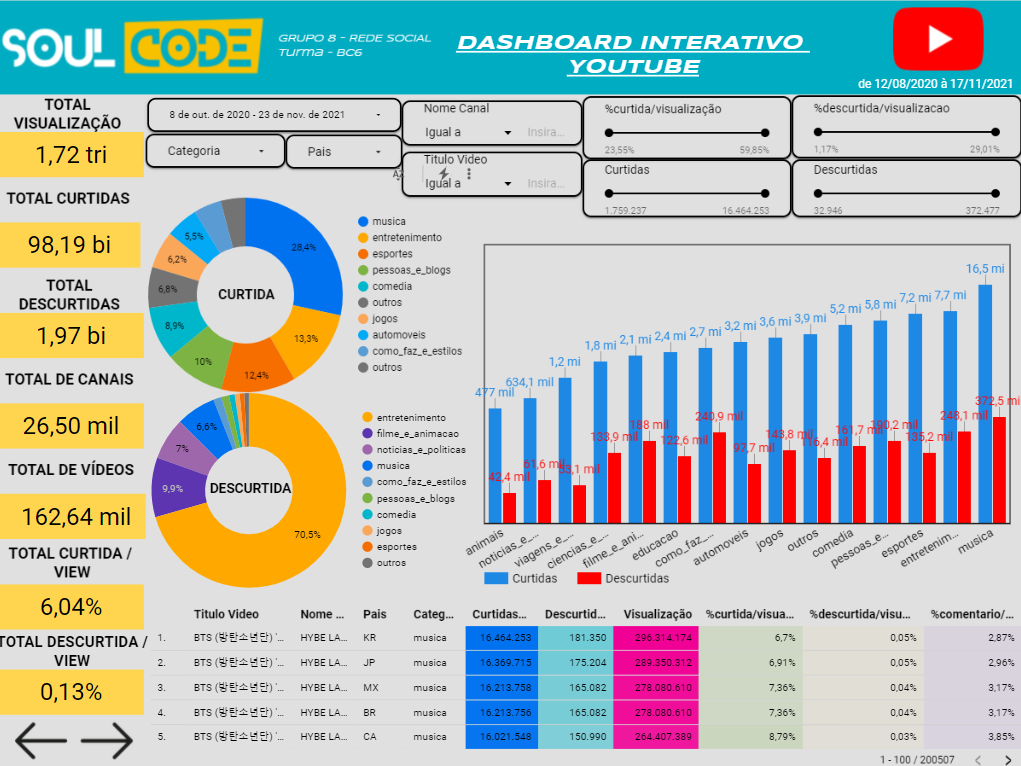
**12.Consulta canais brasileiros com maior número de vídeos publicados**

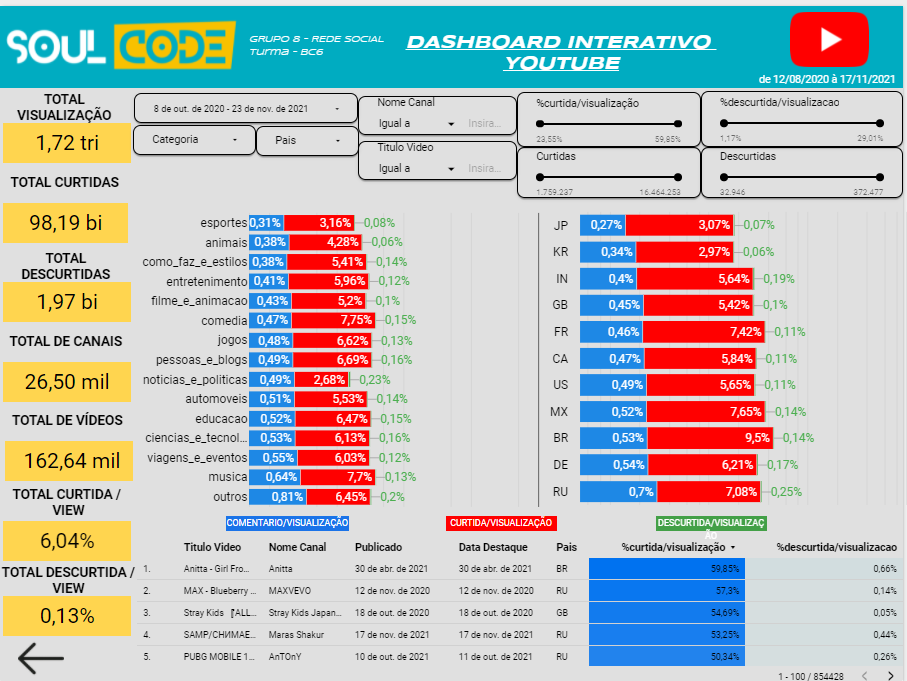


**DATA STUDIO - YOUTUBE**

O Google Data Studio é uma ferramenta online de Bussiness Inteligence, onde é possível criar dashboards interativos, de forma a apresentar as conclusões de forma visual com gráficos e planilhas interagindo de forma simplificada.

Essa sendo a última etapa do projeto, é possível mostrar todas as consultas feitas ao longo do projeto graficamente, trazendo outra forma de apresentar os dados com gráficos e números, tornando a visão do usuário final mais clara e direta. O dashboard pode ser visualizado clicando [aqui](https://datastudio.google.com/s/quWmDEXq278).





**Alguns insights extraídos:**

* Vídeo com mais visualizações em 6 países: BTS (방탄소년단) 'Butter' Official MV
* Categoria que tem mais curtidas respectivamente são: Música, entretenimento e Esporte
* Categoria com mais descurtidas: Entretenimento
* Brasil e Rússia são os países que mais interagem na plataforma.
* O vídeo com mais curtidas em relação ao número de visualizações no Brasil é Anitta - Girl From Rio, já na Alemanha é 𝗸𝗶𝗹𝗹𝗶𝗻𝗴 𝗦𝗧𝗔𝗟𝗞𝗜𝗡𝗚 Episode .O1 [Ref. Video]
* No brasil, o top 3 categoria, respectivamente é: Música, Entretenimento, Pessoas e Blogs
* Os vídeos mais descurtidos são da Índia.

Vale frisar que, na categoria entretenimento há uma mescla de conteúdos que podem varias como jogos, filmes e comédia, por exemplo.

1. **ROTEIRO LINKEDIN**

[Base de dados linkedin](https://www.kaggle.com/omashish/linkedin-profile-data)

A premissa deste banco de dados foi verificar que tipos de informações diferentes podem ser encontradas em uma rede social diferente.

**1º notebook: 001\_LINKEDIN\_SOCIAL\_MEDIA\_TRATAMENTO\_DE\_DADOS\_Nivel\_Pandas**

Seguindo a mesma tratativa da análise exploratória feita no roteiro de youtube, esse dataset do linkedin possuía a princípio 49 colunas e 62706 linhas.

A primeira ação foi remover os “id’s” duplicados da coluna “m\_urn\_id”, resumindo o perfil dos “influencers” deste dataset apenas para a vaga mais atual que eles ocupam.

Logo após foram renomeados os nomes das colunas traduzindo elas do inglês para o português e também o conteúdo das colunas com dados categóricos (texto).

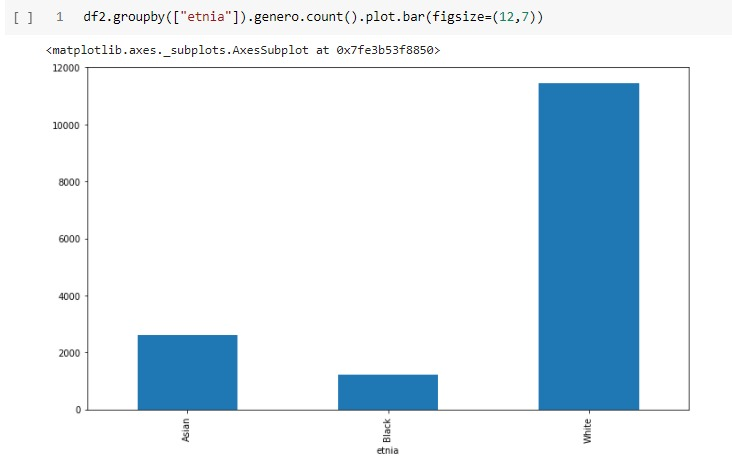
A forma escolhida para tradução, foi criar duas listas, uma com as palavras em inglês e outra com as palavras em português. Depois, utilizamos o comando “replace” com as duas listas para alterar todos os campos de cada coluna.



Para fim de maior compreensão do banco de dados, uma breve explicação sobre alguns grupos regionais:

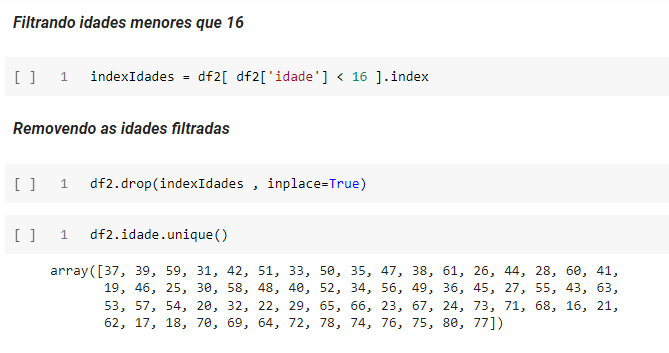
* O Sul asiático (asiatico\_sul) também é conhecido como Ásia Meridional, essa região é composta por um grupo de países, são eles: Bangladesh, Butão, Índia, Maldivas, Nepal, Paquistão e Sri Lanka.
* O Leste asiático (asiatico\_leste) também é conhecido como Ásia Oriental, essa região é composta por: China, Coreia do Sul, Coreia do Norte, Japão, Mongólia e Taiwan.
* O Hispânico (hispânico) são os países da américa latina e Espanha.
* O Nórdico (nordico) são os países: Noruega, Suécia, Dinamarca, Finlândia e Islândia.

**Plotagem de Etnia**



**Plotagem de Idade**Plotagem aqui também foi usada para verificar possíveis desvios.

Nessa linha foram feitos mais alguns filtros, como retirar as pessoas com idade inferior a 16 anos e aquelas com valor negativo no tempo de cargo atual



Por fim validando o tipo dos dados das colunas com Pandera:

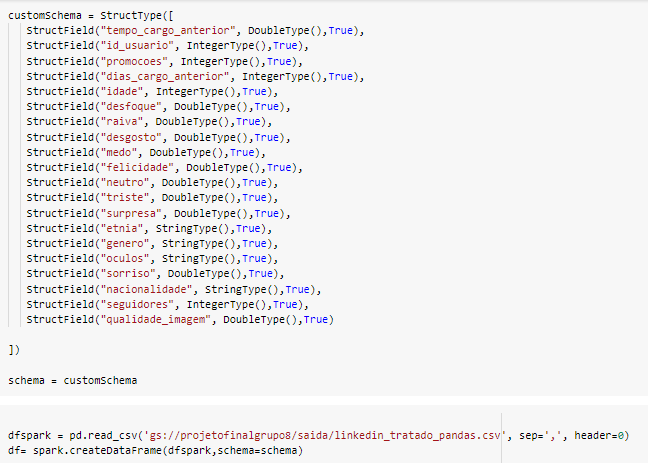


Ao fim seguindo com todo tratamento e validação do dataset, este se resumiu a 20 colunas e 11841 linhas, por fim finalizando esta etapa salvando o dataset como “linkedin\_tratado\_pandas.csv” no Google Cloud Storage.

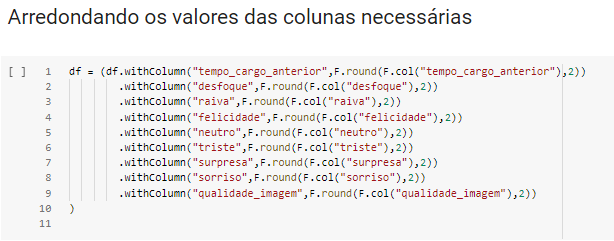
Link do notebook: <https://colab.research.google.com/drive/1WXjY6Jf9ayKAqc29Val1zkP82wwYef4O?usp=sharing>

**2º notebook: 002\_LINKEDIN\_SOCIAL\_MEDIA\_PARAMETROS\_Nivel\_Pyspark**

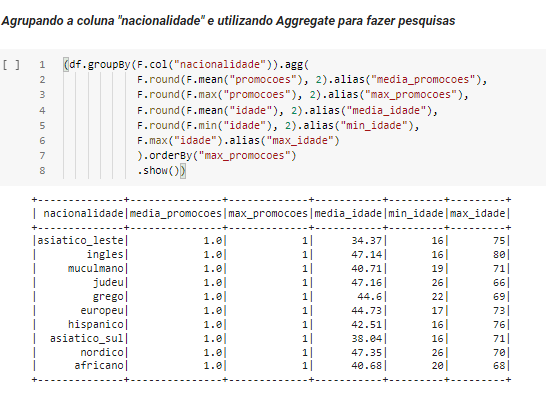
A etapa de manipulação com PySpark se inicia com a validação do tipo dos dados, carregando o arquivo “linkedin\_tratato\_pandas.csv” com a montagem do “schema” do PySpark no carregamento.



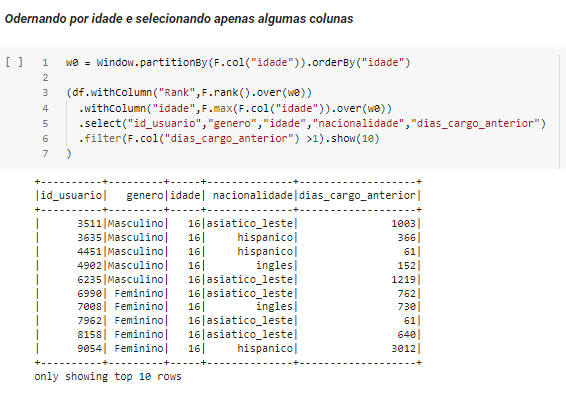
Logo em seguida o primeiro passo feito foi arredondar para duas casas decimais os campos numéricos, a fim de criar um padrão para o dataset.



Foram gerados na sequência alguns insights com a combinação dos comandos “groupBy” e “aggregate”.



Como uma próxima etapa, foram utilizadas as "Window Functions" de forma a demonstrar algumas outras ideias.



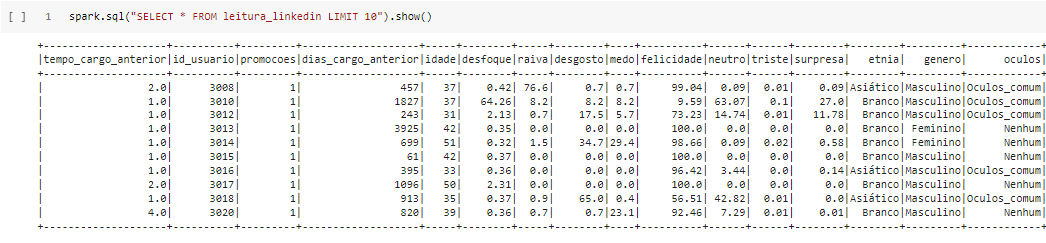
Finalizando mais uma etapa salvando o dataframe no Google Cloud Storage, o data lake usado durante as etapas deste projeto como “linkedin\_tratado\_pyspark”.

Link do notebook: <https://colab.research.google.com/drive/1EunkrtzvQD2AGaTleXVudzmj8BXrF4ja?usp=sharing>

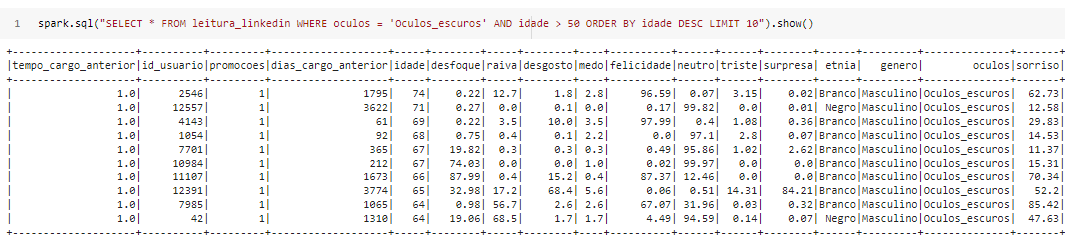
**3º Notebook: 003\_LINKEDIN\_SOCIAL\_MEDIA\_CONSULTAS\_Nivel\_SparkSQL**

Neste notebook foram feitas as consultas SQL através do SparkSQL, estas mesmas consultas estão sendo usadas no BigQuery (Data Warehouse) do projeto logo após este tópico, segue abaixo algumas amostras logo abaixo de todas as consultas.

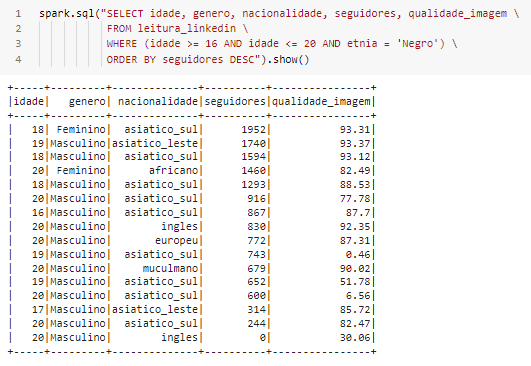
**00.Consulta geral tabela**



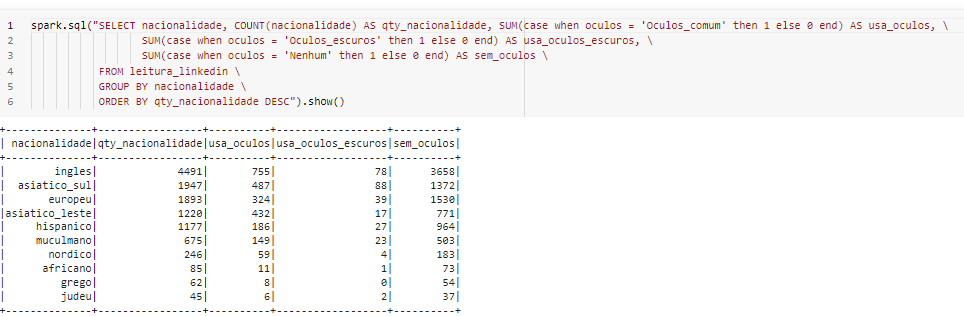
**01.Pessoas com mais de 50 anos que usam óculos escuros**



**02.Consulta pessoas negras entre 16 e 20 anos com suas respectivas nacionalidades e número de seguidores**

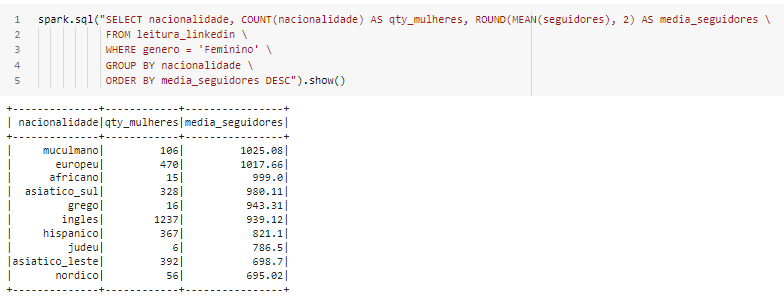


**03.Consulta pessoas que usam ou não óculos por nacionalidade**

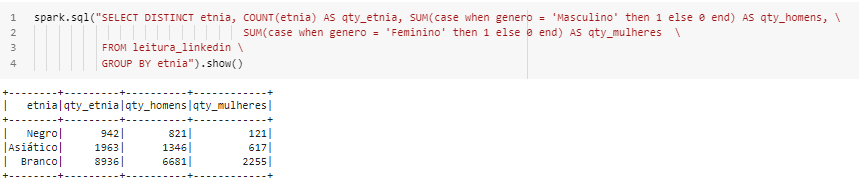


**04.Consulta média de seguidores e quantidade de mulheres influenciadoras no**

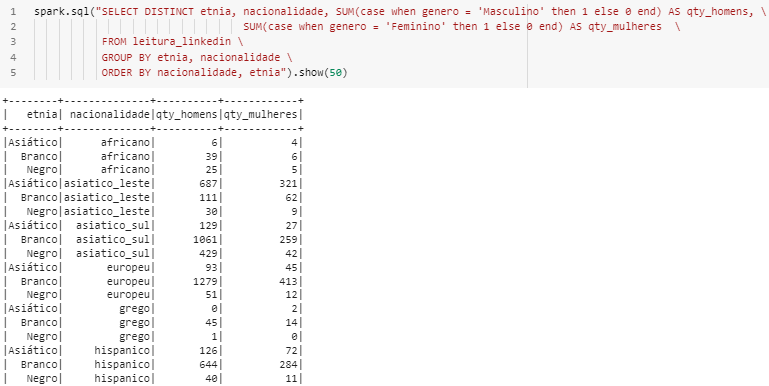
**LinkedIn por nacionalidade**



**05.Consulta quantidade de homens e mulheres por etnia**



**06.Consulta número de homens e mulheres por etnia, nacionalidade e sexo**



**07.Consulta quantidade de homens e mulheres entre 36 e 40 anos de acordo com a etnia**



**08.Consulta idades (mais velho, média e mais novo) por etnia e quantos usam ou não usam óculos**



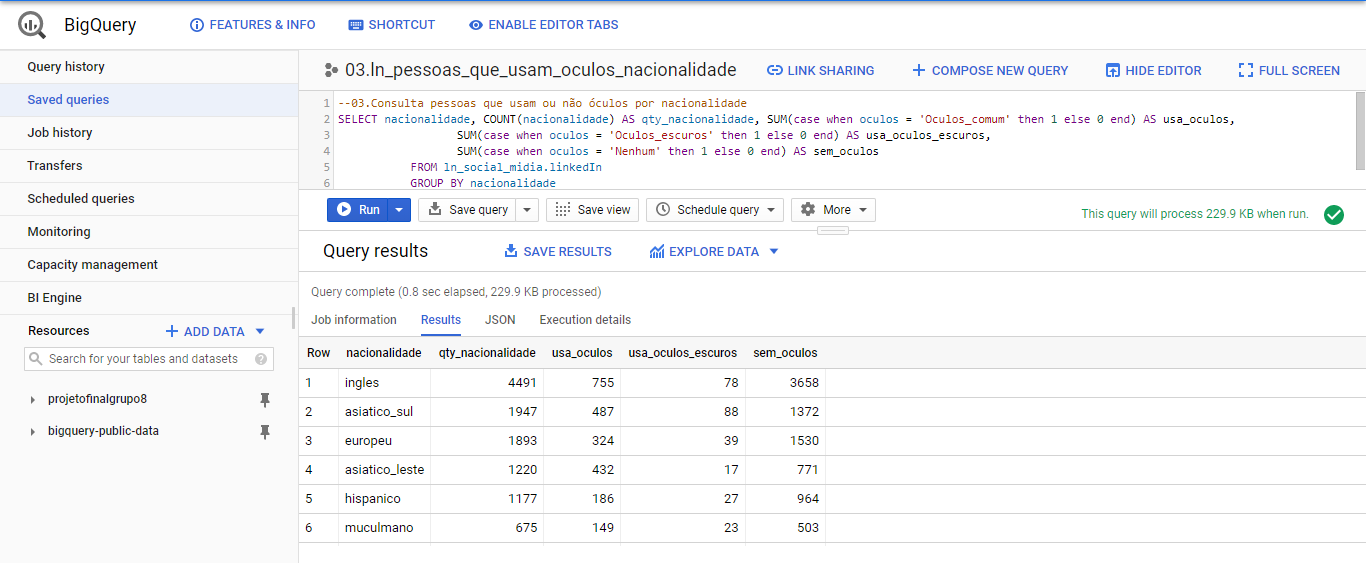
**Link do notebook:** <https://colab.research.google.com/drive/1YYSgxd9lrnmi0e8fQurUsb-eu9LMmdN1?usp=sharing>

**BIGQUERY (DATA WAREHOUSE) - LINKEDIN**

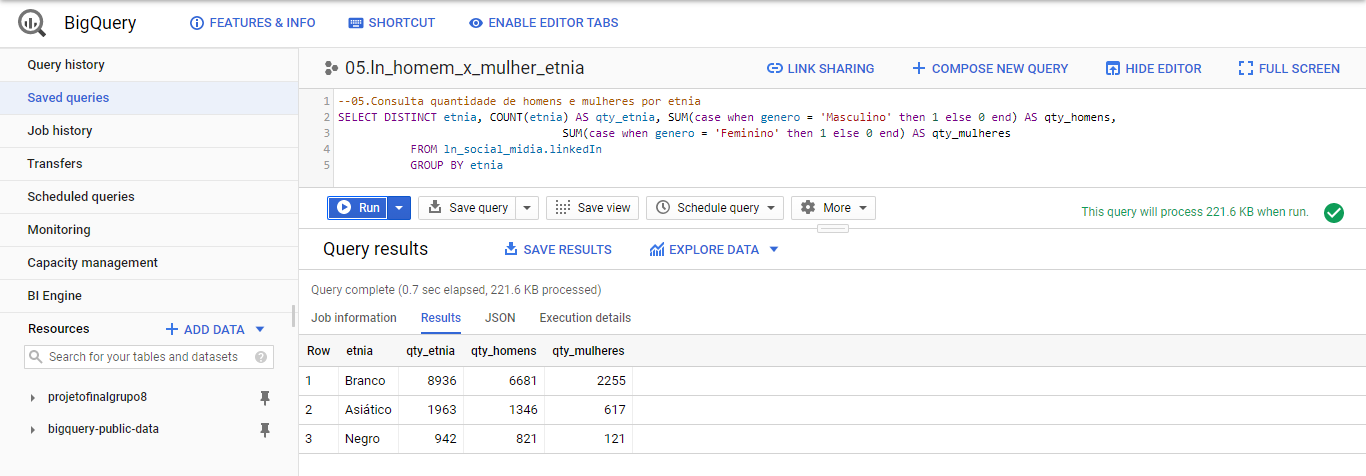
Seguindo a mesma lógica desse tópico para o dataset do youtube já mostrado anteriormente, neste ponto as consultas são feitas através da linguagem SQL dentro do BigQuery.

Logo abaixo será mostrado algumas amostras,de como são as consultas salvas nesse Data Warehouse. Elas estão seguindo a mesma lógica das consultas do SparkSQL.

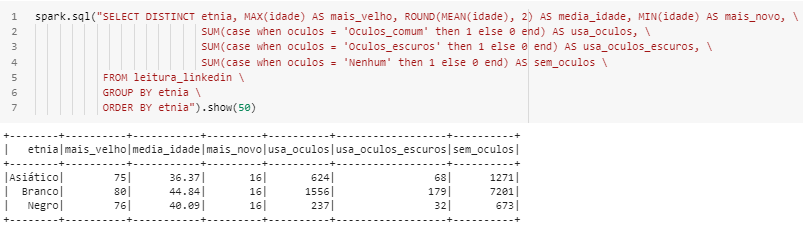
**03.Consulta pessoas que usam ou não óculos por nacionalidade**



**05.Consulta quantidade de homens e mulheres por etnia**



**08.Consulta idades (mais velho, média e mais novo) por etnia e quantos usam ou não usam óculos**

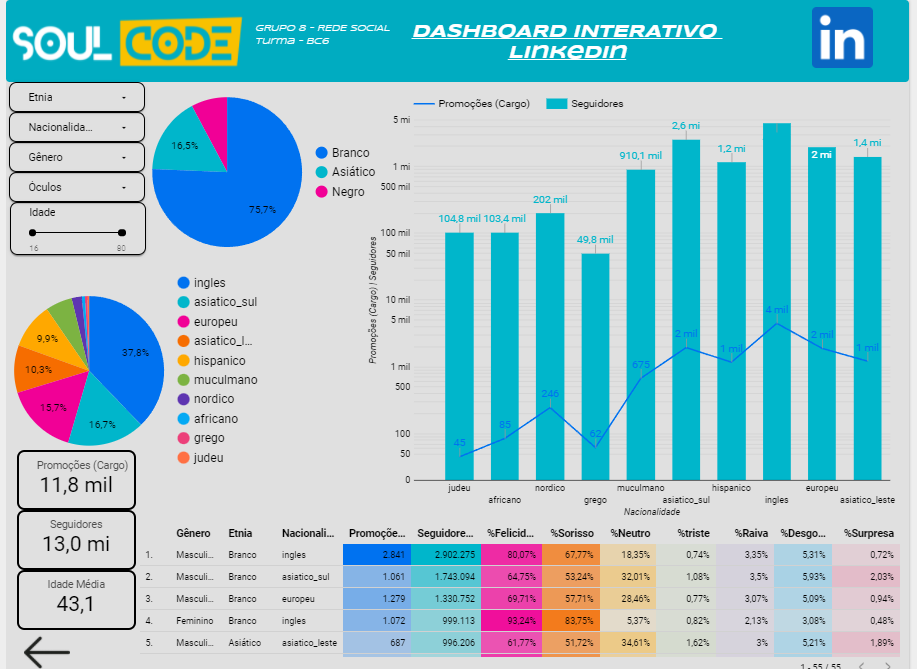


Link do notebook: <https://colab.research.google.com/drive/1YYSgxd9lrnmi0e8fQurUsb-eu9LMmdN1?usp=sharing>

**DATA STUDIO - LINKEDIN**

O Google Data Studio é uma ferramenta online de Bussiness Inteligence, como já explicado neste mesmo tópico do dataset youtube.

Essa sendo a última etapa do projeto, serão mostradas abaixo graficamente as informações derivadas dos insights gerados. Pode acessar o dashboard clicando [aqui](https://datastudio.google.com/s/rZs7mc9cf6Y).



**Alguns insights extraídos:**

* Homens, brancos de nacionalidade Inglesa lideram o primeiro lugar em número de promoções recebidas e de seguidores, seguido de homens, brancos de nacionalidade Ásia Meridional, homens, brancos de nacionalidade europeia, mulheres, brancas de nacionalidade inglesa e Homens, brancos de nacionalidade da Ásia ocidental.
* No contexto geral, a etnia branca domina tem os maiores números de seguidores e promoções.
* Proporcionalmente, as etnias que mais estão registradas nesse banco de dados são branca (75,7%), asiática (16,5%) e negro (7,8%).
* As 5 maiores nacionalidades por quantidade de usuário, respectivamente Ingles, Ásia Meridional, Europeu, Ásia Oriental e Hispanico.
* A média de idade dos usuários é de 43 anos.
* O grupo de usuários que mais tem foto de perfil com característica de felicidade são mulheres, asiáticas e gregas.
* Já as que tiram foto mais sorrindo são mulheres, de etnia negra e muçulmano.
* Fotos com mais raiva são homens, asiáticos judeus.

Analisando bem todas as informações desse banco de dados, foi possível verificar que não há relação com as emoções transmitidas nas fotos com o número de seguidores e promoções. Não foi possível analisar, por exemplo, se uma foto de perfil passando mais alegria tem mais chances de ser chamada para uma entrevista de emprego, tendo em vista que não há esse tempo de informação na base de dados ou outras informações de interação que há dentro do LinkedIn.

1. **CONTATOS**





**Nome: André Victor Moreira Costa  
E-mail:** [**andrevictorm2017@gmail.com**](mailto:andrevictorm2017@gmail.com)

**LinkedIn: andre-victor-moreira-costa**





**Victor Gonçalves**

**E-mail: victor.og17@gmail.com**

**LinkedIn: victor-de-oliveira-goncalves**

****

****

**Eduardo Mathias**

**E-mail: eduardo.mathiass09@gmail.com**

**LinkedIn: eduardo-mathias**



****

**Thiago Regis**

**E-mail: promothiagor@gmail.com**

**LinkedIn: thiagoregis1**