Projeto de Engenharia e Análise de Dados

Aylton Neto - Squad 1 Compass.UOL

Resumo



- Projeto Divido em 4 Partes (Sprints)
 - Parte 1 Carregar arquivos CSV
 - Parte 2 Definição do Projeto e Ingestão de Dados
 - Parte 3 Processamento e Modelagem
 - Parte 4 Consumo e Apresentação
- Sobre o Programa de Bolsas
 - Aprendizado na Compass
 - Agregando aos Clientes

Projeto - Parte 1



Carregar Arquivos na RAW

- Script Python e Dockerfile
 - Carregar Dados para o Bucket

```
Dockerfile ×

Compass.uol > Sprint07 > .DataAnalytics07 > DesafioETL > Dockerfile

PROM python:3.10-slim

COPY . .

WORKDIR /app

RUN pip install boto3

CMD python desafio_ETL.py
```

Projeto - Parte 2

⊌compass.uol

Definição do Projeto

- Squad 1 Animação/Comédia
- Tema: Avaliar o Cenário das Séries de Animação com Boa Avaliação
 - Em qual período foram lançadas mais séries boas?
 - Séries lançadas em períodos anteriores possuem relevância atualmente?
 - Existe predominância de algum país?
 - Existem padrões relacionados a duração?
 - Existem padrões relacionados ao nº de episódios?
 - Quais emissoras e estúdios se destacam?
 - Quais são as melhores animações?



Definição do Projeto

- Definir Amostra
 - Métricas

Resultados:

```
numeroVotos
                                                 notaMedia
summary
                                 summary
                      7764
                                                       7764
  count
                                   count
  mean 2047.3037094281299
                                          6.977807831014931
 stddev | 14651.769885249427 |
                                  stddev | 1.1028955069267738 |
    min
                                     min
                                                         1.0
                    518958
    max
                                     max
```



```
# Trecho 1 - Bibliotecas e Chaves de Acesso
import json
import requests
import pandas as pd
import boto3
def lambda_handler(event, context):
    user = boto3.client(
    service name='s3',
    aws_access_key_id=
    aws_secret_access_key= 🦳
    region name='us-east-1'
    api key =
```



```
# Trecho 2 - Função para obter o TMDB(ID) pelo IMDB(ID) e extrair dados da API, caso encontre algum resultado.

def get_data_TMDB(id_IMDB, api_key):
    url = f'https://api.themoviedb.org/3/find/{id_IMDB}?api_key={api_key}&external_source=imdb_id'
    response1 = requests.get(url)
    data1 = response1.json()

if 'tv_results' in data1 and len(data1['tv_results']) > 0:
    id_TMDB = data1['tv_results'][0]['id']
    url = f'https://api.themoviedb.org/3/tv/{id_TMDB}?api_key={api_key}'
    response2 = requests.get(url)
    data2 = response2.json()
    return data2
    return None
```



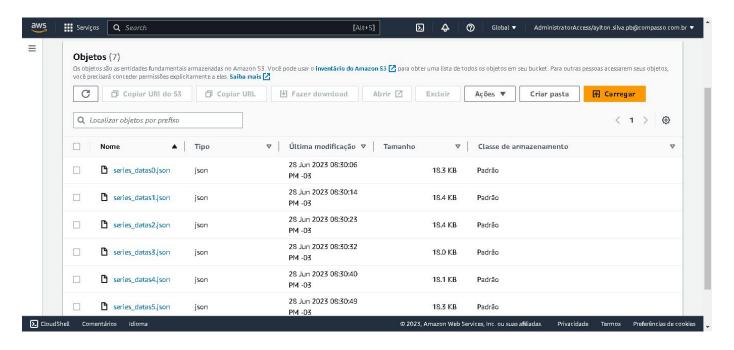
```
# Trecho 4 - Looping para fazer as requisições na API com base nos IDs e nos campos determinados
   series informations = []
   count, num json = 0, 0
   for index, row in enumerate(df.iterrows()):
       id IMDB = str(row[1]['id'])
       series details = get_data_TMDB(id_IMDB, api_key)
       if series details != None:
           title = series details['name']
            origin = series details['origin country'][0] if series details['origin country'] else None
           company = series details['production companies'][0]['name'] if series details['production companies'] else None
           network = series details['networks'][0]['name'] if series details['networks'] else None
            production = series details['in production']
            episodes = series details['number of episodes']
           seasons = series details['number of seasons']
            dic details = {
                'id': id IMDB,
                'titulo': title.
                'pais': origin,
               'estudio': company,
                'emissora': network.
                'mediaEpisodios': episodes//seasons,
                'emProdução': production
           series informations.append(dic details)
            count += 1
```



```
# Trecho 5 - Condição para que um novo arquivo seja criado a cada 100 séries ou quando chegar ao fim da seleção
            if count % 100 == 0 or index == len(df)-1:
                json_informations = json.dumps(series_informations)
                user.put_object(Body=json_informations,
                                Bucket='data-lake-do-aylton',
                                Key=f'Raw/TMDB/JSON/2023/06/28/series_datas{num_json}.json')
                series_informations = []
                num json += 1
            continue
   return {
        "statusCode": 200.
        "body": "Arquivo JSON criado com sucesso."
```



Arquivos JSON no Bucket





Resultados

- Após a filtragem pelo gênero "Animation": 7764 Animações
- Médias: numeroVotos = 2050 e notaMedia = 7.0
- Após a extração no TMDB: 628 Animações
- 7 Arquivos JSON no Bucket

Projeto - Parte 3



1ª JOB GLUE - Arquivos CSV

```
csv_path = "s3://data-lake-do-aylton/Raw/Local/CSV/Series/2023/05/29/series.csv"

csv_data = spark.read.format("csv").option("header", "true").option("delimiter", "|").load(csv_path)

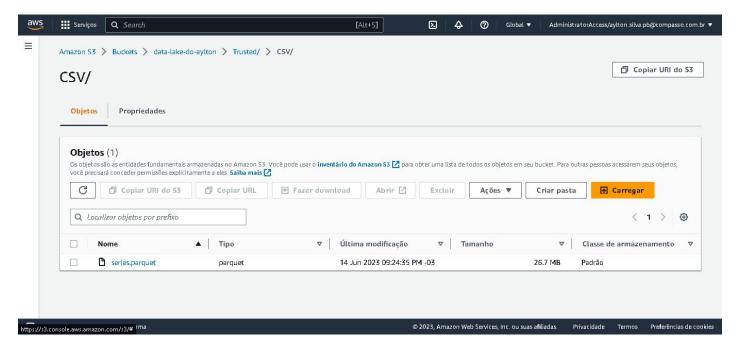
csv_data = csv_data.repartition(1)

csv_data.write.mode("append").parquet("s3://data-lake-do-aylton/Trusted/CSV/")

job.commit()
```



1ª JOB GLUE - Arquivos CSV



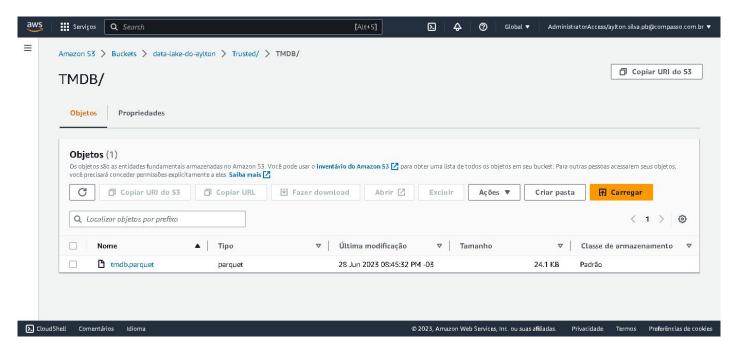


2ª JOB GLUE - Arquivos JSON

```
json paths = [
    "s3://data-lake-do-aylton/Raw/TMDB/JSON/2023/06/28/series datas0.json",
    "s3://data-lake-do-aylton/Raw/TMDB/JSON/2023/06/28/series datas1.json",
    "s3://data-lake-do-aylton/Raw/TMDB/JSON/2023/06/28/series datas2.json",
    "s3://data-lake-do-aylton/Raw/TMDB/JSON/2023/06/28/series datas3.json",
    "s3://data-lake-do-aylton/Raw/TMDB/JSON/2023/06/28/series datas4.json",
    "s3://data-lake-do-aylton/Raw/TMDB/JSON/2023/06/28/series datas5.json",
    "s3://data-lake-do-aylton/Raw/TMDB/JSON/2023/06/28/series datas6.json"
json data = spark.read.json(json paths)
json data = json data.repartition(1)
json_data.write.mode("append").parquet("s3://data-lake-do-aylton/Trusted/TMDB/")
job.commit()
```



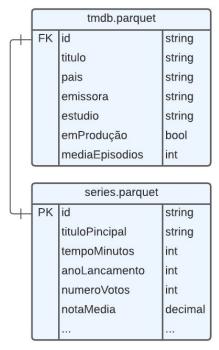
2ª JOB GLUE - Arquivos JSON

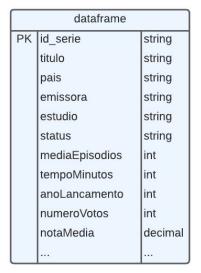




Modelagem de Dados

Modelagem:



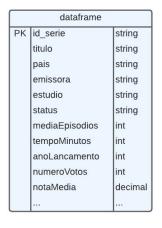


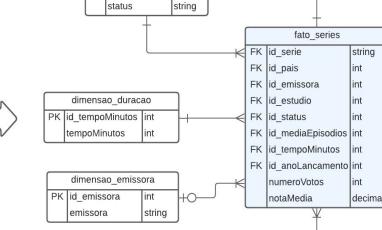
Join



Modelagem de Dados

Modelagem:

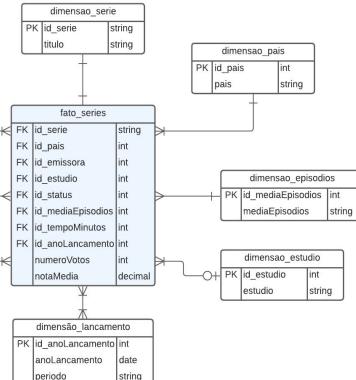




dimensao status

lint

PK id status





```
# 1º Etapa | Bibliotecas e Trecho Padrão do GLUE
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from pyspark.sql.window import Window
import pyspark.sql.functions as func
## @params: [JOB NAME]
args = getResolvedOptions(sys.argv, ['JOB NAME'])
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark session
job = Job(glueContext)
job.init(args['JOB NAME'], args)
```





```
# 5ª Etapa | Criação das Dimensões
# Dimensão Serie
df serie = df fato.select('id serie', 'titulo')
# Dimensão Lancamento
df lancamento = df fato.select('anoLancamento').distinct().orderBy('anoLancamento')
df lancamento = df lancamento.withColumn('id anoLancamento', func.row number().over(Window.orderBy('anoLancamento')))
df lancamento = df lancamento.select('id anoLancamento', 'anoLancamento')
df lancamento = df lancamento.withColumn("periodo", func.when(func.col("anoLancamento").between(1920, 1960), "Era de Ouro Americana")
                                                        .when(func.col("anoLancamento").between(1960, 1980), "Influência dos Animes")
                                                        .when(func.col("anoLancamento").between(1980, 2000), "TV por Assinatura")
                                                        .when(func.col("anoLancamento").between(2000, 2010), "Popularidade dos DVDs")
                                                        .when(func.col("anoLancamento").between(2010, 9999), "Serviços de Streaming"))
# Dimensão Duração
df duracao = df fato.withColumn('tempoMinutos', func.col('tempoMinutos').cast('integer'))
df duracao = df duracao.select('tempoMinutos').distinct().orderBy('tempoMinutos').filter(df_duracao['tempoMinutos'].isNotNull())
df duracao = df duracao.withColumn('id tempoMinutos', func.row number().over(Window.orderBy('tempoMinutos')))
df duracao = df duracao.select('id tempoMinutos', 'tempoMinutos')
```

```
# Dimensão Episodios
df episodios = df fato.select('mediaEpisodios').distinct().orderBy('mediaEpisodios')
df episodios = df episodios.withColumn('id mediaEpisodios', func.row number().over(Window.orderBy('mediaEpisodios')))
df episodios = df episodios.select('id mediaEpisodios', 'mediaEpisodios')
# Dimensão País
df pais = df fato.select('pais').distinct().orderBy('pais').filter(df fato['pais'].isNotNull())
df_pais = df_pais.withColumn('id_pais', func.row_number().over(Window.orderBy('pais')))
df pais = df pais.select('id pais', 'pais')
# Dimensão Status
df status = df fato.select('status').distinct()
df_status = df_status.withColumn('id_status', func.row_number().over(Window.orderBy('status')))
df status = df status.select('id status', 'status')
# Dimensão Estúdio
df estudio = df fato.select('estudio').distinct().orderBy('estudio').filter(df fato['estudio'].isNotNull())
df estudio = df estudio.withColumn('id estudio', func.row number().over(Window.orderBy('estudio')))
df estudio = df estudio.select('id estudio','estudio')
# Dimensão Emissora
df emissora = df fato.select('emissora').distinct().orderBy('emissora').filter(df fato['emissora'].isNotNull())
df emissora = df emissora.withColumn('id emissora', func.row number().over(Window.orderBy('emissora')))
df_emissora = df_emissora.select('id_emissora','emissora')
```





Explicando Código:

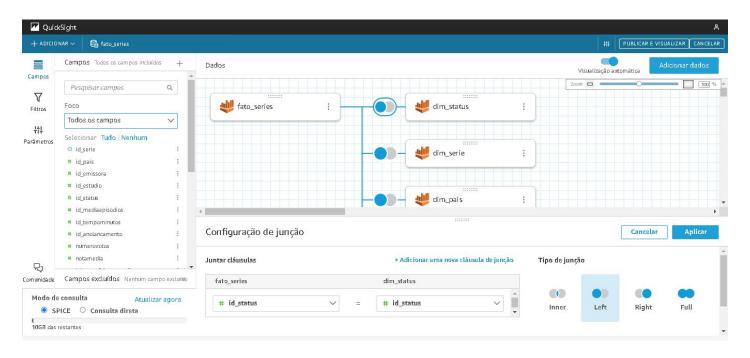
job.commit()

```
# 7 Etapa | Salvar os dataframes como tabelas em formato Parquet na Refined df_serie.repartition(1).write.saveAsTable@name='projeto_animacoes.dim_serie',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Serie',format='parquet' df_lancamento.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_lancamento',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Duracao',format='parquet') df_duracao.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_duracao',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Episodios',format='parquet') df_pais.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_pais',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Pais',format='parquet') df_status.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_status',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Status',format='parquet') df_emissora.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_emissora',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Emissora',format='parquet') df_estudio.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_estudio',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Estudio',format='parquet') df_fato.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_estudio',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Estudio',format='parquet') df_fato.repartition(1).write.saveAsTable(name='projeto_animacoes.dim_estudio',mode='overwrite',path='s3://data-lake-do-aylton/Refined/Estudio',format='parquet')
```

Projeto - Parte 4

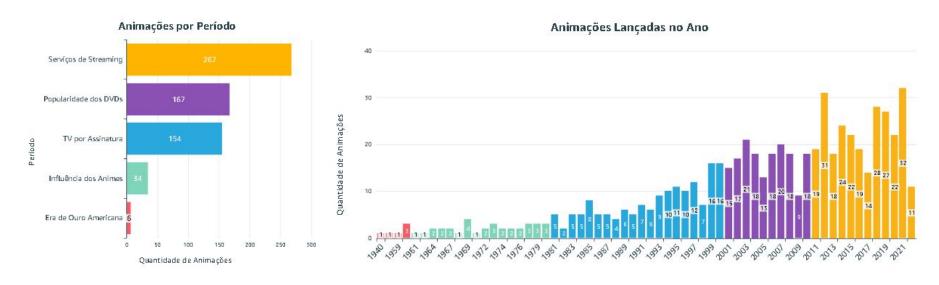


Acessar dados no QuickSight



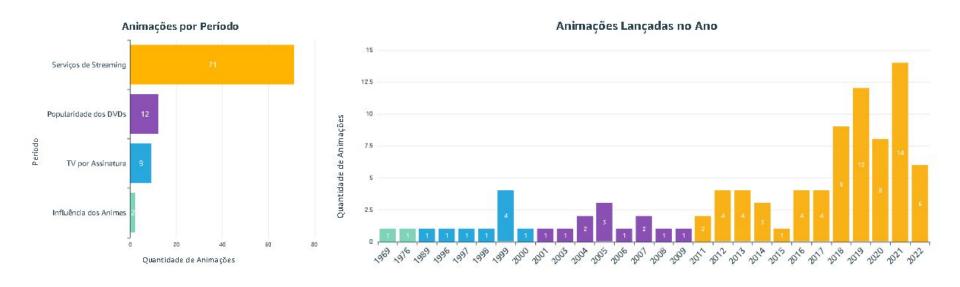


- Qual foi o período em que foram lançadas mais séries boas?
 - Esperado: Maior quantidade de séries boas lançadas no último período





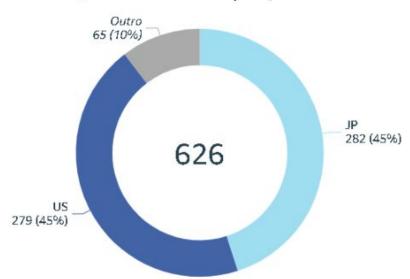
- Séries lançadas em períodos anteriores possuem relevância atualmente?
 - Esperado: Quantidade pequena de séries antigas (pouca relevância)





- Existe predominância de algum país?
 - Esperado: Animações Japonesas devido a popularidade dos Animes

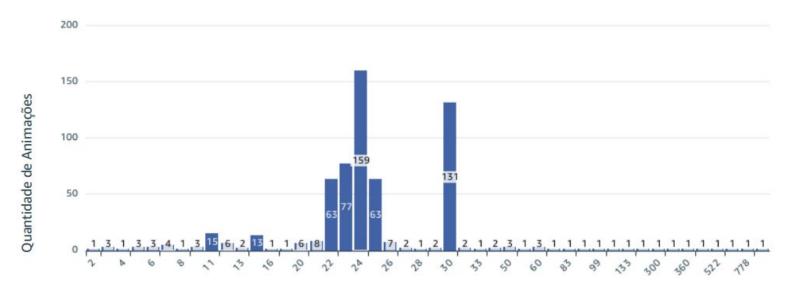
Quantidade de Animações por País





- Existem padrões relacionados a duração?
 - Esperado: 24mim de duração (Baseado nos Animes)

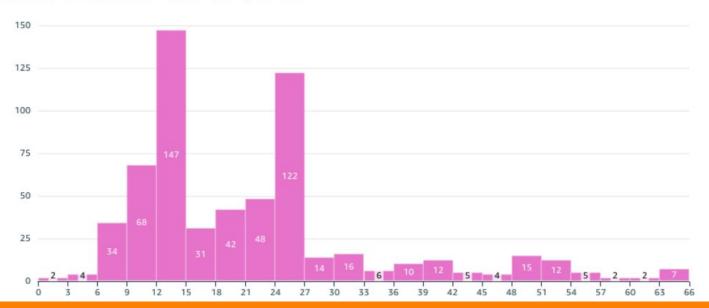
Frequencia de Duração das Animações





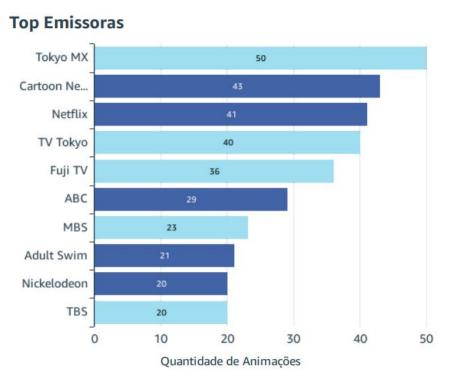
- Existem padrões relacionados ao nº de episódios?
 - Esperado: 12 episódios por temporada (Baseado nos Animes)

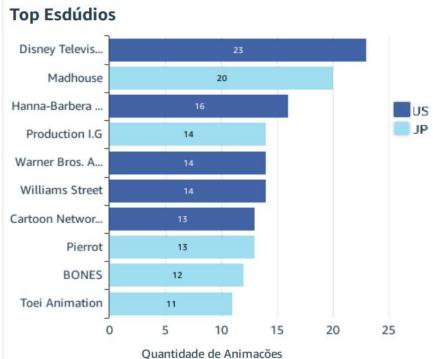
Média de Episódios por Temporada





Quais emissoras e estúdios se destacam?







Quais são as melhores animações de acordo com a análise?

Top Séries

País	Título	Emissora	Estúdio	Status	Ano	Nota	Numero de Votos	Duração
⊟ US	⊟ Arcane	─ Netflix	☐ Fortiche	Em Produção	2021	9,0	205,071	41
	☐ Avatar: The Last	⊟ Nickelodeon	□ Nickelodeon	Finalizado	2005	9,3	314,371	23
	☐ Batman: The	⊟ FOX	☐ DC Comics	Finalizado	1992	9,0	104,459	23
	☐ Rick and Morty	☐ Adult Swim	⊟ Williams Street	Em Produção	2013	9,1	518,958	23
⊟ JP	☐ Attack on Titan	⊟ MBS	☐ Production I.G	Em Produção	2013	9,0	371,218	24
	☐ Death Note	☐ Nippon TV	⊟ Madhouse	Finalizado	2006	9,0	323,143	24
	☐ Fullmetal	⊟ MBS	⊟ BONES	Finalizado	2009	9,1	167,814	24
	☐ Hunter x Hunter	☐ Nippon TV	⊟ Madhouse	Finalizado	2011	9,0	100,138	24
□ IR	☐ The Heroes	⊟ null	□ null	Em Produção	2008	9,4	165,984	30

Sobre o Programa de Bolsas















































- Não somente Hardskills!
 - Trabalho em Equipe
 - Equilíbrio Emocional
 - Pontualidade
 - Organização





Agregando aos Clientes

- Como agregar?
 - Análises
 - Arquiteturas
 - Soluções
 - o Insights



Muito Obrigado!