BC 17 - Engenharia de Dados

PROJETO FINAL
EQUIPE 3

ACIDENTES TERRESTRES

Ana Paula Guimarães
Andréa Goulart
Bruno Yaporandy
Carlos Dudas

Projeto Final – Equipe 3

Este relatório contém uma descrição das atividades realizadas para a execução do Projeto Final da Equipe 3 da turma BC17 – Engenharia de Dados da SoulCode Academy.

Para esse projeto foi proposto à equipe o tema "Acidentes Terrestres", tendo como orientação a aplicação de conceitos vistos durante o curso, tais como: tratamento, organização e modelagem de dados de 2 ou mais datasets, sendo um, fornecido pela SoulCode e o(s) outro(s) escolhido(s) pela equipe. Também como orientação foi exigido a utilização das seguintes tecnologias: Google Cloud Platform (Cloud Storage), Python, Pandas, PySpark, SparkSQL, Apache Beam, Data Studio, Big Query e NoSQL. Além das orientações acima, foram estabelecidos os seguintes requisitos:

Requisitos Obrigatórios:

- Os datasets devem ter formatos diferentes (CSV / Json / Parquet / Sql / NoSql) e 1 deles obrigatoriamente o que for fornecido para o projeto (o mesmo j está em CSV).
- Converter e normalizar os dados via SPARK (csv/parquet)
- Haver utilização de triggers e procedures para o banco SQL
- Entregar todos os scripts (DDL//DML)
- Utilizar o banco NoSQL (MongoDB ou Cassandra) como um datalake
- 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 datalake (Gstorage) , DW(BigQuery) ou em ambos.
- Os datasets originais devem ser armazenados em MySql ou PostgresSQL
- Os Dataframe(s) resultante(s) deve(m) estar em uma coleção do mongoDb atlas (informar a key de acesso ao cluster) e preferencialmente criar o usuário (soulcode) e senha (a1b2c3) no cluster
- Deve ser feito análises dentro do Big Query utilizando a linguagem padrão SQL com a descrição das consultas feitas.
- Deve ser criado no datastudio um dashboard para exibição gráfica dos dados tratados trazendo insights importantes
- E deve ser demonstrado em um workflow simples (gráfico) as etapas de ETL com suas respectivas ferramentas.

Requisitos desejáveis:

- Implementar captura e ingestão de dados por meio de uma PIPELINE com modelo criado em apache beam usando o dataflow para o work
- Utilizar o dataflow com algum modelo pré-definido
- Criar plotagens usando pandas para alguns insights durante o processo de Transformação
- Por meio de uma PIPELINE fazer o carregamento dos dados normalizados diretamente para um DW ou DataLake ou ambos
- Montar um relatório completo com os insights que justificam todo o processo de ETL utilizado
- Levantar custos com a utilização do google cloud no período do projeto e possíveis otimizações de custo.

No intuito de evidenciar a execução das atividades, listaremos e demonstraremos através de *prints,* algumas etapas do nosso trabalho, seguindo a orientação dos professores:

1) Datasets trabalhados

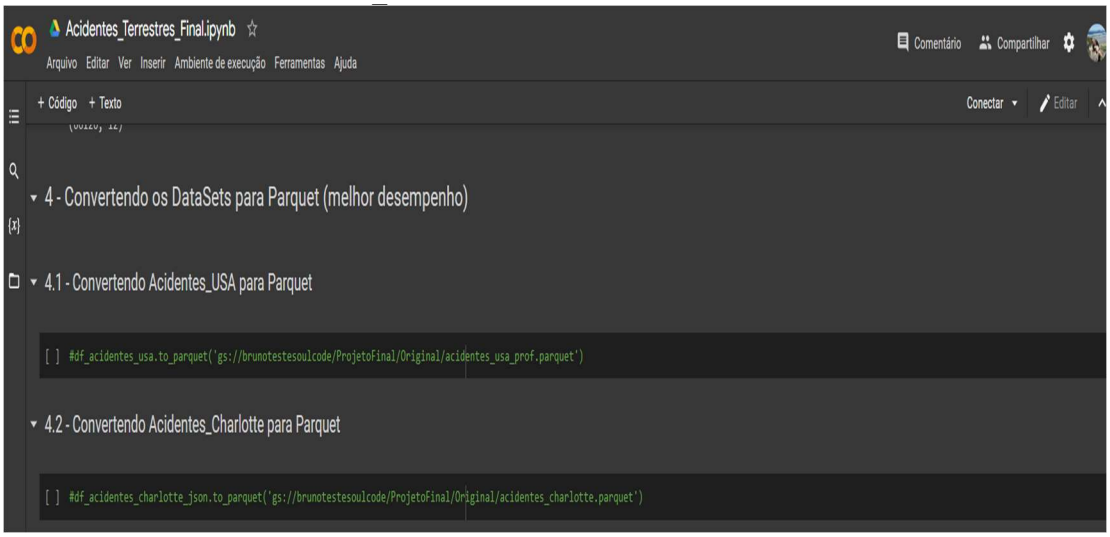
- Conforme notebook Acidentes_terrestres item 3

Equipe3.csv (Fonte: Soulcode Academy) e

Accidents.json – (Fonte: https://www.kaggle.com/datasets/rapiddev/charlotte-nc-traffic-accidents-20182019)

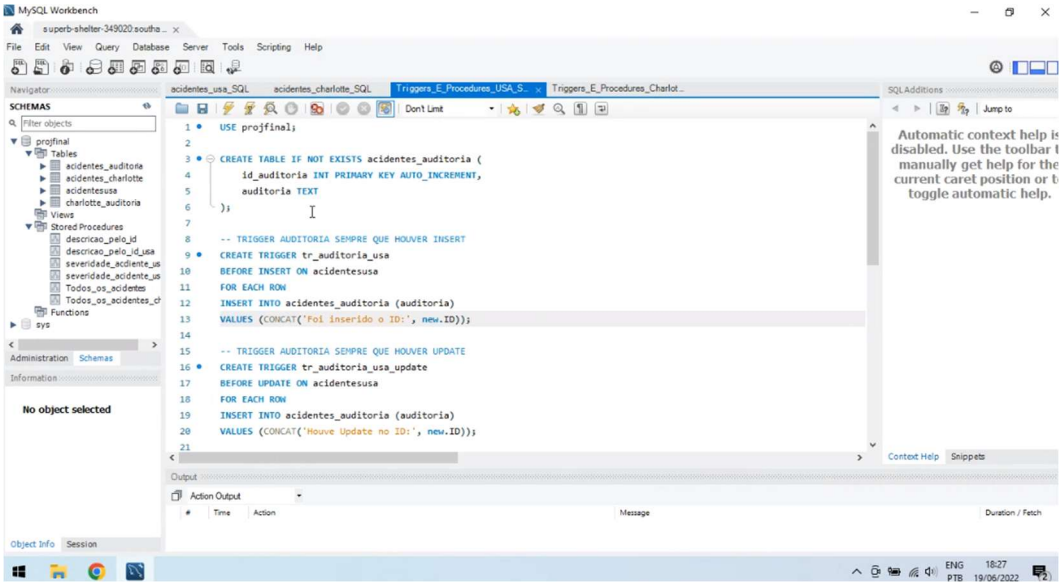
2. Conversão e normalização dos datasets no formato parquet

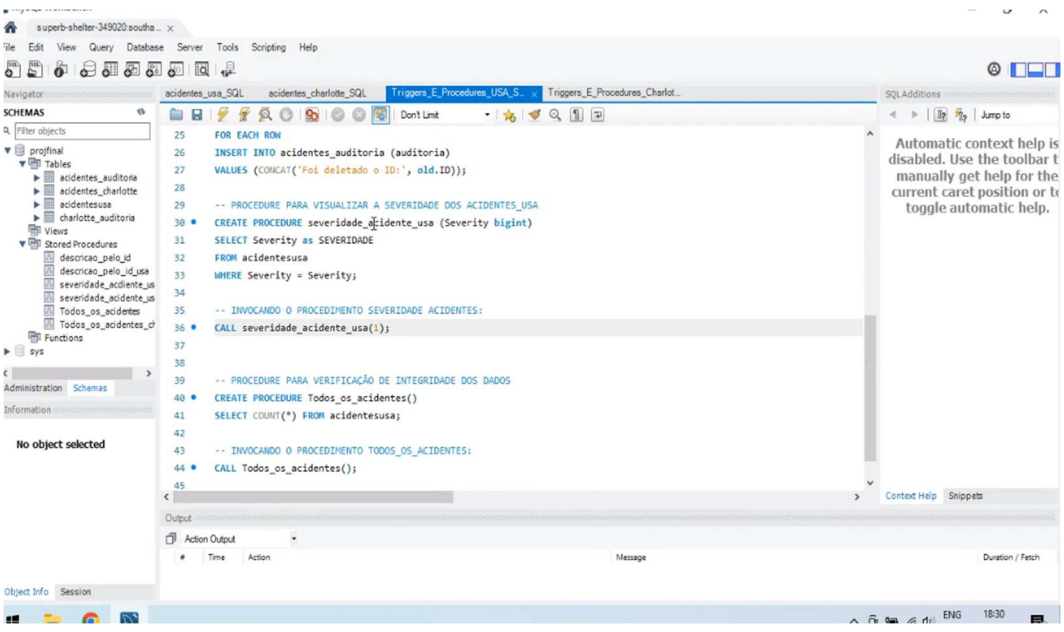
- Conforme notebook Acidentes_terrestres item 4



3. Criação de triggers e procedures para o banco SQL

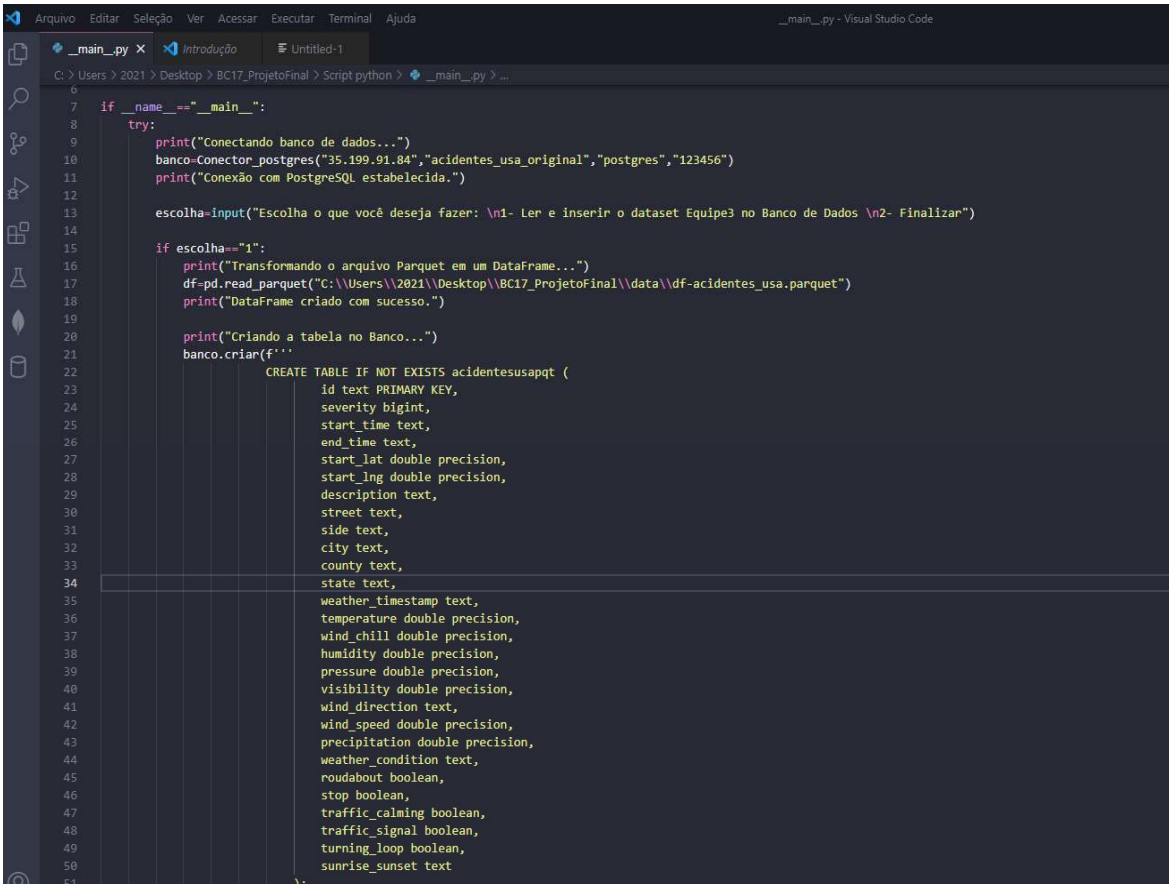
- Conforme anexo





4. Criação do script (DDL e DML) do dataset Equipe3(acidenteusa)

- Conforme anexo



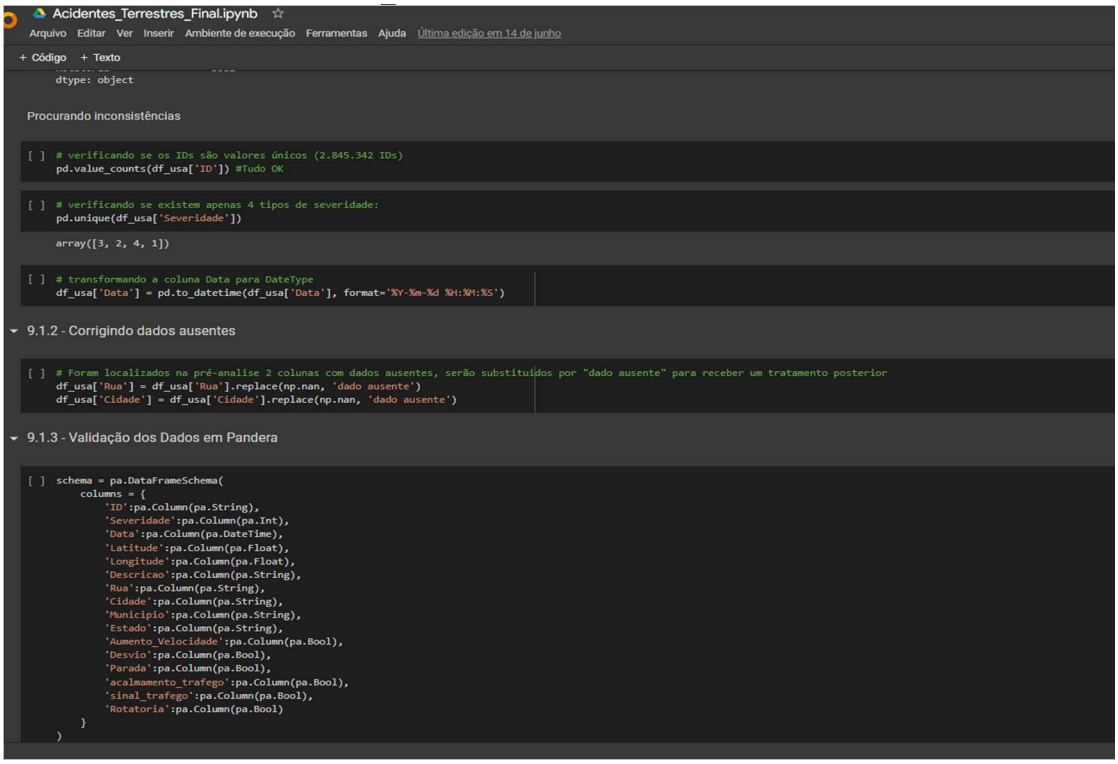
```
if __name__=="__main__":
    try:
        print("Conectando banco de dados...")
        banco=Conector_postgres("35.199.91.84","acidentes_usa_original","postgres","123456")
        print("Conexão com PostgreSQL estabelecida.")

        escolha=input("Escolha o que você deseja fazer: \n1- Ler e inserir o dataset Equipe3 no Banco de Dados \n2- Finalizar")

        if escolha=="1":
            print("Transformando o arquivo Parquet em um DataFrame...")
            df=pd.read_parquet("C:\\Users\\2021\\Desktop\\BC17_ProjetoFinal\\data\\df-acidentes_usa.parquet")
            print("DataFrame criado com sucesso.")

            print("Criando a tabela no Banco...")
            banco.criar(f'''
                        CREATE TABLE IF NOT EXISTS acidentesusapqt (
                                id text PRIMARY KEY,
                                severity bigint,
                                start_time text,
                                end_time text,
                                start_lat double precision,
                                start_lng double precision,
                                description text,
                                street text,
                                side text,
                                city text,
                                county text,
                                state text,
                                weather_timestamp text,
                                temperature double precision,
                                wind_chill double precision,
                                humidity double precision,
                                pressure double precision,
                                visibility double precision,
                                wind_direction text,
                                wind_speed double precision,
                                precipitation double precision,
                                weather_condition text,
                                roudabout boolean,
                                stop boolean,
                                traffic_calming boolean,
                                traffic_signal boolean,
                                turning_loop boolean,
                                sunrise_sunset text
                            );
                ''')
            print("Tabela Criada com sucesso.")

            print("Inserindo o DataFrame no Banco...")
            for i,x in df.iterrows():
                banco.inserir(f"INSERT INTO acidentesusapqt (id, severity, start_time, end_time, start_lat, start_lng, description, street, side, city, county, state, weather_timestamp, temperature, wind_chill, humidi
            print("DataFrame inserido com sucesso.")


            escolha="2"

        elif escolha=="2":
            print("\n-------------------------------------------------------------\n")


    except Exception as e:
        print(str(e))
```

5. Utilizar o banco NoSQL (MongoDB ou Cassandra) como um datalake.

- Conforme notebook Acidentes_terrestres item 12

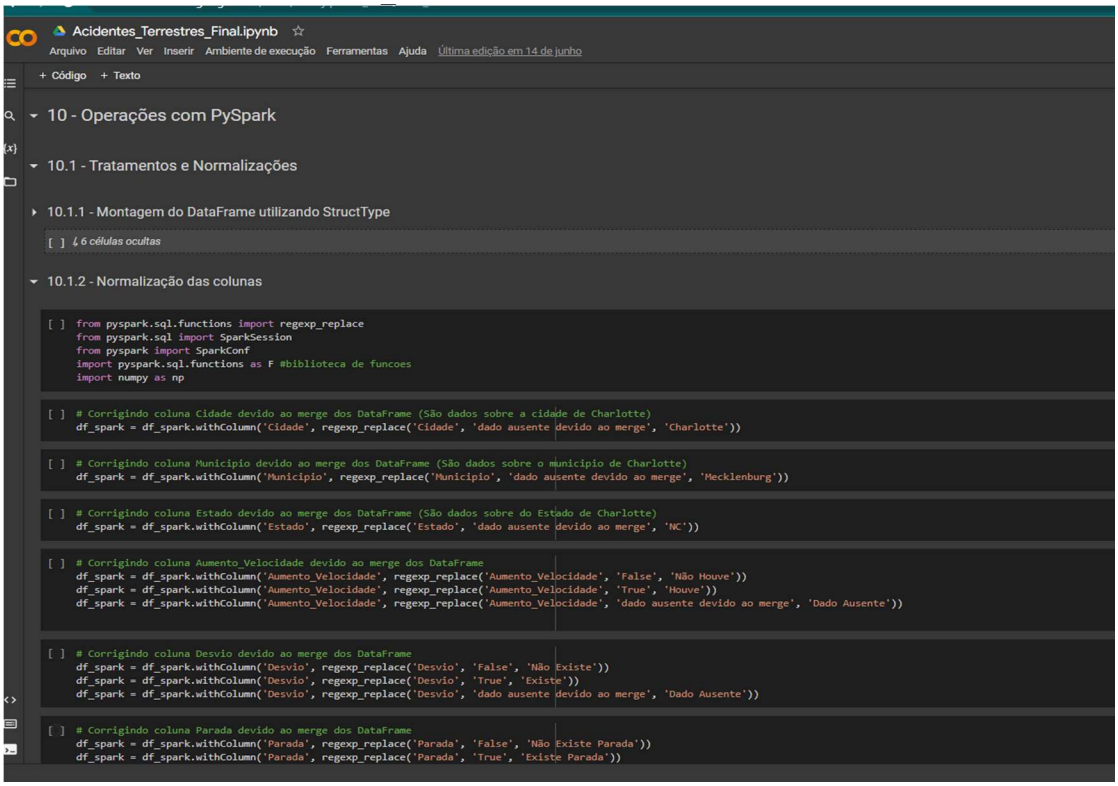
6. Operações com Pandas (limpeza, transformações e normalizações)

- Conforme notebook Acidentes_terrestres item 10



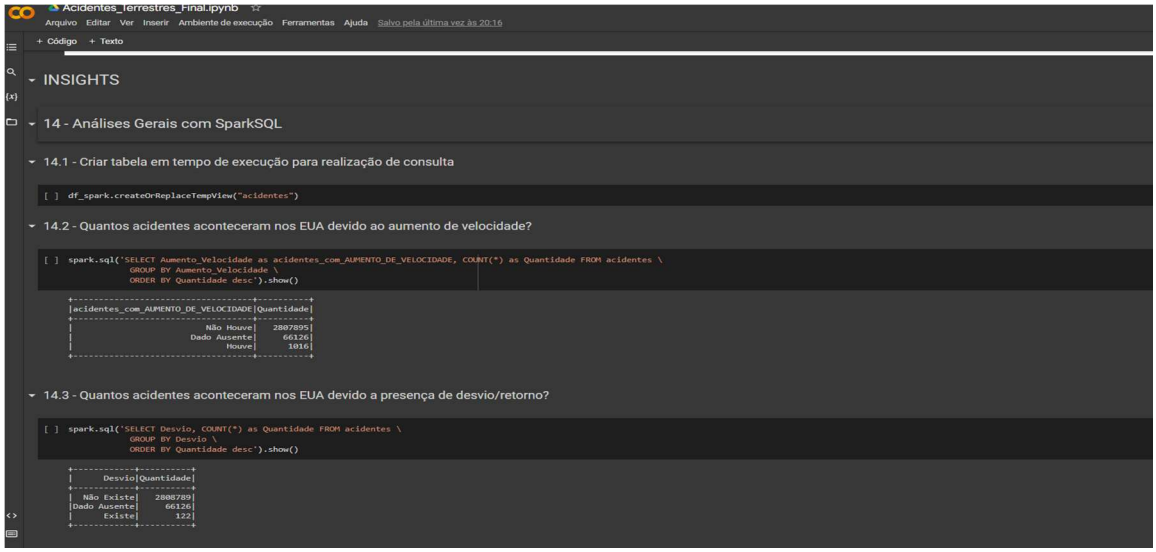
7. Operações usando PySpark para o tratamento e normalização dos dados

- Conforme notebook Acidentes_terrestres item 11



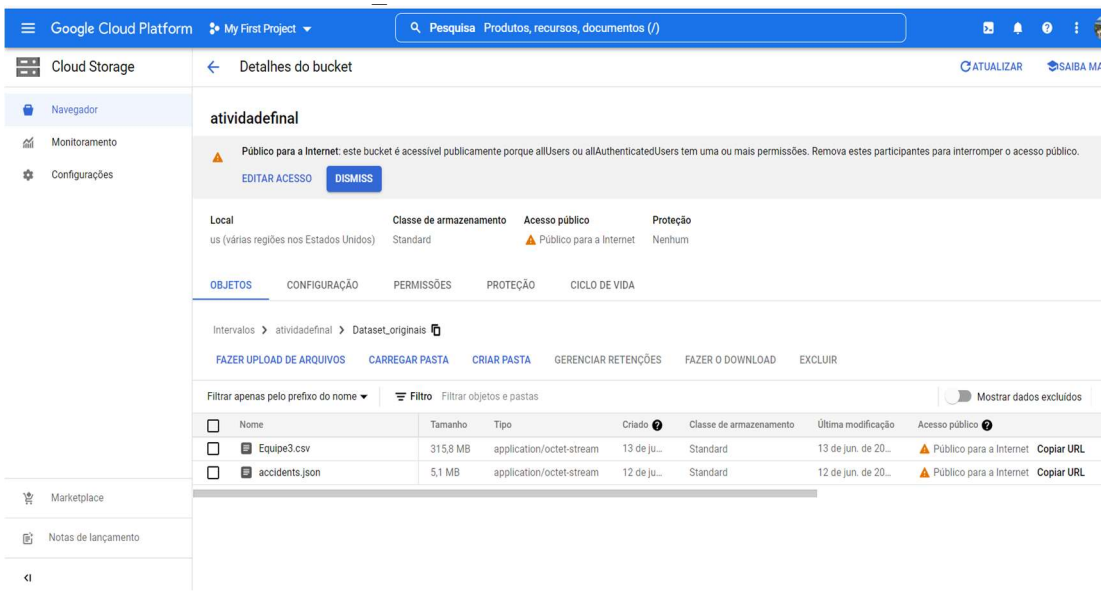
8. Operações usando SparkSQL com a descrição de cada uma das operações

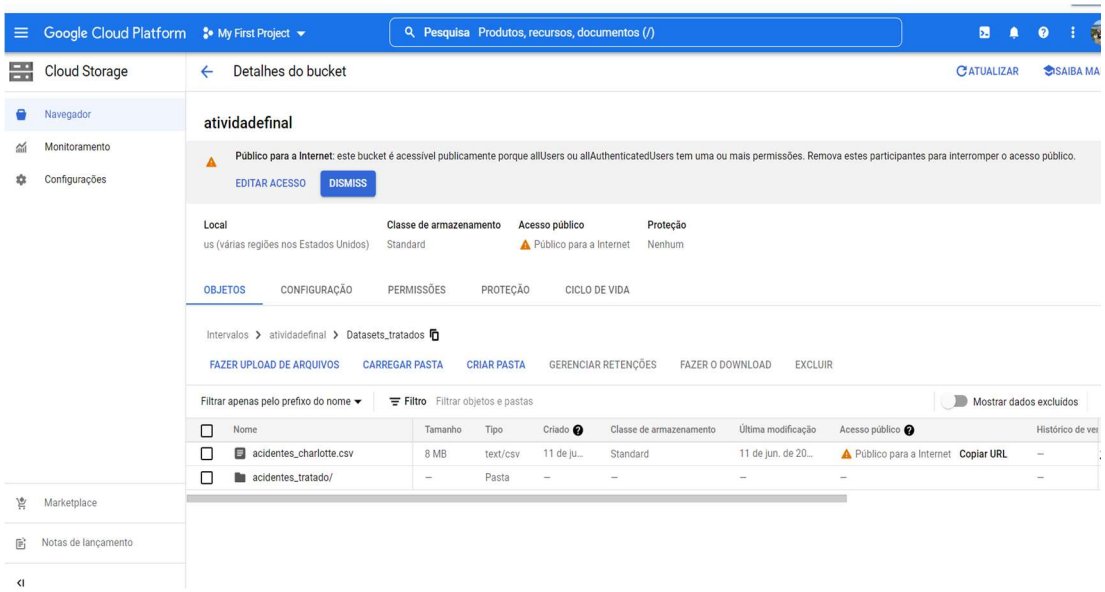
- Conforme notebook Acidentes_terrestres item 14



9. Datasets originais salvos e operados no GCP

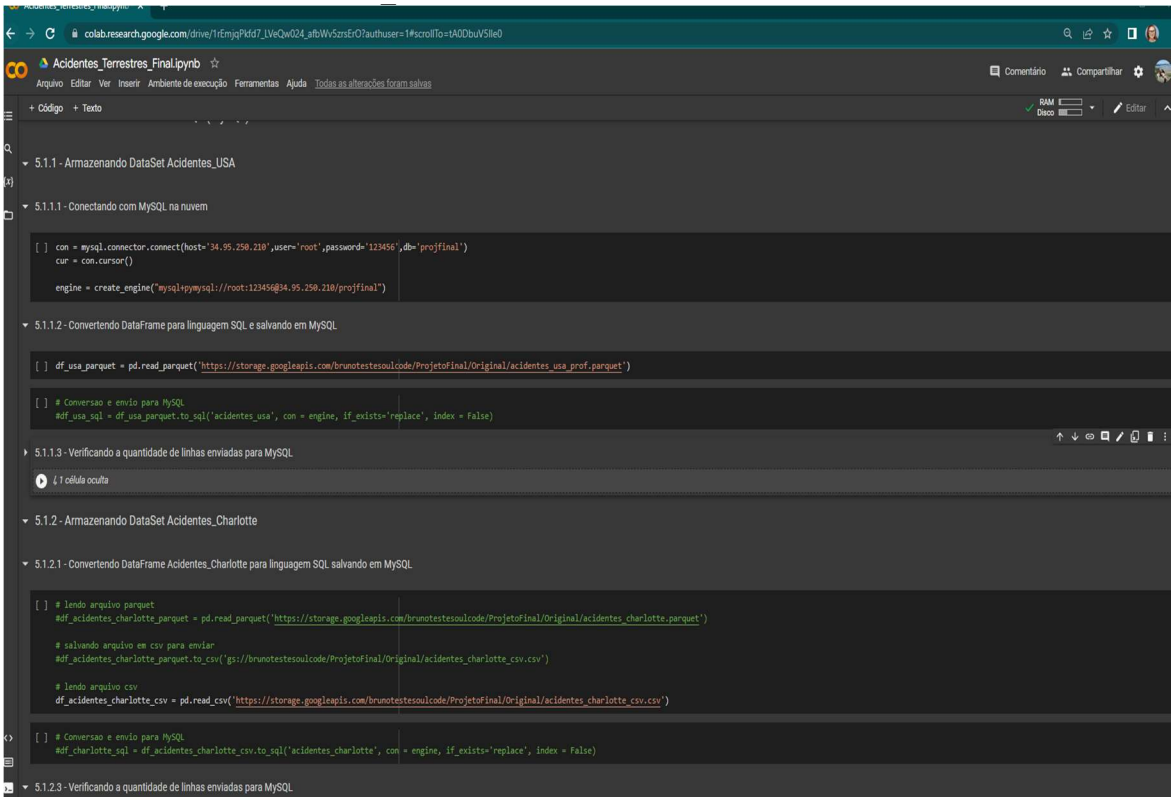
- Conforme notebook Acidentes_terrestres itens 4 e 11.2





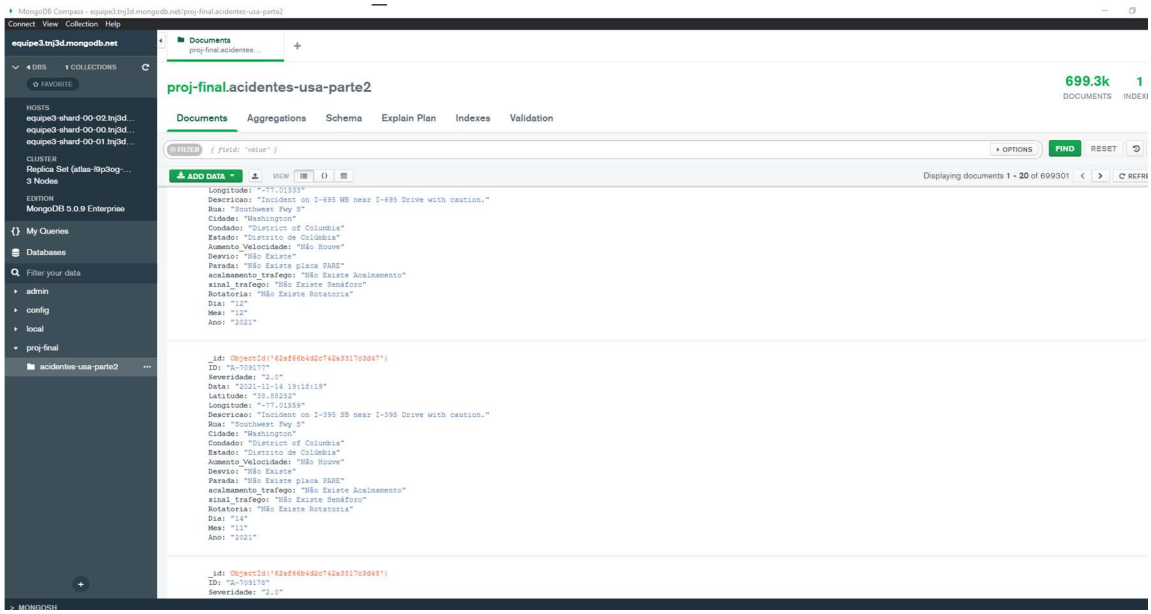
10. Datasets originais armazenados em MySql

- Conforme notebook Acidentes_terrestres item 5



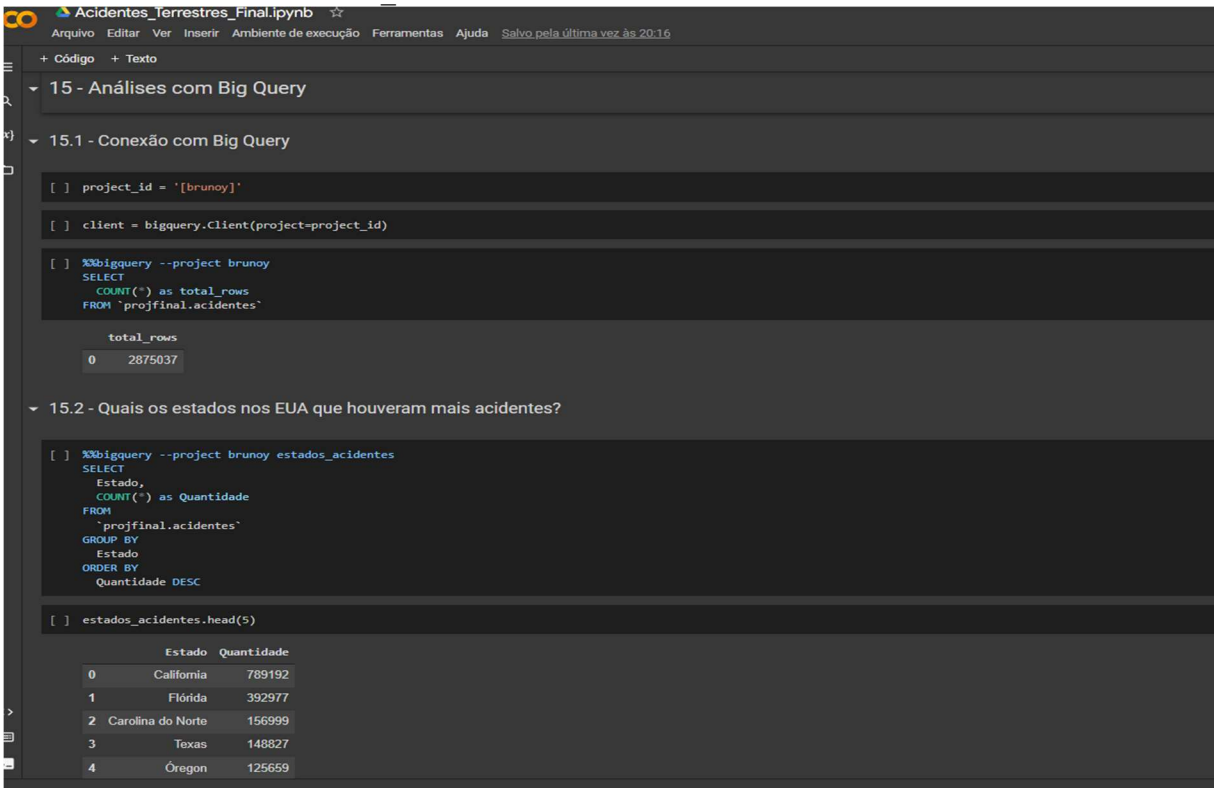
11. Os Dataframes resultantes inseridos na coleção mongodb Atlas

- Conforme notebook Acidentes_terrestres item 12



12. Análises dentro do Big Query

- Conforme notebook Acidentes_terrestres item 15

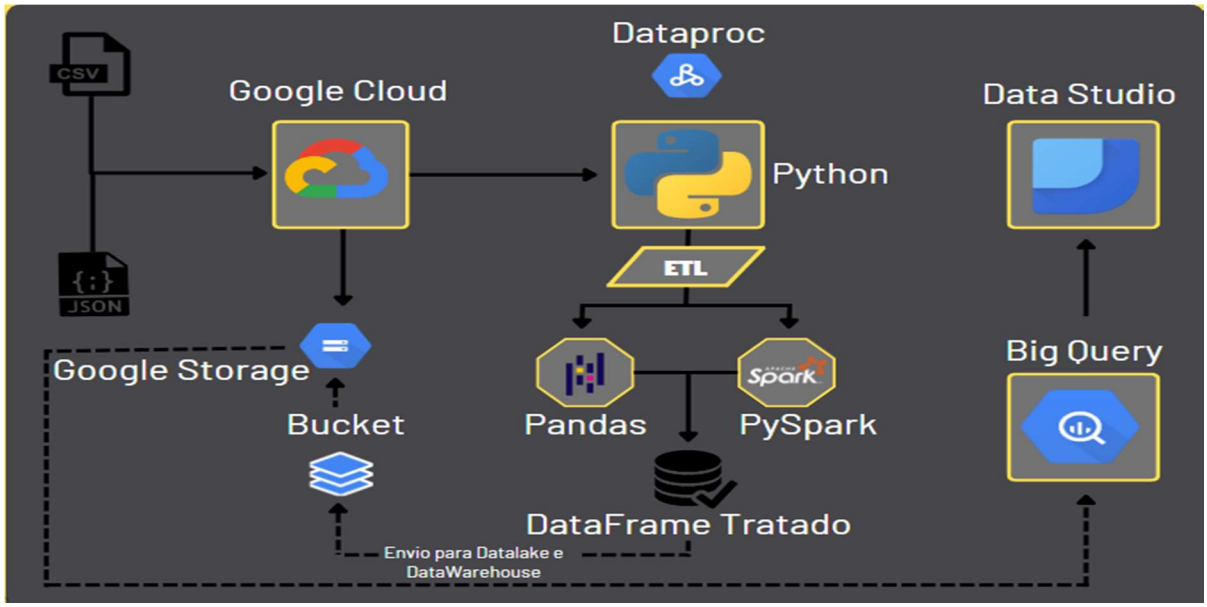


13. O Dashboard no DataStudio poderá ser acessado pelo link:

https://datastudio.google.com/u/1/reporting/dc8c47ad-927a-40f0-a906-604faf2ecd53/page/DANvC

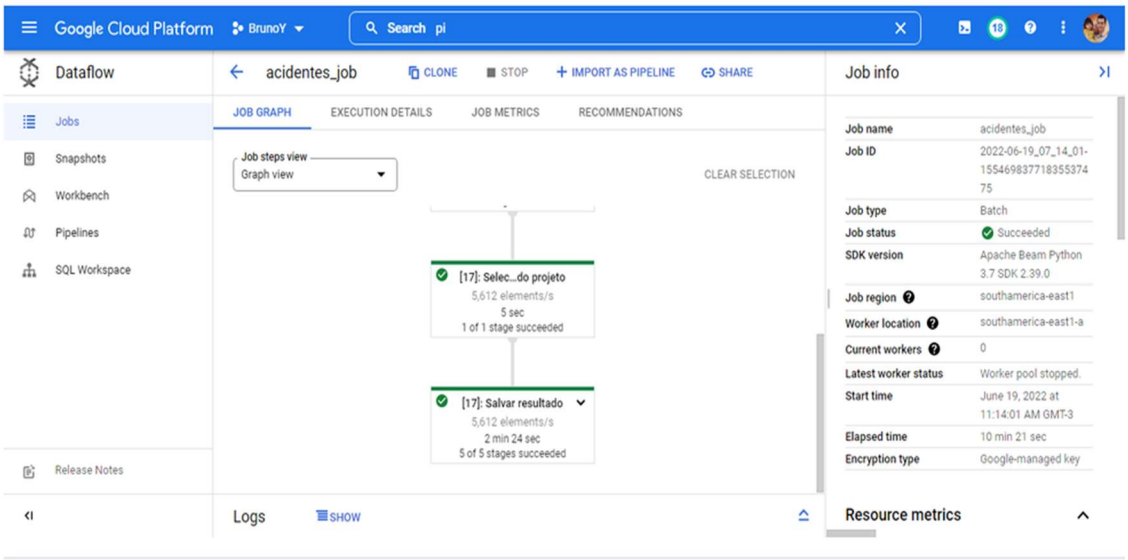
14. Workflow com as etapas de ETL e as respectivas ferramentas utilizadas

- Conforme anexo



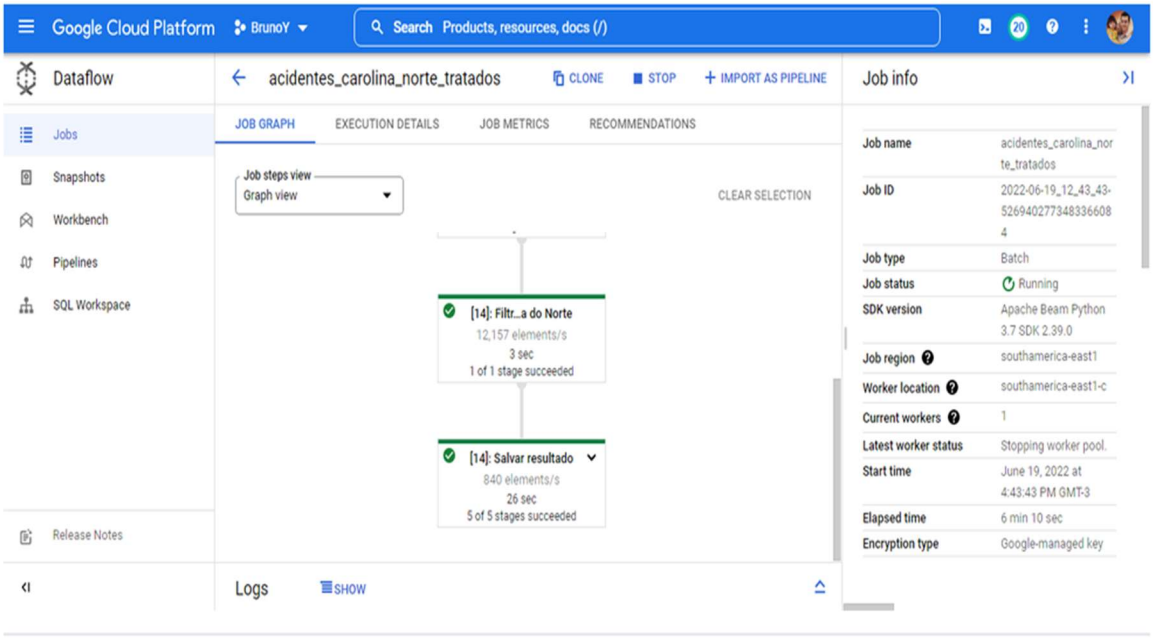
15. Pipeline com modelo criado em apache beam usando dataflow para work

- Conforme notebook Acidentes_terrestres item 6



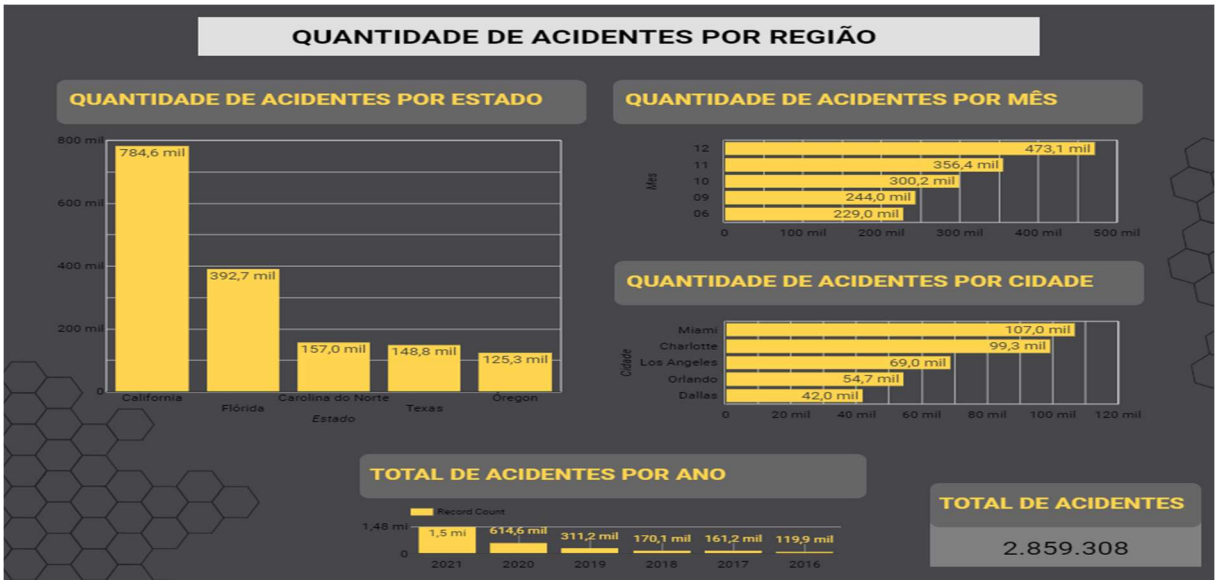
16. Pipeline dos dados normalizados

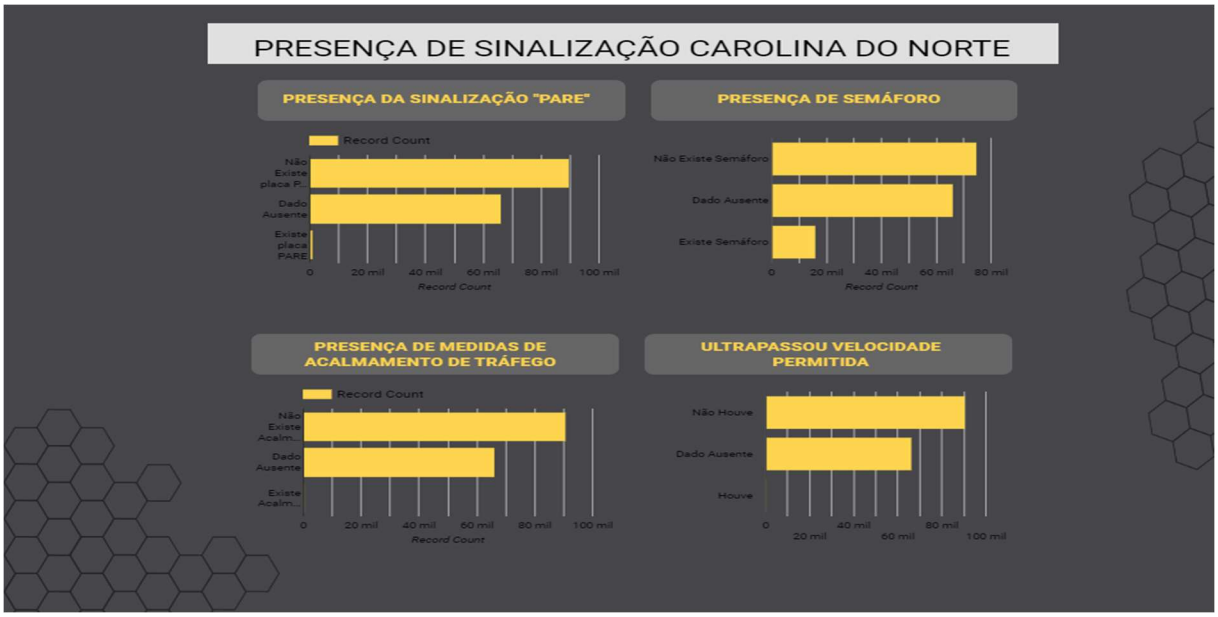
- Conforme notebook Acidentes_terrestres item 3



Insights

A empresa Soul Traffic Company, especialista em sinalização de transito está estudando a possibilidade de abrir uma nova filial na cidade de Charlotte (Estados Unidos). E para a tomada dessa decisão, empresa entrou em contato com a Equipe3 para assessorar o time de dados na normalização e geração de instghs por meio de um processo de ETL. Após os trabalhos de normalização, a Equipe3 encaminhou os seguintes insights:





Também sugere as seguintes ações:

- Redução da velocidade
- Educação no transito (conscientização do público adulto)
- Estruturas adequadas
- Campanhas educativas nas escolas