SOUL CODE

Projeto Final

Tema: Redes Sociais

SOUL CODE

Integrantes

André Victor

Eduardo Mathias

Thiago Regis

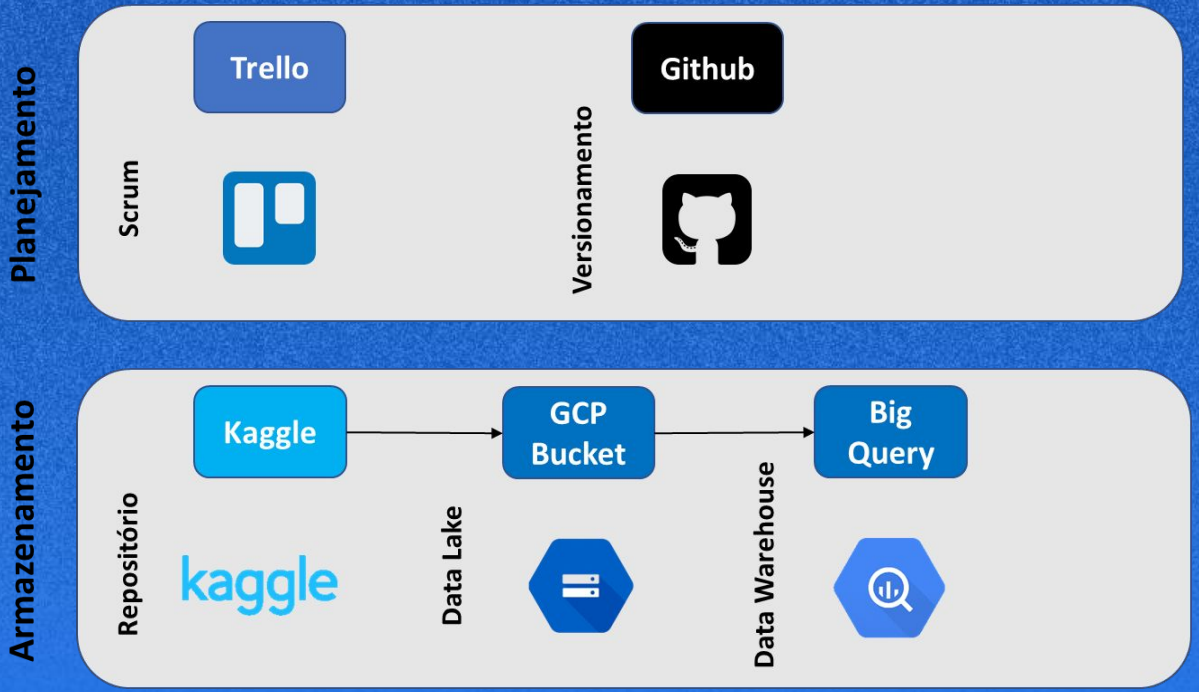
Victor Gonçalves

Objetivo

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

Motivação

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.


Planejamento
Scrum
Trello
Versionamento
Github
Armazenamento
Repositório
Kaggle
kaggle
Data Lake
GCP Bucket
Data Warehouse
Big Query

SOUL CODE

Transformação

Python → Pandas → Pandera → PySpark

Linguagem: Python

Tratamento: Pandas

Validação: Pandera

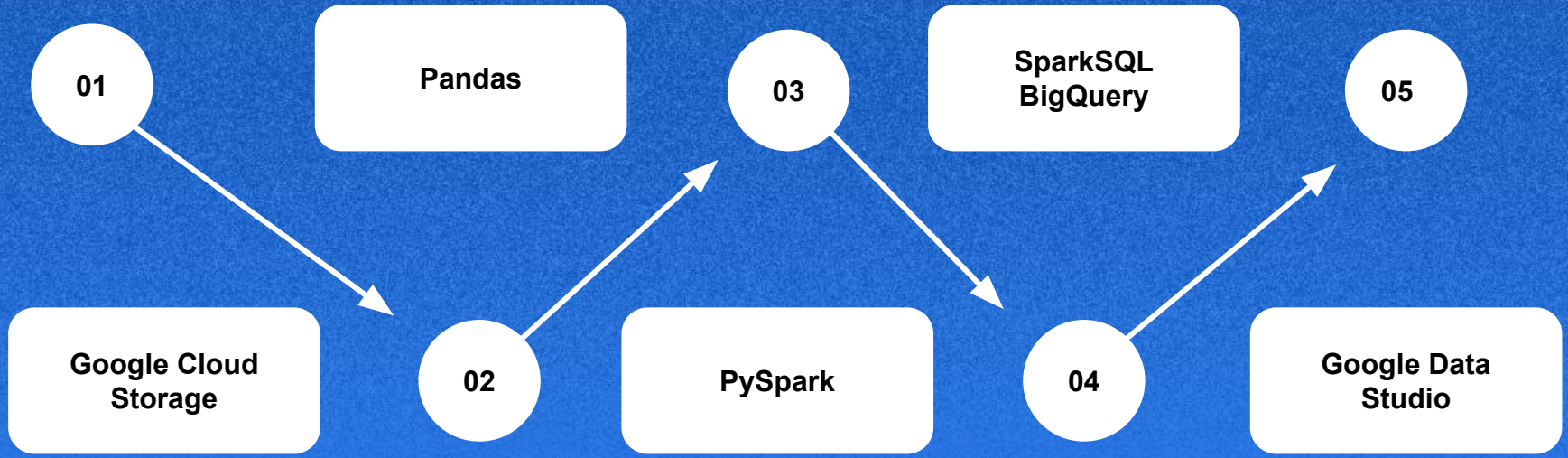
Manipulação: PySpark

Visualização

Dashboard: Data Studio

SOUL CODE

Fluxograma



Extração

Datasets retirados do Kaggle

SOUL CODE

01

Google Cloud Storage

02

Pandas

PySpark

03

SparkSQL
BigQuery

04

05

Google Data Studio

Google Cloud Storage

Data Lake

Buckets > projetofinalgrupo8 > entrada

UPLOAD FILES | UPLOAD FOLDER | CREATE FOLDER | MANAGE HOLDS | DOWNLOAD | DELETE

Filter by name prefix only | Filter | Filter objects and folders

Name	Size	Type	Created	Storage class
BR_category_id.json	9.9 KB	application/json	Nov 17, 2...	Standard
BR_youtube_trending_data.csv	122.1 MB	text/csv	Nov 17, 2...	Standard
CA_category_id.json	9.9 KB	application/json	Nov 17, 2...	Standard
CA_youtube_trending_data.csv	133.8 MB	text/csv	Nov 17, 2...	Standard
DE_category_id.json	9.9 KB	application/json	Nov 17, 2...	Standard
DE_youtube_trending_data.csv	149.9 MB	text/csv	Nov 17, 2...	Standard
FR_category_id.json	9.9 KB	application/json	Nov 17, 2...	Standard
FR_youtube_trending_data.csv	123.5 MB	text/csv	Nov 17, 2...	Standard

Buckets > projetofinalgrupo8 > saida

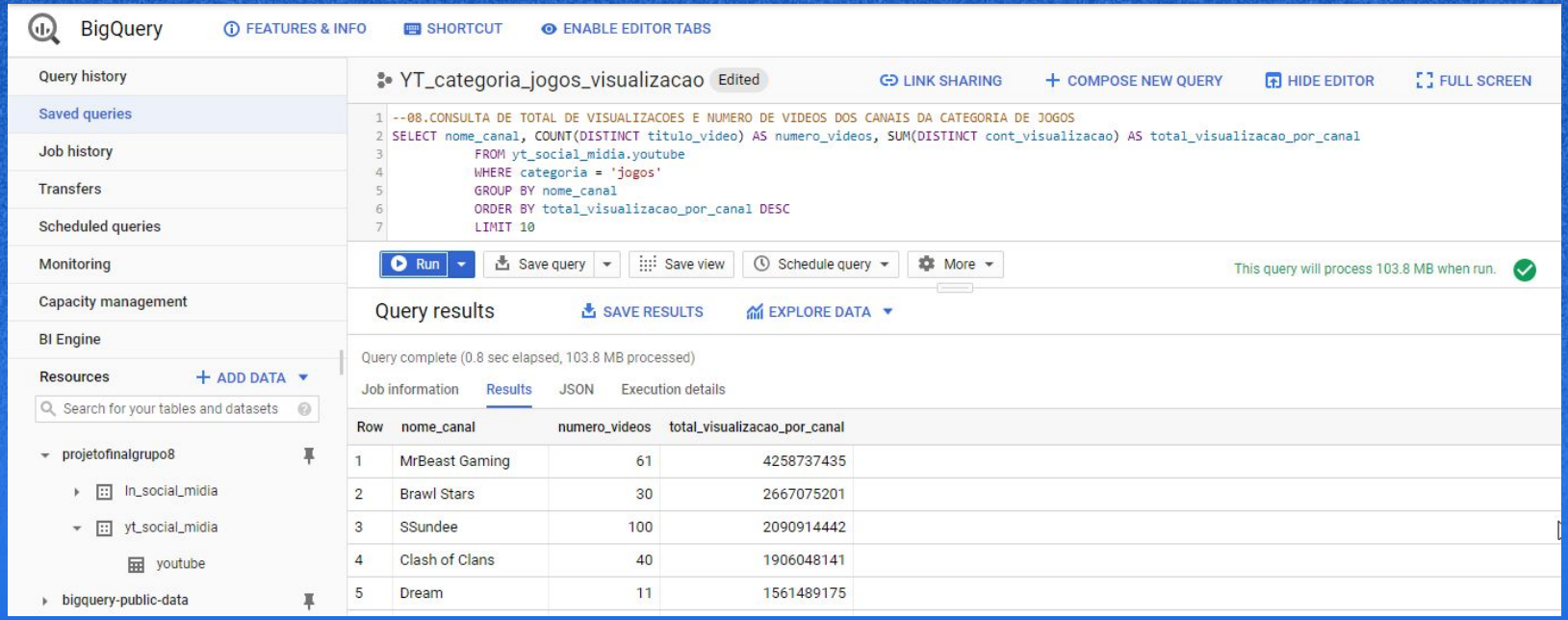
UPLOAD FILES | UPLOAD FOLDER | CREATE FOLDER | MANAGE HOLDS | DOWNLOAD | DELE

Filter by name prefix only | Filter | Filter objects and folders

Name	Size	Type	Created	Storage class
data_tratado_pyspark.csv	201.4 MB	text/csv	Nov 22, 2...	Standard
linkedin_tratado_pandas.csv	1.5 MB	text/csv	Nov 25, 2...	Standard
linkedin_tratado_pyspark.csv	1.2 MB	text/csv	Nov 25, 2...	Standard
youtube_data_base.csv	205.5 MB	text/csv	Nov 23, 2...	Standard
youtube_tratado_pandas.csv	204.4 MB	text/csv	Nov 23, 2...	Standard
youtube_tratado_pyspark.csv	201.4 MB	text/csv	Nov 23, 2...	Standard

BigQuery

Data Warehouse



SOUL CODE

01

Pandas

03

SparkSQL
BigQuery

05

Google Cloud
Storage

02

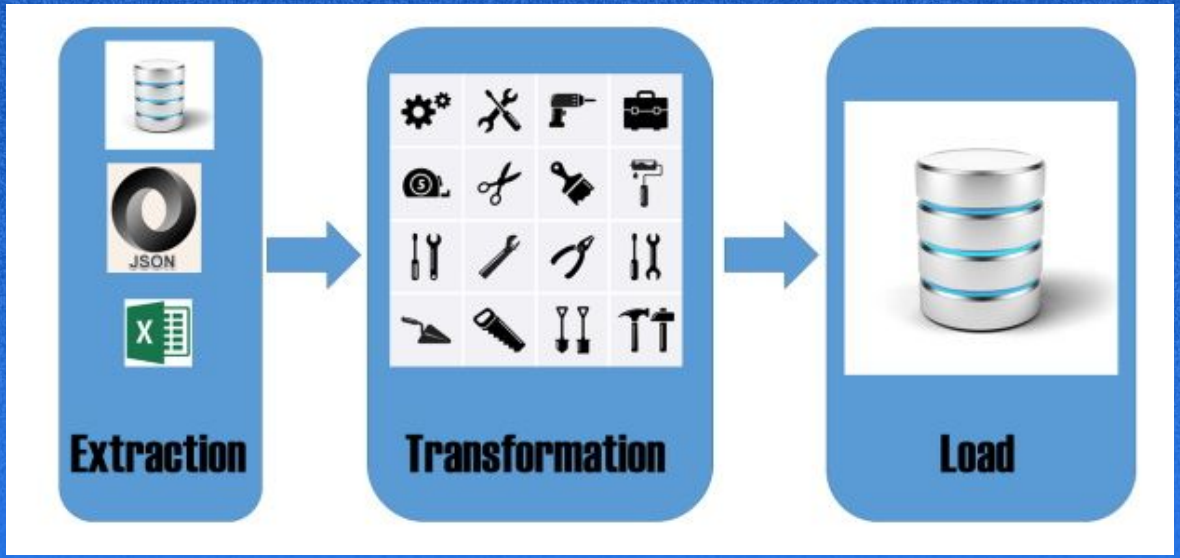
PySpark

04

Google Data
Studio

SOUL CODE

ETL



SOUL CODE

Descomprimindo

```
[ ]  1  dados_json['items'][0]
```

```
{'etag': 'IfWa37JGcqZs-jZeAyFGkbeh6bc',
 'id': '1',
 'kind': 'youtube#videoCategory',
 'snippet': {'assignable': True,
  'channelId': 'UCBR8-60-B28hp2BmDPdntcQ',
  'title': 'Film & Animation'}}
```

DESCOMPRIMINDO JSON

```
[ ]  1  new_dados = []
     2
     3  for i in range(len(dados_json['items'])):
     4    #print(dados_json['items'][i])
     5    new_data = {}
     6    new_data['json_kind'] = dados_json['items'][i]['kind']
     7    new_data['json_etag'] = dados_json['items'][i]['etag']
     8    new_data['categoryId'] = dados_json['items'][i]['id']
     9    new_data['json_title'] = dados_json['items'][i]['snippet']['title']
    10    new_data['json_assignable'] = dados_json['items'][i]['snippet']['assignable']
    11    new_data['json_channelId'] = dados_json['items'][i]['snippet']['channelId']
    12    new_dados.append(new_data)
```

SOUL CODE

Unificando

```
[ ]  1  dados_br_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/BR_youtube_trending_data.csv')
     2  dados_ca_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/CA_youtube_trending_data.csv')
     3  dados_de_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/DE_youtube_trending_data.csv')
     4  dados_fr_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/FR_youtube_trending_data.csv')
     5  dados_gb_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/GB_youtube_trending_data.csv')
     6  dados_in_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/IN_youtube_trending_data.csv')
     7  dados_jp_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/JP_youtube_trending_data.csv')
     8  dados_kr_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/KR_youtube_trending_data.csv')
     9  dados_mx_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/MX_youtube_trending_data.csv')
    10  dados_us_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/US_youtube_trending_data.csv')
    11  dados_ru_csv = pd.read_csv('gs://projetofinalgrupo8/entrada/RU_youtube_trending_data.csv')
```

```
[ ]  1  dados_br_csv.shape
```

```
[1]  1  dataframes = [dados_br_csv, dados_ca_csv, dados_de_csv, dados_fr_csv, dados_gb_csv, dados_in_csv, dados_jp_csv,
     2                dados_kr_csv, dados_mx_csv, dados_us_csv, dados_ru_csv]
     3  dados_world_csv = pd.concat(dataframes)
```

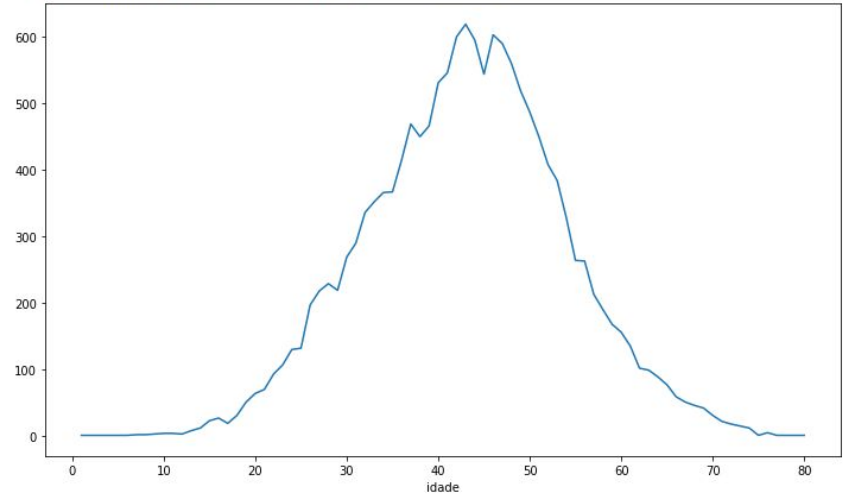
```
[ ]  1  data_raw = pd.merge(dados_world_csv, new_df, on=['categoryId'], how='left')
```

SOUL CODE

Analisando

```
[42]  1  df2.groupby(["idade"]).idade.count().plot.line(figsize=(12,7))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe3b4d7b3d0>
```



SOUL CODE

Colunas e dados inválidos

```
[ ]  1  df2 = df2.drop(['c_id','m_urn','avg_current_position_length','no_of_promotions','head_pitch','head_roll','head_yaw',
     2                  'mouth_close', 'mouth_mask', 'mouth_open','mouth_other','beauty',
     3                  'skin_acne', 'skin_dark_circle', 'skin_health','skin_stain',
     4                  'african', 'celtic_english','east_asian', 'european', 'greek',
     5                  'hispanic', 'jewish', 'muslim','nordic', 'south_asian','beauty_female',
     6                  'beauty_male','avg_previous_position_length'], axis=1)
```

```
[ ]  1  df2 = df2.drop_duplicates(subset=['m_urn_id'])
```

SOUL CODE

Renomeando

Renomeando colunas e dados Categoricos

Criando listas para renomear colunas

```
[ ]  1  col_old = ['video_id', 'title', 'publishedAt', 'channelId', 'channelTitle',
     2             'trending_date', 'view_count', 'likes', 'dislikes', 'comment_count',
     3             'comments_disabled', 'ratings_disabled', 'country', 'json_title']
     4
     5  col_new = ['id_video', 'titulo_video', 'publicado_em', 'id_canal', 'nome_canal',
     6             'data_destaque', 'cont_visualizacao', 'curtidas', 'nao_curtidas', 'cont_comentarios',
     7             'comentarios_desabilitados', 'curtidas_desabilitadas', 'pais', 'categoria']
```

```
[ ]  1  df2.columns = col_new
```

```
[ ]  1  df2.head(2)
```

	id_video	titulo_video	publicado_em	id_canal	nome_canal	data_destaque	cont_visualizacao	curtidas	nao_curtidas	cont_comentarios	c
0	s9FH4rDMvds	LEVEI UM FORA? FINGI ESTAR APAIXONADO	2020-08-11T22:21:49Z	UCGfBwrCoi9ZJjKiUK8MmJNw	Pietro Guedes	2020-08-12T00:00:00Z	263835	85095	487	4500	

SOUL CODE

Traduzindo

```
[32]  1  generos_antigos = ['Male','Female']
      2  generos_novos = ['Masculino','Feminino']

[33]  1  df2['genero'] = df2['genero'].replace(generos_antigos, generos_novos)
```

SOUL CODE

Validando

```
schema = pa.DataFrameSchema(
    columns = {
        "id_video":pa.Column(pa.String),
        "titulo_video":pa.Column(pa.String),
        "publicado_em":pa.Column(pa.DateTime),
        "id_canal":pa.Column(pa.String),
        "nome_canal":pa.Column(pa.String),
        "data_destaque":pa.Column(pa.DateTime),
        "cont_visualizacao":pa.Column(pa.Int),
        "curtidas":pa.Column(pa.Int),
        "nao_curtidas":pa.Column(pa.Int),
        "cont_comentarios":pa.Column(pa.Int),
        "comentarios_desabilitados":pa.Column(pa.Bool),
        "curtidas_desabilitadas":pa.Column(pa.Bool),
        "pais":pa.Column(pa.String),
        "categoria":pa.Column(pa.String)
    }
)
```

Exportando

```
[ ]
from google.cloud import storage
import os
serviceAccount = '/content/projetofinalgrupo8-2dcd866c3f46.json'
os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = serviceAccount


client = storage.Client()
bucket = client.get_bucket('projetofinalgrupo8')

bucket.blob('saida/youtube_tratado_pandas.csv').upload_from_string(df2.to_csv(index=False), 'text/csv')
```

SOUL CODE

01

Pandas

03

SparkSQL
BigQuery

05

Google Cloud
Storage

02

PySpark

04

Google Data
Studio

Validando

```
customSchema = StructType([
    StructField("id_video", StringType(),True),
    StructField("titulo_video", StringType(),True),
    StructField("publicado_em", StringType(),True),
    StructField("id_canal", StringType(),True),
    StructField("nome_canal", StringType(),True),
    StructField("data_destaque", StringType(),True),
    StructField("cont_visualizacao", IntegerType(),True),
    StructField("curtidas", IntegerType(),True),
    StructField("nao_curtidas", IntegerType(),True),
    StructField("cont_comentarios", IntegerType(),True),
    StructField("comentarios_desabilitados", StringType(),True),
    StructField("curtidas_desabilitadas", StringType(),True),
    StructField("pais", StringType(),True),
    StructField("categoria", StringType(),True)

])

schema = customSchema
```

```
dfspark = pd.read_csv('gs://projetofinalgrupo8/saida/youtube_tratado_pandas.csv', sep=',',encoding='UTF-8', header=0)
df= spark.createDataFrame(dfspark,schema=schema)
```

SOUL CODE

Arredondando

```
df = (df.withColumn("tempo_cargo_anterior",F.round(F.col("tempo_cargo_anterior"),2))
        .withColumn("desfoque",F.round(F.col("desfoque"),2))
        .withColumn("raiva",F.round(F.col("raiva"),2))
        .withColumn("felicidade",F.round(F.col("felicidade"),2))
        .withColumn("neutro",F.round(F.col("neutro"),2))
        .withColumn("triste",F.round(F.col("triste"),2))
        .withColumn("surpresa",F.round(F.col("surpresa"),2))
        .withColumn("sorriso",F.round(F.col("sorriso"),2))
        .withColumn("qualidade_imagem",F.round(F.col("qualidade_imagem"),2))
)
```

SOUL CODE

Datas

```
[17]  1  df = df.withColumn("publicado_em_data", F.col("publicado_em_data").cast("date"))
      2  df = df.withColumn("data_destaque", F.col("data_destaque").cast("date"))
      3
      4  df.select(F.col("publicado_em_data"), F.col("data_destaque")).show(5)
      5  df.printSchema()
```

```
+-----------------+-------------+
|publicado_em_data|data_destaque|
+-----------------+-------------+
|       2020-08-11|   2020-08-12|
|       2020-08-11|   2020-08-12|
|       2020-08-10|   2020-08-12|
|       2020-08-11|   2020-08-12|
|       2020-08-11|   2020-08-12|
+-----------------+-------------+
only showing top 5 rows

root
 |-- id_video: string (nullable = true)
 |-- titulo_video: string (nullable = true)
 |-- id_canal: string (nullable = true)
 |-- nome_canal: string (nullable = true)
 |-- cont_visualizacao: integer (nullable = true)
 |-- curtidas: integer (nullable = true)
 |-- nao_curtidas: integer (nullable = true)
 |-- cont_comentarios: integer (nullable = true)
 |-- comentarios_desabilitados: string (nullable = true)
 |-- curtidas_desabilitadas: string (nullable = true)
 |-- pais: string (nullable = true)
 |-- categoria: string (nullable = true)
 |-- publicado_em_data: date (nullable = true)
 |-- data_destaque: date (nullable = true)
```

SOUL CODE

01

Pandas

03

SparkSQL
BigQuery

05

Google Cloud
Storage

02

PySpark

04

Google Data
Studio

Consultas SQL BigQuery

- Apresentação Organização Google Cloud Storage
- BigQuery:
  - Query’s salvas
  - Consultas SQL

SOUL CODE

01

Pandas

03

SparkSQL
BigQuery

05

Google Cloud
Storage

02

PySpark

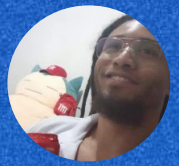
04

Google Data
Studio

SOUL CODE

Dúvidas?

Contatos



André Victor
E-mail: andrevictorm2017@gmail.com
linkedin: andre-victor-moreira-costa

Eduardo Mathias
E-mail: eduardo.mathiass09@gmail.com
Linkedin: eduardo-mathias

Thiago Regis
E-mail: promothiagor@gmail.com
linkedin: thiagoregis1

Victor Gonçalves
E-mail: victor.og17@gmail.com
Linkedin: victor-de-oliveira-goncalves

SOUL CODE

Muito Obrigado!