

## **MVP - Sprint: Engenharia de Dados**

*Esse documento será apresentado da seguinte forma:*

- 1- Figuras com representações Databricks para entendimento inicial;
- 2- Notebook master simulando pipeline;
  - a. OBJETIVO: ANÁLISE DE CONFLITOS SOCIAIS NOS ESTADOS BRASILEIROS ENTRE 2018-2023
  - b. BUSCA DOS DADOS (Kaggle)
  - c. COLETA E ARMAZENAMENTO DE DADOS
  - d. CARGA DE DADOS - ROW (CAMADA BRONZE)
  - e. TRANSFORMAÇÃO DOS DADOS (SILVER)
  - f. JUNTANDO INFORMAÇÕES DE CONFLITOS E CIDADES (GOLD)
    - i. CATÁLOGO DE DADOS
  - g. ANÁLISE DOS DADOS - SQL DASHBOARD
- 3- Notebooks periféricos representando camadas Bronze, Silver e Gold;
- 4- Notebook de Análise para resposta às 5 perguntas de negócio.
- 5- Análise e Autoavaliação.
- 6- O código construído está disponibilizado no repositório público do GitHub através do link [igorbrasil1978/engenharia\\_dados\\_mvp: Sprint 2 - Engenharia de Dados \(github.com\)](https://github.com/igorbrasil1978/engenharia_dados_mvp: Sprint 2 - Engenharia de Dados)

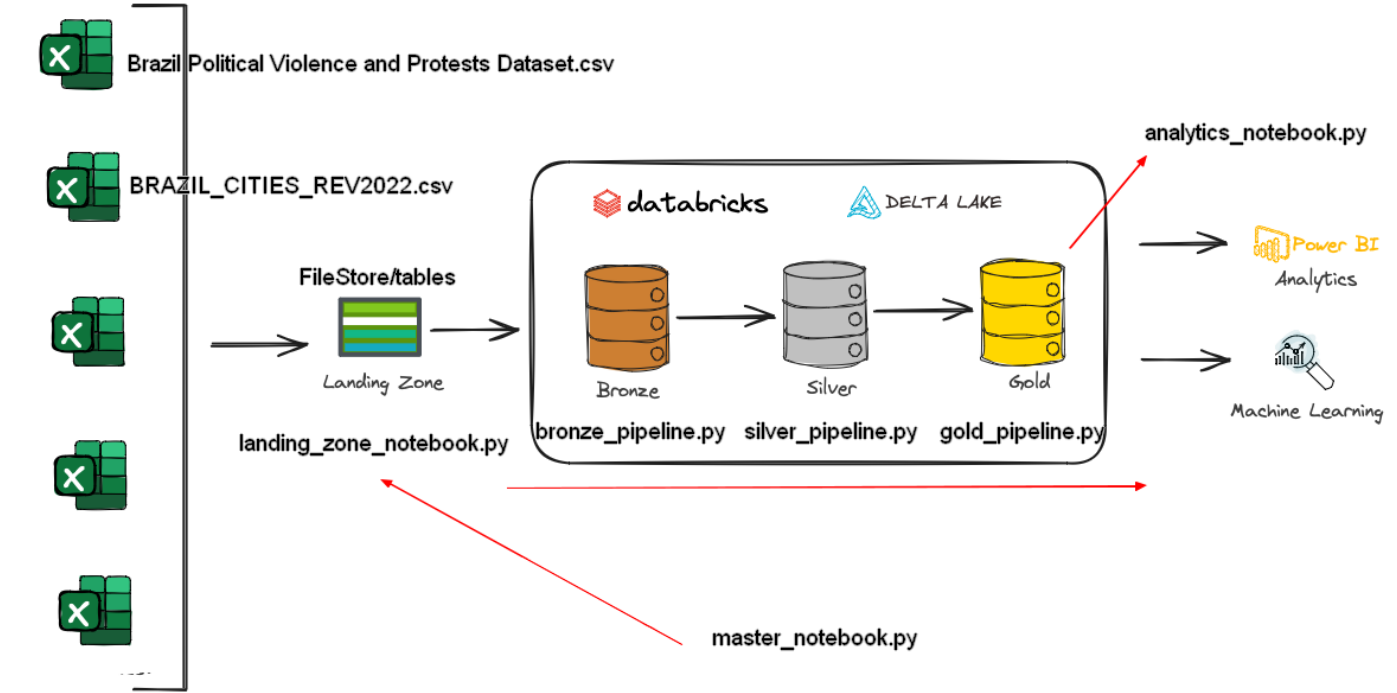
**Estudante:** IGOR PIROLA BRASIL

**Matrícula:** 4052024000351

**Curso:** CIÊNCIA DE DADOS E ANALYTICS

**Unidade:** PUC-RIO 100% ONLINE

PIPELINE



NOTEBOOKS (PASTAS)

+

Workspace

Inicio

Workspace

Shared

Users

notebook\_mvp\_engenharia\_dados

analytics

bronze

gold

landing\_zone

silver

Lixo

Workspace >

notebook\_mvp\_engenharia\_dados

Nome

Tipo

Proprietário

Criada às

analytics

bronze

gold

landing\_zone

silver

master\_notebook

Pasta

Pasta

Pasta

Pasta

Pasta

Notebook

Igor Brasil

Igor Brasil

Igor Brasil

Igor Brasil

Igor Brasil

Igor Brasil

2024-06-25 10:05:36

2024-06-25 09:54:40

2024-06-25 09:54:57

2024-06-25 10:05:08

2024-06-25 09:54:51

2024-06-20 15:38:07

CATÁLOGO DE DADOS (DATABASES)

Novo

Espaço de trabalho

Recentes

Pesquisar

Catálogo

Fluxos de trabalho

Compute

Machine Learning

Dados

Bases de dados

Filtrar bases de dados

bronze

default

gold

silver

Tabelas

Filtrar tabelas

cidade

conflito

Criar tabela

Interromper

Igor Brasil's Cluster (clo...)

Compartilhar

Publicar

- ROW (CAMADA BRONZE)

rasil.

rac

## 1- OBJETIVO: ANÁLISE DE CONFLITOS SOCIAIS NOS ESTADOS BRASILEIROS NO ANO DE 2018-2023

O objetivo desse Notebook é analisar um conjunto de dados de conflitos pacíficos e violentos no Brasil, a fim de responder algumas perguntas:

- 1- Quais as TOP 10 cidades mais Violentas do Brasil?
- 2- Que tipo de conflitos são mais comuns nessas cidades ?
- 3- Quais principais atores estão envolvidos nesses conflitos nessas cidades?
- 4- Quais as cidades mais letais do Brasil?
- 5- Qual ano obteve a maior fatalidade ?

## 2 - BUSCA DOS DADOS

CONFLITO (Brazil Political Violence and Protests Dataset.csv)

<https://www.kaggle.com/datasets/justin2028/brazil-conflict-tracker-20182023/data>

CIDADE (BRAZIL\_CITIES\_REV2022.csv)

<https://www.kaggle.com/datasets/crisparada/brazilian-cities>

<https://www.kaggle.com/datasets/justin2028/brazil-conflict-tracker-20182023/data>

<https://www.kaggle.com/datasets/crisparada/brazilian-cities>

## 3 - COLETA E ARMAZENAMENTO DE DADOS

- 1- Download dos arquivos CSV do kaggle e armazenamento no FileStore/tables

```
%run
./landing_zone/landing_zone_notebook
```

## landing\_zone\_notebook.py

Armazenamento dos arquivos CSV diretamente no FileStore/tables. Processo manual de carga de dados CSV

----- Fim Notebook landing\_zone\_notebook.py -----

## 4 - CARGA DE DADOS - ROW (CAMADA BRONZE)

- 1- Criação do DATABASE Bronze
- 2- Carga de dados dos CONFLITOS no Brasil.
- 3- Carga de dados das CIDADES Brasileiras.

```
%run
./bronze/bronze_pipeline
```

## bronze\_pipeline.py



bronze\_pipeline (Python)

Import Notebook

```
%sql
CREATE DATABASE IF NOT EXISTS bronze
```

OK

```
# 1 - Criação do DATABASE Bronze.
```

```
print("DATABASE BRONZE CRIADO COM SUCESSO!")
```

DATABASE BRONZE CRIADO COM SUCESSO!

```
# 2 - Carga de dados dos CONFLITOS no Brasil.
```

```
# File location and type
```

```
file_location = "/FileStore/tables/Brazil_Political_Violence_and_Protests_Dataset.csv"
```

```
file_type = "csv"
```

```
# CSV options
```

```
infer_schema = "false"
```

```
first_row_is_header = "true"
```

```
delimiter = ","
```

```
# The applied options are for CSV files. For other file types, these will be ignored.
```

```
df_conflicto = spark.read.format(file_type) \
```

```
    .option("inferSchema", infer_schema) \
```

```
    .option("header", first_row_is_header) \
```

```
    .option("sep", delimiter) \
```

```
    .load(file_location)
```

```
%sql
DROP TABLE IF EXISTS bronze.conflicto
```

OK

```
dbutils.fs.rm('dbfs:/user/hive/warehouse/bronze.db/conflicto',True)
```

```
df_conflicto.write.format("delta").mode("append").saveAsTable("bronze.conflicto")
```

```
print("CARGA DADOS DE CONFLITO CRIADO COM SUCESSO NA DATABASE BRONZE!")
```

CARGA DADOS DE CONFLITO CRIADO COM SUCESSO NA DATABASE BRONZE!

```
# 3- Carga de dados das CIDADES Brasileiras.
```

```
# File location and type
```

```
file_location = "/FileStore/tables/BRAZIL_CITIES_REV2022.CSV"
```

```
file_type = "CSV"
```

```
# CSV options
```

```
infer_schema = "false"
```

```
first_row_is_header = "true"
```

```
delimiter = ","
```

```
# The applied options are for CSV files. For other file types, these will be ignored.
```

```
df_cidade = spark.read.format(file_type) \
```

```
    .option("inferSchema", infer_schema) \
```

```
    .option("header", first_row_is_header) \
```

```
    .option("sep", delimiter) \
```

```
    .load(file_location)
```

```
%sql
DROP TABLE IF EXISTS bronze.cidade
```

OK

```
## primeiro tratamento de dados: colunas com espaço não são aceitas para criar tabela no DATABASE
df_cidade = df_cidade.withColumnRenamed('IDHM Ranking 2010', 'IDHM_Ranking_2010')

dbutils.fs.rm('dbfs:/user/hive/warehouse/bronze.db/cidade', True)

df_cidade.write.format("delta").mode("append").saveAsTable("bronze.cidade")

print("CARGA DADOS DE CIDADE CRIADO COM SUCESSO NA DATABASE BRONZE!")
```

CARGA DADOS DE CIDADE CRIADO COM SUCESSO NA DATABASE BRONZE!

```
display(df_conflito)

display(df_cidade)
```

Table					
	EVENT_DATE	EVENT_TYPE	SUB_EVENT_TYPE	ACTOR1	ACTOR2
1	01-January-2018	Protests	Peaceful protest	Protesters (Brazil)	null
2	01-January-2018	Protests	Peaceful protest	Protesters (Brazil)	null
3	01-January-2018	Violence against civilians	Attack	Unidentified Armed Group (Brazil)	Civilians (Brazil)
4	01-January-2018	Violence against civilians	Attack	Unidentified Gang and/or Police Militia	Civilians (Brazil)
5	01-January-2018	Violence against civilians	Attack	Unidentified Armed Group (Brazil)	Civilians (Brazil)
6	01-January-2018	Violence against civilians	Attack	Unidentified Gang and/or Police Militia	Civilians (Brazil)
7	01-January-2018	Battles	Armed clash	Unidentified Armed Group (Brazil)	Unidentified Armed Group (Brazil)
8	01-January-2018	Violence against civilians	Attack	Unidentified Armed Group (Brazil)	Civilians (Brazil)
9	01-January-2018	Violence against civilians	Attack	Unidentified Armed Group (Brazil)	Civilians (Brazil)

Table								
	CITY	STATE	CAPITAL	IBGE_RES_POP	IBGE_RES_POP_BRAS	IBGE_RES_POP_ESTR	IBGE_DU	IBGE_DU_URBA
1	Abadia De Goiás	GO	0	6876	6876	0	2137	1546
2	Abadia Dos Dourados	MG	0	6704	6704	0	2328	1481
3	Abadiânia	GO	0	15757	15609	148	4655	3233
4	Abaetetuba	PA	0	141100	141040	60	31061	19057
5	Abaeté	MG	0	22690	22690	0	7694	6667
6	Abaíara	CE	0	10496	10496	0	2791	1251
7	Abaré	BA	0	17064	17064	0	4332	2379
8	Abatiá	PR	0	7764	7764	0	2499	1877
9	Abaira	BA	0	8316	8316	0	2572	1193
10	Abdon Batista	SC	0	2653	2653	0	848	234
11	Abel Figueiredo	PA	0	6780	6780	0	1880	1650
12	Abelardo Luz	SC	0	17100	17084	16	4739	2694
13	Abre Campo	MG	0	13311	13294	17	3930	2202
14	Abreu E Lima	PE	0	94429	94407	22	28182	25944

3,570+ rows | Truncated data due to byte limit

----- Fim Notebook bronze\_pipeline.py -----

## 5 - TRANSFORMAÇÃO DOS DADOS (SILVER)

- 1- Tratamento do campo LOCATION na tabela Conflito - Removento acentos
- 2- Tratamento do campo EVENT\_DATE na tabela Conflito - Transformando em mês e ano
- 3- Tratamento do campo CITY na tabela Cidades - Removento acentos
- 4- Limpeza e transformação dos dados: tradução, eliminação de colunas, rename de colunas
- 5- Criação do DATABASE Silver e Carga das tabelas Conflito e Cidade

```
%run
./silver/silver_pipeline
```

```
from pyspark.sql.types import IntegerType
#from pyspark.sql.functions import translate, upper
from pyspark.sql.functions import *
from datetime import datetime
```

```
%sql
select upper(LOCATION), count(*) from bronze.conflicto where UPPER(LOCATION) like '%ZONE%' group by LOCATION order by LOCATION
```

Table



	upper(LOCATION)	count(1)
1	RIO DE JANEIRO - CENTRAL ZO...	926
2	RIO DE JANEIRO - NORTH ZONE	3038
3	RIO DE JANEIRO - SOUTH ZONE	590
4	RIO DE JANEIRO - WEST ZONE	1859
5	SAO PAULO - CENTRAL ZONE	476
6	SAO PAULO - EAST ZONE	271
7	SAO PAULO - NORTH ZONE	172
8	SAO PAULO - SOUTH ZONE	422
9	SAO PAULO - WEST ZONE	179

9 rows

```
## carregando os dados do DATABASE bronze para tratamento dos dados de Conflito
```

```
df_conflicto_bronze_sql = spark.sql('select * from bronze.conflicto')
```

```
print("TABELA CONFLITO CARREGADA COM SUCESSO DO DATABASE BRONZE!")
```

TABELA CONFLITO CARREGADA COM SUCESSO DO DATABASE BRONZE!

```
## carregando os dados do DATABASE bronze para tratamento dos dados de Cidade
```

```
df_cidade_bronze_sql = spark.sql('select * from bronze.cidade')
```

```
print("TABELA CIDADE CARREGADA COM SUCESSO DO DATABASE BRONZE!")
```

TABELA CIDADE CARREGADA COM SUCESSO DO DATABASE BRONZE!

```
## 1- Tratamento do campo LOCATION na tabela Conflito - Removento acentos
```

```
acento = 'áãäåæéëïíóôõöùü'
```

```
sem_acento = 'aaaaaeiiiiooooouuu'
```

```
df_conflicto2 = df_conflicto_bronze_sql.withColumn('LOCATION_2', upper(translate(df_conflicto_bronze_sql['LOCATION'], acento, sem_acento)))
```

```
df_conflicto3 = df_conflicto2.withColumn('LOCATION_3', when(col('LOCATION_2').like ('%RIO DE JANEIRO%'), 'RIO DE JANEIRO').otherwise(col('LOCATION_2')))
```

```
df_conflicto4 = df_conflicto3.withColumn('ID_CITY', when(col('LOCATION_3').like ('%SAO PAULO%'), 'SAO PAULO').otherwise(col('LOCATION_3')))
```

```
#df_conflicto4.display()
```

```
# 2- Tratamento do campo EVENT_DATE na tabela Conflito - Transformando em mês e ano
```

```
split_cols = split(df_conflicto4['EVENT_DATE'], '-')
```

```
df_conflicto4 = df_conflicto4.withColumn('EVENT_MONTH', split_cols.getItem(1))
```

```
df_conflicto4 = df_conflicto4.withColumn('EVENT_YEAR', split_cols.getItem(2))
```

```
df_conflicto4 = df_conflicto4.replace(['January', 'February', 'March', 'April', 'May', 'June', 'July', 'August', 'September', 'October', 'November', 'December'],  
['1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12'])
```

```
df_conflicto4 = df_conflicto4.withColumn('EVENT_MONTH', df_conflicto4['EVENT_MONTH'].cast(IntegerType()))
```

```
df_conflicto4 = df_conflicto4.withColumn('EVENT_YEAR', df_conflicto4['EVENT_YEAR'].cast(IntegerType()))
```

```
#display(df_conflicto5)
```

```
# 3- Tratamento do campo CITY na tabela Cidade - Removento acentos

acento = 'áâãäåæçèéíîïóôõöùú'
sem_acento = 'aaaaaeiiiiooooouuu'
df_cidade2 = df_cidade_bronze_sql.withColumn('ID_CITY', upper(translate(df_cidade_bronze_sql['CITY'], acento, sem_acento)))

#print(df_cidade2.columns)
#print(df_conflito4.columns)
```

```
# 4 - Limpeza e transformação dos dados: tradução, eliminação de colunas, rename de colunas

lista_conflito = ['EVENT_DATE', 'LOCATION', 'LATITUDE', 'LONGITUDE', 'LOCATION_2', 'LOCATION_3']
df_conflito4 = df_conflito4.drop(*lista_conflito)

df_conflito4 = df_conflito4\
    .withColumnRenamed('EVENT_TYPE', 'TIPO_EVENTO')\
    .withColumnRenamed('SUB_EVENT_TYPE', 'SUB_TIPO_EVENTO')\
    .withColumnRenamed('ACTOR1', 'ATOR_PRIMARIO')\
    .withColumnRenamed('ACTOR2', 'ATOR_SECUNDARIO')\
    .withColumnRenamed('COUNTRY', 'PAIS')\
    .withColumnRenamed('SOURCE_SCALE', 'ESCALA_GEOGRAFICA')\
    .withColumnRenamed('NOTES', 'DESCRICAO')\
    .withColumnRenamed('FATALITIES', 'FATALIDADE')\
    .withColumnRenamed('EVENT_MONTH', 'MES')\
    .withColumnRenamed('EVENT_YEAR', 'ANO')

df_conflito4 = df_conflito4.withColumn('FATALIDADE', df_conflito4['FATALIDADE'].cast(IntegerType()))
```

```
lista_cidade = ['CITY', 'IBGE_RES_POP', 'IBGE_RES_POP_BRAS', 'IBGE_RES_POP_ESTR', 'IBGE_DU', 'IBGE_DU_URBAN', 'IBGE_DU_RURAL', 'IBGE_POP', 'IBGE_1', 'IBGE_1-4',
'IBGE_5-9', 'IBGE_10-14', 'IBGE_15-59', 'IBGE_60+', 'IBGE_PLANTED_AREA', 'IBGE_CROP_PRODUCTION_$', 'IDHM_Ranking_2010', 'IDHM', 'IDHM_Renda', 'IDHM_Longevidade',
'IDHM_Educacao', 'LONG', 'LAT', 'ALT', 'PAY_TV', 'FIXED_PHONES', 'AREA', 'REGIAO_TUR', 'CATEGORIA_TUR', 'ESTIMATED_POP', 'RURAL_URBAN', 'GVA_AGROPEC',
'GVA_INDUSTRY', 'GVA_SERVICES', 'GVA_PUBLIC', 'GVA_TOTAL', 'TAXES', 'GDP', 'POP_GDP', 'GDP_CAPITA', 'GVA_MAIN', 'MUN_EXPENDIT', 'COMP_TOT', 'COMP_A', 'COMP_B',
'COMP_C', 'COMP_D', 'COMP_E', 'COMP_F', 'COMP_G', 'COMP_H', 'COMP_I', 'COMP_J', 'COMP_K', 'COMP_L', 'COMP_M', 'COMP_N', 'COMP_O', 'COMP_P', 'COMP_Q', 'COMP_R',
'COMP_S', 'COMP_T', 'COMP_U', 'HOTELS', 'BEDS', 'Pr_Agencies', 'Pu_Agencies', 'Pr_Bank', 'Pu_Bank', 'Pr_Assets', 'Pu_Assets', 'Cars', 'Motorcycles',
'Wheeled_tractor', 'UBER', 'MAC', 'WAL-MART', 'POST_OFFICES']
df_cidade2 = df_cidade2.drop(*lista_cidade)

df_cidade2 = df_cidade2\
    .withColumnRenamed('STATE', 'ESTADO')
```

```
%sql
CREATE DATABASE IF NOT EXISTS silver
```

OK

```
print("DATABASE SILVER CRIADO COM SUCESSO!")
```

DATABASE SILVER CRIADO COM SUCESSO!

```
%sql
DROP TABLE IF EXISTS silver.conflito
```

OK

```
%sql
DROP TABLE IF EXISTS silver.cidade
```

OK

```
# 5- Criação do DATABASE Silver e Carga das tabelas Conflito e Cidade

dbutils.fs.rm('dbfs:/user/hive/warehouse/silver.db/conflito', True)

df_conflito4.write.format("delta").mode("append").saveAsTable("silver.conflito")

print("CARGA DADOS DE CONFLITO CRIADO COM SUCESSO NO DATABASE SILVER!")

dbutils.fs.rm('dbfs:/user/hive/warehouse/silver.db/cidade', True)

df_cidade2.write.format("delta").mode("append").saveAsTable("silver.cidade")

print("CARGA DADOS DE CIDADE CRIADO COM SUCESSO NO DATABASE SILVER!")
```

CARGA DADOS DE CONFLITO CRIADO COM SUCESSO NO DATABASE SILVER!  
CARGA DADOS DE CIDADE CRIADO COM SUCESSO NO DATABASE SILVER!

display(df\_conflito4)

display(df\_cidade2)

Table

	A <sub>C</sub> TIPO_EVENTO	A <sub>C</sub> SUB_TIPO_EVENTO	A <sub>C</sub> ATOR_PRIMARIO	A <sub>C</sub> ATOR_SECUNDARIO	A <sub>C</sub> PAIS	A <sub>C</sub> ESCALA_GEOGR
1	Battles	Armed clash	Unidentified Gang and/or Police Militia	Military Forces of Brazil (2016-2018) Military Police	Brazil	Subnational
2	Explosions/Remote violence	Remote explosive/landmine/IED	Unidentified Gang (Brazil)	null	Brazil	Subnational
3	Battles	Armed clash	Unidentified Gang and/or Police Militia	Unidentified Gang (Brazil)	Brazil	National
4	Violence against civilians	Attack	Unidentified Gang and/or Police Militia	Civilians (Brazil)	Brazil	National
5	Strategic developments	Looting/property destruction	Police Forces of Brazil (2016-2018) Federal Police	Unidentified Gang (Brazil)	Brazil	National
6	Battles	Armed clash	Military Forces of Brazil (2016-2018) Military Police	Unidentified Gang and/or Police Militia	Brazil	National-International
7	Explosions/Remote violence	Remote explosive/landmine/IED	Unidentified Gang (Brazil)	null	Brazil	National
8	Strategic developments	Looting/property destruction	Police Forces of Brazil (2016-2018) Federal Police	Unidentified Gang (Brazil)	Brazil	Subnational-National
9	Violence against civilians	Attack	Unidentified Gang and/or Police Militia	Civilians (Brazil)	Brazil	Subnational
10	Battles	Armed clash	Unidentified Gang (Brazil)	Unidentified Gang (Brazil)	Brazil	National
11	Battles	Armed clash	Unidentified Armed Group (Brazil)	Unidentified Armed Group (Brazil)	Brazil	New media
12	Battles	Armed clash	Unidentified Gang and/or Police Militia	Military Forces of Brazil (2016-2018) Military Police	Brazil	Subnational
13	Battles	Armed clash	Unidentified Gang and/or Police Militia	Military Forces of Brazil (2016-2018) Military Police	Brazil	National

Table

	A <sub>C</sub> ESTADO	A <sub>C</sub> CAPITAL	A <sub>C</sub> ID_CITY
1	GO	0	ABADIA DE GOIAS
2	MG	0	ABADIA DOS DOURADOS
3	GO	0	ABADIANIA
4	PA	0	ABAETETUBA
5	MG	0	ABAETE
6	CE	0	ABAIARA
7	BA	0	ABARE
8	PR	0	ABATIA
9	BA	0	ABAIRA
10	SC	0	ABDON BATISTA
11	PA	0	ABEL FIGUEIREDO
12	SC	0	ABELARDO LUZ
13	MG	0	ABRE CAMPO
14	PE	0	ABREU E LIMA
15	TO	0	ABREULANDIA

5,578 rows

----- Fim Notebook silver\_pipeline.py -----

6 - JUNTANDO INFORMAÇÕES DE CONFLITOS E CIDADES (GOLD)

1- Dados tratados para uso de processamento analítico ou BI

2- Criação do DATABASE gold e Carga das tabela Conflito

3- Descrição do catálogo de dados

```
%run
./gold/gold_pipeline
```

gold\_pipeline.py

databricks

gold\_pipeline (Python)

Import Notebook

```
## carregando os dados do DATABASE silver para tratamento dos dados de Conflito e Cidade

df_conflito_silver_sql = spark.sql('select cidade.ESTADO, conflito.* from silver.conflito, silver.cidade where conflito.ID_CITY = cidade.ID_CITY')

print('TABELA CONFLITO E CIDADE CARREGADA COM SUCESSO DO DATABASE SILVER!')
```

TABELA CONFLITO E CIDADE CARREGADA COM SUCESSO DO DATABASE SILVER!



```
%sql
CREATE DATABASE IF NOT EXISTS gold

OK

print('DATABASE GOLD CRIADO COM SUCESSO!')

DATABASE GOLD CRIADO COM SUCESSO!

%sql
DROP TABLE IF EXISTS gold.conflicto

OK
```

```
# 2- Criação do DATABASE gold e Carga das tabela Conflito

dbutils.fs.rm('dbfs:/user/hive/warehouse/gold.db/conflicto',True)

df_conflicto_silver_sql.write.format("delta").mode("append").saveAsTable("gold.conflicto")

print('CARGA DADOS DE CONFLITO CRIADO COM SUCESSO NO DATABASE GOLD!')

CARGA DADOS DE CONFLITO CRIADO COM SUCESSO NO DATABASE GOLD!
```

display(df\_conflicto\_silver\_sql)

Table						
		FATALIDADE	ID_CITY	MES	ANO	
1	in Viamao, Rio Grande do Sul, the military police tried to approach a suspicious car and an armed clash took place b...	0	VIAMAO	6	2018	
2	Caldas Novas, Goias, a bank was blown up by at least five armed individuals. Three suspects were arrested.	0	CALDAS NOVAS	6	2018	
3	in Teresina, Piaui, a man was killed and another was injured by armed men on a motorcycle in Parque Brasil II neighb...	1	TERESINA	6	2018	
4	in Santo Andre, Sao Paulo, a man was killed and his body was dismembered and abandoned in a wheelbarrow in Vil...	1	SANTO ANDRE	6	2018	
5	in Santo Andre, Sao Paulo, a man was killed and his body was dismembered and abandoned in a wheelbarrow in Vil...	1	SANTO ANDRE	6	2018	
6	in Borborema, Sao Paulo, 1,4 tones of cocaine were seized by the Federal Highway Police on km 458 of the BR-245 h...	0	BORBOREMA	6	2018	
7	in Borborema, Sao Paulo, 1,4 tones of cocaine were seized by the Federal Highway Police on km 458 of the BR-245 h...	0	BORBOREMA	6	2018	
8	in Angra dos Reis, Rio de Janeiro, the BOPE military police engaged in a shoot-out with drug traffickers during a sec...	2	ANGRA DOS REIS	6	2018	
9	Imaculada, Paraiba, a National Postal Service branch was blown up by at least four men.	0	IMACULADA	6	2018	
10	in Ponta Pora, Mato Grosso do Sul, 1.1 tons of marijuana were found inside a vehicle and seized by the Federal High...	0	PONTA PORA	6	2018	
11	2018 (as reported), in Rio Largo, Alagoas, a woman was killed and her body was found without the head in Mata do R...	1	RIO LARGO	6	2018	
12	in Coelho Neto, Maranhao, a drug trafficking group invaded the area of a rival drug trafficking group in order to gai...	1	COELHO NETO	6	2018	

# Catálogo de dados

```
%sql

SELECT MAX(LEN(conflicto.ESTADO)), MAX(LEN(conflicto.TIPO_EVENTO)), MAX(LEN(conflicto.SUB_TIPO_EVENTO)), MAX(LEN(conflicto.ATOR_PRIMARIO)), MAX(LEN(conflicto.
ATOR_SECUNDARIO)), MAX(LEN(conflicto.PAIS)), MAX(LEN(conflicto.ESCALA_GEOGRAFICA)), MAX(LEN(conflicto.DESCRICAO)), MAX(LEN(conflicto.FATALIDADE)), MAX(LEN(conflicto.
ID_CITY)), MAX(LEN(conflicto.MES)), MAX(LEN(conflicto.ANO))
FROM gold.conflicto
```

Table						
	max(len(ESTADO))	max(len(TIPO_EVENTO))	max(len(SUB_TIPO_EVENTO))	max(len(ATOR_PRIMARIO))	max(len(ATOR_SECUNDARIO))	max(len(PAIS))
1	2	26	34	101	101	101

1 row

```
print('\033[1m<< CATÁLOGO DE DADOS >>\033[0m')

df_catalogo = spark.createDataFrame(
    [
        ("ESTADO", "STRING", "2", "Representa a silga do Estado associado a Cidade", "Não Nulo", "Sigla válida dos estados brasileiros"),
        ("TIPO_EVENTO", "STRING", "30", "Tipo de evento primário do conflito", "Não nulo", "Máximo 30 caracteres"),
        ("SUB_TIPO_EVENTO", "STRING", "40", "Subtipo de evento associado ao conflito", "Nulo", "Máximo de 30 caracateres"),
        ("ATOR_PRIMARIO", "STRING", "150", "Agente primário causador do conflito", "Não nulo", "Máximo de 150 caracateres"),
        ("ATOR_SECUNDARIO", "STRING", "150", "Agente secundário causador do conflito", "Nulo", "Máximo de 150 caracateres"),
        ("PAIS", "STRING", "10", "País de origem", "Não nulo", "Máximo de 10 caracateres"),
        ("ESCALA_GEOGRAFICA", "STRING", "25", "Posição geográfica associada ao País", "Nulo", "Máximo de 25 caracateres"),
        ("DESCRICAO", "STRING", "2000", "Descrição detalhada do conflito", "Nulo", "Máximo de 2000 caracateres"),
        ("FATALIDADE", "INTEGER", "3", "Quantidade de mortes associado ao conflito", "Não nulo", "Máximo de 3 caracateres"),
        ("ID_CITY", "STRING", "20", "Cidade de origem do conflito", "Não Nulo", "Máximo de 20 caracateres. Chave estrangeira da tabela de origem Cidade"),
        ("MES", "INTEGER", "2", "Mês de ocorrência do conflito", "Não nulo", "Numérico que representa cada mês do ano"),
        ("ANO", "INTEGER", "4", "Ano de ocorrência do conflito", "Não nulo", "Exatamente 4 cacateres")
    ],
    ["Campo", "Tipo de dados", "Tamanho", "Descrição do dado", "Observação", "Regra de validação"]
)

display(df_catalogo)
```

display(df\_catalogo)

<< CATÁLOGO DE DADOS >>

Table

	<div>⌵</div> Campo	<div>⌵</div> Tipo de d...	<div>⌵</div> Tamanho	<div>⌵</div> Descrição do dado	<div>⌵</div> Observação	<div>⌵</div> Regra de validação
1	ESTADO	STRING	2	Representa a silga do Estado associado a Cida...	Não Nulo	Sigla válida dos estados brasileiros
2	TIPO_EVENTO	STRING	30	Tipo de evento primário do conflito	Não nulo	Máximo 30 caracteres
3	SUB_TIPO_EVENTO	STRING	40	Subtipo de evento associado ao conflito	Nulo	Máximo de 30 caracateres
4	ATOR_PRIMARIO	STRING	150	Agente primário causador do conflito	Não nulo	Máximo de 150 caracateres
5	ATOR_SECUNDARIO	STRING	150	Agente secundário causador do conflito	Nulo	Máximo de 150 caracateres
6	PAIS	STRING	10	País de origem	Não nulo	Máximo de 10 caracateres
7	ESCALA_GEOGRAFI...	STRING	25	Posição geográfica associada ao País	Nulo	Máximo de 25 caracateres
8	DESCRICAO	STRING	2000	Descrição detalhada do conflito	Nulo	Máximo de 2000 caracateres
9	FATALIDADE	INTEGER	3	Quantidade de mortes associado ao conflito	Não nulo	Máximo de 3 caracateres
10	ID_CITY	STRING	20	Cidade de origem do conflito	Não Nulo	Máximo de 20 caracateres. Chave estrangeira da tabela de c
11	MES	INTEGER	2	Mês de ocorrência do conflito	Não nulo	Númerico que representa cada mês do ano
12	ANO	INTEGER	4	Ano de ocorrência do conflito	Não nulo	Exatamente 4 caracateres

12 rows

Campo	Tipo de dados	Tama nho	Descrição do dado	Observ ação	Regra de validação
ESTADO	STRING	2	Representa a silga do Estado associado a Cidade	Não Nulo	Sigla válida dos estados brasileiros
TIPO_EVENTO	STRING	30	Tipo de evento primário do conflito	Não nulo	Máximo 30 caracteres
SUB_TIPO_EVENTO	STRING	40	Subtipo de evento associado ao conflito	Nulo	Máximo de 30 caracateres
ATOR_PRIMARIO	STRING	150	Agente primário causador do conflito	Não nulo	Máximo de 150 caracateres
ATOR_SECUNDARIO	STRING	150	Agente secundário causador do conflito	Nulo	Máximo de 150 caracateres
PAIS	STRING	10	País de origem	Não nulo	Máximo de 10 caracateres
ESCALA_GEOGRAFICA	STRING	25	Posição geográfica associada ao País	Nulo	Máximo de 25 caracateres
DESCRICAO	STRING	2000	Descrição detalhada do conflito	Nulo	Máximo de 2000 caracateres
FATALIDADE	INTEGER	3	Quantidade de mortes associado ao conflito	Não nulo	Máximo de 3 caracateres
ID_CITY	STRING	20	Cidade de origem do conflito	Não Nulo	Máximo de 20 caracateres. Chave estrangeira da tabela de origem Cidade
MES	INTEGER	2	Mês de ocorrência do conflito	Não nulo	Númerico que representa cada mês do ano
ANO	INTEGER	4	Ano de ocorrência do conflito	Não nulo	Exatamente 4 cacateres


----- Fim Notebook gold\_pipeline.py -----

# 7 - ANALISE DOS DADOS - SQL DASHBOARD

- 1- Quais as TOP 10 cidades mais Violentas do Brasil?
- 2- Que tipo de conflitos são mais comuns nessas cidades ?
- 3- Quais principais atores estão envolvidos nesses conflitos nessas cidades?
- 4- Quais as cidades mais letais do Brasil?
- 5- Qual ano obteve a maior fatalidade ?

```
%run
./analytics/analytics_notebook
```

## analytics\_notebook.py

 analytics\_notebook (Python) Import Notebook

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
from pyspark.sql import functions as F
import matplotlib.patches as mpatches
```

```
#1 - Quais as TOP 10 cidades mais Violentas do Brasil?
#Removendo da consulta os conflitos considerados protestos pacíficos

df_conflito_gold_sql_10 = spark.sql('''select conflito.ID_CITY, count(*) as CONFLITOS_VIOLENTOS from gold.conflito
where conflito.SUB_TIPO_EVENTO <> 'Peaceful protest'
group by conflito.ID_CITY order by 2 desc
LIMIT 10''')

print('\033[1m1 - Quais as TOP 10 cidades mais Violentas do Brasil?\033[0m')
display(df_conflito_gold_sql_10)
```

1 - Quais as TOP 10 cidades mais Violentas do Brasil?

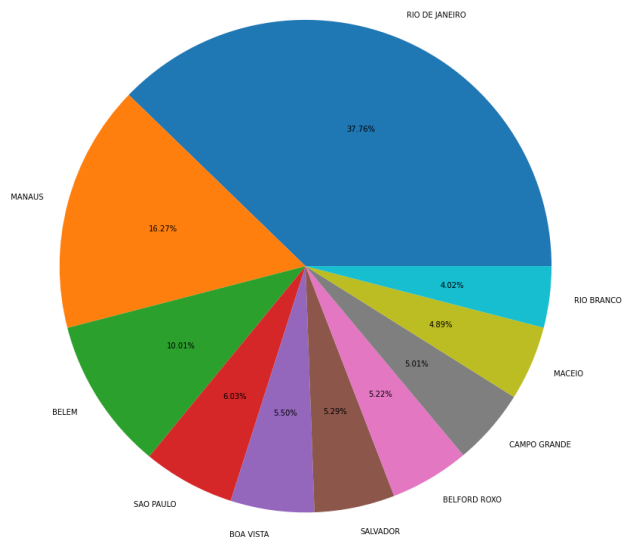
Table

	ID_CITY	CONFLITOS_VIOLENTOS
1	RIO DE JANEIRO	5973
2	MANAUS	2574
3	BELEM	1584
4	SAO PAULO	954
5	BOA VISTA	870
6	SALVADOR	836
7	BELFORD ROXO	825
8	CAMPO GRANDE	792
9	MACEIO	774
10	RIO BRANCO	636

10 rows

```
labels = df_conflito_gold_sql_10.agg(F.collect_list(F.col('ID_CITY'))).collect()[0][0]
vals = df_conflito_gold_sql_10.agg(F.collect_list(F.col('CONFLITOS_VIOLENTOS'))).collect()[0][0]
explode = (0.1,0,0,0,0,0,0,0,0,0)
fig, ax = plt.subplots(figsize=(22,15))
ax.pie(vals, labels=labels, autopct='%.2f%%')
ax.set_title('Quais as TOP 10 cidades mais Violentas do Brasil?', fontsize='26')
```

Quais as TOP 10 cidades mais Violentas do Brasil?



```
# 2- Que tipo de conflitos são mais comuns nessas cidades ?

top_10_cidades = df_conflito_gold_sql_10.select('ID_CITY')
top_10_cidades_list = top_10_cidades.agg(F.collect_list(F.col("ID_CITY"))).collect()[0][0]

df_conflito_gold_sql_tipo = spark.sql("""select conflito.ID_CITY, conflito.TIPO_EVENTO, conflito.SUB_TIPO_EVENTO, count(*)
from gold.conflito
where conflito.SUB_TIPO_EVENTO <> 'Peaceful protest'
group by conflito.ID_CITY, conflito.TIPO_EVENTO, conflito.SUB_TIPO_EVENTO order by 1,4 desc""")

print('\033[1m2 - Que tipo de conflitos são mais comuns nessas cidades?\033[0m')

for row in top_10_cidades.collect():
    df = df_conflito_gold_sql_tipo.where(df_conflito_gold_sql_tipo.ID_CITY == row["ID_CITY"])
    display(df)
```

2 - Que tipo de conflitos são mais comuns nessas cidades?

Table					Q F □	
	A <sub>0</sub> ID_CITY	A <sub>0</sub> TIPO_EVENTO	A <sub>0</sub> SUB_TIPO_EVENTO	1 <sub>2</sub> 3 count(1)		
1	RIO DE JANEIRO	Battles	Armed clash	5379		
2	RIO DE JANEIRO	Violence against civilians	Attack	268		
3	RIO DE JANEIRO	Riots	Violent demonstration	98		
4	RIO DE JANEIRO	Strategic developments	Looting/property destruction	51		
5	RIO DE JANEIRO	Strategic developments	Arrests	45		
6	RIO DE JANEIRO	Riots	Mob violence	38		
7	RIO DE JANEIRO	Explosions/Remote violence	Remote explosive/landmine/IED	27		
8	RIO DE JANEIRO	Protests	Protest with intervention	20		
9	RIO DE JANEIRO	Violence against civilians	Abduction/forced disappearance	18		
10	RIO DE JANEIRO	Strategic developments	Change to group/activity	12		
11	RIO DE JANEIRO	Strategic developments	Disrupted weapons use	7		
12	RIO DE JANEIRO	Strategic developments	Other	4		
13	RIO DE JANEIRO	Violence against civilians	Sexual violence	4		
14	RIO DE JANEIRO	Explosions/Remote violence	Grenade	2		

14 rows

Table

	📄 ID_CITY	📄 TIPO_EVENTO	📄 SUB_TIPO_EVENTO	📄 count(1)
1	MANAUS	Violence against civilia...	Attack	1512
2	MANAUS	Battles	Armed clash	703
3	MANAUS	Riots	Mob violence	260
4	MANAUS	Riots	Violent demonstration	44
5	MANAUS	Strategic developments	Looting/property destruction	19
6	MANAUS	Strategic developments	Arrests	11
7	MANAUS	Violence against civilia...	Sexual violence	7
8	MANAUS	Violence against civilia...	Abduction/forced disappearance	7
9	MANAUS	Protests	Protest with intervention	4
10	MANAUS	Strategic developments	Disrupted weapons use	3
11	MANAUS	Strategic developments	Other	3
12	MANAUS	Protests	Excessive force against proteste...	1

12 rows

Table

	📄 ID_CITY	📄 TIPO_EVENTO	📄 SUB_TIPO_EVENTO	📄 count(1)
1	BELEM	Violence against civilia...	Attack	900
2	BELEM	Battles	Armed clash	429
3	BELEM	Riots	Violent demonstration	177
4	BELEM	Riots	Mob violence	36
5	BELEM	Protests	Protest with intervention	18
6	BELEM	Strategic developments	Arrests	15
7	BELEM	Strategic developments	Looting/property destruction	6
8	BELEM	Strategic developments	Disrupted weapons use	3

8 rows

Table

	📄 ID_CITY	📄 TIPO_EVENTO	📄 SUB_TIPO_EVENTO	📄 count(1)
1	SAO PAULO	Battles	Armed clash	465
2	SAO PAULO	Violence against civilians	Attack	256
3	SAO PAULO		Violent demonstration	86
4	SAO PAULO	Riots	Mob violence	57
5	SAO PAULO	Protests	Protest with intervention	23
6	SAO PAULO	Strategic developments	Arrests	22
7	SAO PAULO	Strategic developments	Looting/property destruction	20
8	SAO PAULO	Violence against civilians	Abduction/forced disappearance	13
9	SAO PAULO	Explosions/Remote violence	Remote explosive/landmine/IED	4
10	SAO PAULO	Protests	Excessive force against proteste...	3
11	SAO PAULO	Strategic developments	Other	3
12	SAO PAULO	Strategic developments	Disrupted weapons use	1
13	SAO PAULO	Violence against civilians	Sexual violence	1

13 rows

Table

	📄 ID_... ⋮ ↕	📄 TIPO_EVENTO	📄 SUB_TIPO_EVENTO	📄 count(1)
1	BOA VISTA	Violence against civilians	Attack	514
2	BOA VISTA	Battles	Armed clash	206
3	BOA VISTA	Riots	Mob violence	100
4	BOA VISTA	Strategic developments	Looting/property destruction	14
5	BOA VISTA	Riots	Violent demonstration	12
6	BOA VISTA	Violence against civilians	Abduction/forced disappearance	10
7	BOA VISTA	Explosions/Remote violence	Remote explosive/landmine/IED	4
8	BOA VISTA	Violence against civilians	Sexual violence	4
9	BOA VISTA	Strategic developments	Arrests	2
10	BOA VISTA	Explosions/Remote violence	Grenade	2
11	BOA VISTA	Protests	Protest with intervention	2

11 rows

Table

	Ⓐ <sub>C</sub> ID_CITY	Ⓐ <sub>C</sub> TIPO_EVENTO	Ⓐ <sub>C</sub> SUB_TIPO_EVENTO	± <sub>3</sub> count(1)
1	SALVADOR	Battles	Armed clash	529
2	SALVADOR	Violence against civilians	Attack	180
3	SALVADOR	Riots	Violent demonstration	80
4	SALVADOR	Explosions/Remote violence	Remote explosive/landmine/IED	15
5	SALVADOR	Riots	Mob violence	10
6	SALVADOR	Strategic developments	Looting/property destruction	10
7	SALVADOR	Protests	Protest with intervention	4
8	SALVADOR	Violence against civilians	Abduction/forced disappearance	4
9	SALVADOR	Strategic developments	Arrests	3
10	SALVADOR	Violence against civilians	Sexual violence	1

10 rows

Table

	Ⓐ <sub>C</sub> ID_CITY	Ⓐ <sub>C</sub> TIPO_EVENTO	Ⓐ <sub>C</sub> SUB_TIPO_EVENTO	± <sub>3</sub> count(1)
1	BELFORD ROXO	Battles	Armed clash	785
2	BELFORD ROXO	Violence against civilia...	Attack	26
3	BELFORD ROXO	Riots	Violent demonstration	5
4	BELFORD ROXO	Strategic developments	Other	2
5	BELFORD ROXO	Riots	Mob violence	2
6	BELFORD ROXO	Violence against civilia...	Sexual violence	1
7	BELFORD ROXO	Strategic developments	Non-violent transfer of territory	1
8	BELFORD ROXO	Strategic developments	Arrests	1
9	BELFORD ROXO	Strategic developments	Change to group/activity	1
10	BELFORD ROXO	Protests	Excessive force against proteste...	1

10 rows

Table

	Ⓐ <sub>C</sub> ID_CITY	Ⓐ <sub>C</sub> TIPO_EVENTO	Ⓐ <sub>C</sub> SUB_TIPO_EVENTO	± <sub>3</sub> count(1)
1	CAMPO GRANDE	Violence against civilia...	Attack	340
2	CAMPO GRANDE	Battles	Armed clash	214
3	CAMPO GRANDE	Strategic developments	Looting/property destruction	106
4	CAMPO GRANDE	Riots	Mob violence	92
5	CAMPO GRANDE	Riots	Violent demonstration	20
6	CAMPO GRANDE	Violence against civilia...	Sexual violence	8
7	CAMPO GRANDE	Protests	Protest with intervention	4
8	CAMPO GRANDE	Violence against civilia...	Abduction/forced disappearance	4
9	CAMPO GRANDE	Strategic developments	Arrests	2
10	CAMPO GRANDE	Strategic developments	Disrupted weapons use	2

10 rows

Table

	Ⓐ <sub>C</sub> ID_CITY	Ⓐ <sub>C</sub> TIPO_EVENTO	Ⓐ <sub>C</sub> SUB_TIPO_EVENTO	± <sub>3</sub> count(1)
1	MACEIO	Violence against civilia...	Attack	369
2	MACEIO	Battles	Armed clash	231
3	MACEIO	Riots	Violent demonstration	85
4	MACEIO	Riots	Mob violence	72
5	MACEIO	Violence against civilia...	Abduction/forced disappearance	5
6	MACEIO	Strategic developments	Arrests	4
7	MACEIO	Violence against civilia...	Sexual violence	4
8	MACEIO	Protests	Protest with intervention	3
9	MACEIO	Strategic developments	Disrupted weapons use	1

9 rows

Table

	ID_CITY	TIPO_EVENTO	SUB_TIPO_EVENTO	count(1)
1	RIO BRANCO	Violence against civilia...	Attack	400
2	RIO BRANCO	Battles	Armed clash	170
3	RIO BRANCO	Riots	Violent demonstration	16
4	RIO BRANCO	Strategic developments	Arrests	12
5	RIO BRANCO	Riots	Mob violence	12
6	RIO BRANCO	Protests	Protest with intervention	10
7	RIO BRANCO	Violence against civilia...	Abduction/forced disappearance	6
8	RIO BRANCO	Strategic developments	Looting/property destruction	4
9	RIO BRANCO	Protests	Excessive force against proteste...	2
10	RIO BRANCO	Strategic developments	Other	2
11	RIO BRANCO	Strategic developments	Non-violent transfer of territory	2

11 rows

```
# 3 - Quais principais atores estão envolvidos nesses conflitos?

from pyspark.sql.functions import *

print('\033[1m3 - Quais atores principais estão envolvidos nesses conflitos?\033[0m')

df_conflito_gold_sql_ator = spark.sql("""select conflito.ID_CITY, conflito.ATOR_PRIMARIO, count(*) as QTD
from gold.conflito
where conflito.SUB_TIPO_EVENTO <> 'Peaceful protest'
group by conflito.ID_CITY, conflito.ATOR_PRIMARIO order by 3 desc
""")

for row in top_10_cidades.collect():
    df = df_conflito_gold_sql_ator.where(df_conflito_gold_sql_ator.ID_CITY == row["ID_CITY"])
    display(df.limit(10))
```

3 - Quais atores principais estão envolvidos nesses conflitos?

	ID_CITY	ATOR_PRIMARIO	QTD
1	RIO DE JANEIRO	Unidentified Armed Group (Brazil)	3215
2	RIO DE JANEIRO	Unidentified Gang and/or Police Militia	812
3	RIO DE JANEIRO	Military Forces of Brazil (2019-2022) Military Police	467
4	RIO DE JANEIRO	CV: Red Command	415
5	RIO DE JANEIRO	TCP: Pure Third Command	184
6	RIO DE JANEIRO	Military Forces of Brazil (2016-2018) Military Police	142
7	RIO DE JANEIRO	Rioters (Brazil)	136
8	RIO DE JANEIRO	Police Forces of Brazil (2019-2022)	133
9	RIO DE JANEIRO	Military Forces of Brazil (2019-2022) UPP: Pacifying Police Unit	114
10	RIO DE JANEIRO	Police Forces of Brazil (2019-2022) Civil Police	73

10 rows

Table

	ID_CITY	ATOR_PRIMARIO	QTD
1	MANAUS	Unidentified Gang and/or Police Militia	1038
2	MANAUS	Unidentified Gang (Brazil)	526
3	MANAUS	Unidentified Armed Group (Brazil)	362
4	MANAUS	Rioters (Brazil)	304
5	MANAUS	Military Forces of Brazil (2019-2022) Military Police	181
6	MANAUS	CV: Red Command	53
7	MANAUS	FDN: Family of the North	28
8	MANAUS	PCC: First Capital Command	17
9	MANAUS	Military Forces of Brazil (2016-2018) Military Police	16
10	MANAUS	Police Forces of Brazil (2019-2022) Civil Police	13

10 rows

Table

	A <sup>B</sup> <sub>C</sub> ID_CITY	A <sup>B</sup> <sub>C</sub> ATOR_PRIMARIO	1 <sup>2</sup> <sub>3</sub> QTD
1	BELEM	Unidentified Gang and/or Police Militia	807
2	BELEM	Rioters (Brazil)	213
3	BELEM	Unidentified Armed Group (Brazil)	201
4	BELEM	Military Forces of Brazil (2019-2022) Military Police	177
5	BELEM	Unidentified Gang (Brazil)	93
6	BELEM	Military Forces of Brazil (2016-2018) Military Police	21
7	BELEM	Protesters (Brazil)	18
8	BELEM	Police Forces of Brazil (2019-2022) Civil Police	18
9	BELEM	Police Forces of Brazil (2016-2018) Civil Police	9
10	BELEM	Police Militia	6

10 rows

Table

	A <sup>B</sup> <sub>C</sub> ID_CITY	A <sup>B</sup> <sub>C</sub> ATOR_PRIMARIO	1 <sup>2</sup> <sub>3</sub> QTD
1	SAO PAULO	Unidentified Armed Group (Brazil)	375
2	SAO PAULO	Rioters (Brazil)	143
3	SAO PAULO	Unidentified Gang and/or Police Militia	133
4	SAO PAULO	Military Forces of Brazil (2019-2022) Military Police	124
5	SAO PAULO	Unidentified Gang (Brazil)	37
6	SAO PAULO	Police Forces of Brazil (2019-2022) Civil Police	28
7	SAO PAULO	Protesters (Brazil)	26
8	SAO PAULO	PCC: First Capital Command	17
9	SAO PAULO	Military Forces of Brazil (2016-2018) Military Police	13
10	SAO PAULO	Police Forces of Brazil (2019-2022)	11

10 rows

Table

	A <sup>B</sup> <sub>C</sub> ID_CITY	A <sup>B</sup> <sub>C</sub> ATOR_PRIMARIO	1 <sup>2</sup> <sub>3</sub> QTD
1	BOA VISTA	Unidentified Gang and/or Police Militia	314
2	BOA VISTA	Unidentified Gang (Brazil)	202
3	BOA VISTA	Rioters (Brazil)	108
4	BOA VISTA	Unidentified Armed Group (Brazil)	102
5	BOA VISTA	PCC: First Capital Command	44
6	BOA VISTA	Military Forces of Brazil (2019-2022) Military Police	40
7	BOA VISTA	Unidentified Armed Group (Venezuela)	14
8	BOA VISTA	Military Forces of Brazil (2016-2018) Military Police	12
9	BOA VISTA	CV: Red Command	8
10	BOA VISTA	Unidentified Gang and/or Colectivo	6

10 rows

Table

	A <sup>B</sup> <sub>C</sub> ID_CITY	A <sup>B</sup> <sub>C</sub> ATOR_PRIMARIO	1 <sup>2</sup> <sub>3</sub> QTD
1	SALVADOR	Military Forces of Brazil (2019-2022) Military Police	289
2	SALVADOR	Unidentified Gang (Brazil)	153
3	SALVADOR	Unidentified Gang and/or Police Militia	112
4	SALVADOR	Rioters (Brazil)	90
5	SALVADOR	Unidentified Armed Group (Brazil)	74
6	SALVADOR	Military Forces of Brazil (2016-2018) Military Police	46
7	SALVADOR	BDM: Bonde do Maluco	17
8	SALVADOR	Police Forces of Brazil (2019-2022) Civil Police	15
9	SALVADOR	CV: Red Command	10
10	SALVADOR	Katiara Gang	5

10 rows



Table

	Ⓐ ID_CITY	Ⓐ ATOR_PRIMARIO	1 <sup>2</sup> QTD
1	BELFORD ROXO	Unidentified Armed Group (Brazil)	625
2	BELFORD ROXO	Military Forces of Brazil (2019-2022) Military Police	65
3	BELFORD ROXO	Unidentified Gang and/or Police Militia	49
4	BELFORD ROXO	Police Forces of Brazil (2019-2022)	29
5	BELFORD ROXO	TCP: Pure Third Command	20
6	BELFORD ROXO	CV: Red Command	15
7	BELFORD ROXO	Rioters (Brazil)	7
8	BELFORD ROXO	Military Forces of Brazil (2016-2018) Military Police	5
9	BELFORD ROXO	Police Forces of Brazil (2016-2018)	4
10	BELFORD ROXO	Police Forces of Brazil (2016-2018) Civil Police	1

10 rows

Table

	Ⓐ ID_CITY	Ⓐ ATOR_PRIMARIO	1 <sup>2</sup> QTD
1	CAMPO GRANDE	Unidentified Gang and/or Police Militia	264
2	CAMPO GRANDE	Rioters (Brazil)	112
3	CAMPO GRANDE	Unidentified Gang (Brazil)	104
4	CAMPO GRANDE	Military Forces of Brazil (2019-2022) Military Police	88
5	CAMPO GRANDE	Unidentified Armed Group (Brazil)	64
6	CAMPO GRANDE	PCC: First Capital Command	62
7	CAMPO GRANDE	Police Forces of Brazil (2019-2022) Civil Police	30
8	CAMPO GRANDE	Police Forces of Brazil (2019-2022) Federal Police	22
9	CAMPO GRANDE	Police Forces of Brazil (2016-2018) Federal Police	16
10	CAMPO GRANDE	Military Forces of Brazil (2016-2018) Military Police	10

10 rows

Table

	Ⓐ ID_CITY	Ⓐ ATOR_PRIMARIO	1 <sup>2</sup> QTD
1	MACEIO	Unidentified Gang and/or Police Militia	240
2	MACEIO	Rioters (Brazil)	157
3	MACEIO	Unidentified Armed Group (Brazil)	152
4	MACEIO	Unidentified Gang (Brazil)	110
5	MACEIO	Military Forces of Brazil (2019-2022) Military Police	67
6	MACEIO	Military Forces of Brazil (2016-2018) Military Police	28
7	MACEIO	Police Forces of Brazil (2016-2018)	5
8	MACEIO	Protesters (Brazil)	3
9	MACEIO	Police Forces of Brazil (2019-2022)	2
10	MACEIO	Police Forces of Brazil (2019-2022) Federal Police	2

10 rows

Table

	Ⓐ ID_CITY	Ⓐ ATOR_PRIMARIO	1 <sup>2</sup> QTD
1	RIO BRANCO	Unidentified Gang and/or Police Militia	258
2	RIO BRANCO	Unidentified Gang (Brazil)	188
3	RIO BRANCO	Unidentified Armed Group (Brazil)	60
4	RIO BRANCO	Rioters (Brazil)	28
5	RIO BRANCO	Military Forces of Brazil (2019-2022) Military Police	24
6	RIO BRANCO	CV: Red Command	22
7	RIO BRANCO	B13: Tram of 13	16
8	RIO BRANCO	Military Forces of Brazil (2016-2018) Military Police	14
9	RIO BRANCO	Protesters (Brazil)	12
10	RIO BRANCO	Police Forces of Brazil (2019-2022) Federal Police	6

10 rows

```
# 4- Qual as cidades mais letais do Brasil?

print('\033[1m4- Qual as cidades mais letais do Brasil?\033[0m')

df_conflito_gold_sql_letal = spark.sql('''select conflito.ID_CITY, sum(conflito.FATALIDADE) as TOTAL_MORTES
from gold.conflito
where conflito.SUB_TIPO_EVENTO <> 'Peaceful protest' and conflito.FATALIDADE <> 0
group by conflito.ID_CITY order by 2 desc
limit 10
''')

display(df_conflito_gold_sql_letal)
```

4- Qual as cidades mais letais do Brasil?

Table

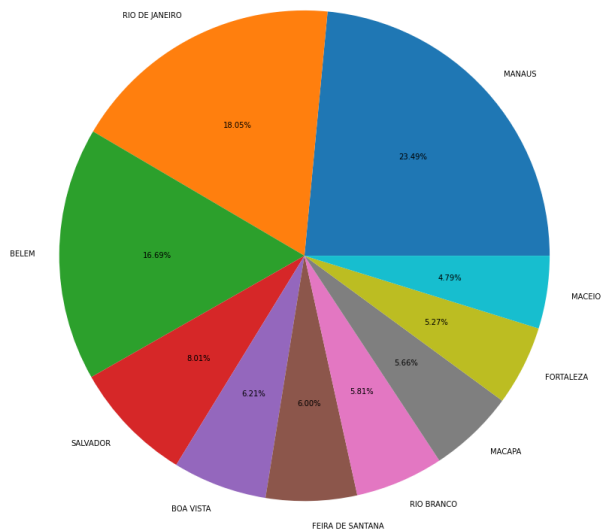
Q F □

	1 ID_CITY	2 TOTAL_MORTES
1	MANAUS	2344
2	RIO DE JANEIRO	1801
3	BELEM	1665
4	SALVADOR	799
5	BOA VISTA	620
6	FEIRA DE SANTANA	599
7	RIO BRANCO	580
8	MACAPA	565
9	FORTALEZA	526
10	MACEIO	478

10 rows

```
labels = df_conflito_gold_sql_letal.agg(F.collect_list(F.col('ID_CITY'))).collect()[0][0]
vals = df_conflito_gold_sql_letal.agg(F.collect_list(F.col('TOTAL_MORTES'))).collect()[0][0]
explode = (0.1,0,0,0,0,0,0,0,0,0)
fig, ax = plt.subplots(figsize=(22,15))
ax.pie(vals, labels=labels, autopct='%2f%%')
ax.set_title('4- Quais as cidades mais letais do Brasil?', fontsize='26')
```

4- Quais as cidades mais letais do Brasil?



```
# 5- Qual ano obteve maior fatalidade ?

print('\033[1m5- Qual ano obteve maior fatalidade?\033[0m')

df_conflito_gold_sql_ano = spark.sql('''select conflito.ANO, sum(conflito.FATALIDADE) as TOTAL_MORTES
from gold.conflito
where conflito.SUB_TIPO_EVENTO <> 'Peaceful protest' and conflito.FATALIDADE <> 0
group by conflito.ANO order by 1
''')

display(df_conflito_gold_sql_ano)
```

5- Qual ano obteve maior fatalidade?

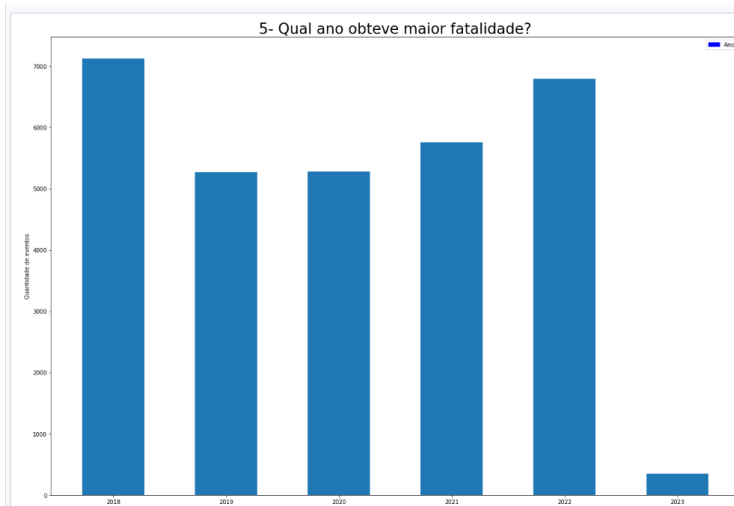
Table

	ANO	TOTAL_MORTES
1	2018	7121
2	2019	5274
3	2020	5283
4	2021	5759
5	2022	6793
6	2023	347

6 rows

```
labels = df_conflito_gold_sql_ano.agg(F.collect_list(F.col('ANO'))).collect()[0][0]
vals = df_conflito_gold_sql_ano.agg(F.collect_list(F.col('TOTAL_MORTES'))).collect()[0][0]
fig, ax = plt.subplots(figsize=(22,15))

ax.bar(labels,vals, 0.55, label='Ano')
ax.set_title('5- Qual ano obteve maior fatalidade?', fontsize='26')
ax.set_ylabel('Quantidade de eventos')
red_patch = mpatches.Patch(color='blue', label='Ano')
ax.legend(handles=[red_patch])
plt.show()
```



----- Fim Notebook analytics\_notebook.py -----

## Análise e Autoavaliação

### 1- Qualidade de dados

Durante a análise dos dados foi verificado que alguns atributos estavam comprometidos para modelagem e análise dos dados. Esses problemas foram tratados na camada Silver.

O ano de 2023 possui apenas dados de janeiro de 2023. Logo, não podemos fazer uma análise comparativa do ano inteiro.

Na camada Gold apenas foi necessário unificar a tabela CONFLITO, incluindo a coluna ESTADO. Dessa forma uma consulta simples a essa tabela é suficiente para responder a uma pergunta de negócio envolvendo Estado, sem necessidade de uma junção, por exemplo.

### 2- Solução do problema

Todas as perguntas enumeradas antes de iniciar as etapas anteriores foram respondidas com sucesso.

### **3- Autoavaliação**

Minha autoavaliação a respeito do problema proposto e da solução obtida foi satisfatória. Conforme sugestão da equipe de docentes, usei somente a Plataforma Databricks para implementar a solução. Foi possível simular um PIPELINE de dados na ferramenta, passando pelas etapas de busca, coleta, modelagem, carga e análise dos dados. Estou ciente que existem outras plataformas de serviço de computação em nuvem, como AWS, Google Cloud e Microsoft Azure que permitiriam uma integração mais automatizada do processo. Não obtive dificuldades no conteúdo, mas para realizar o trabalho fiz complemento do estudo através de vídeos explicativos sobre os temas Databricks e PySpark, no YOUTUBE. As dúvidas extraídas pela ferramenta DISCORD e as Transmissões Ao Vivo – MVP também foram suficientes para execução do trabalho.