PROJETO FINAL

Acidentes Terrestres

Soul Code - Turma BC17

Engenharia de Dados

INTEGRANTES

Ana Paula Guimarães

Andréa Goulart

Bruno Yaporandy

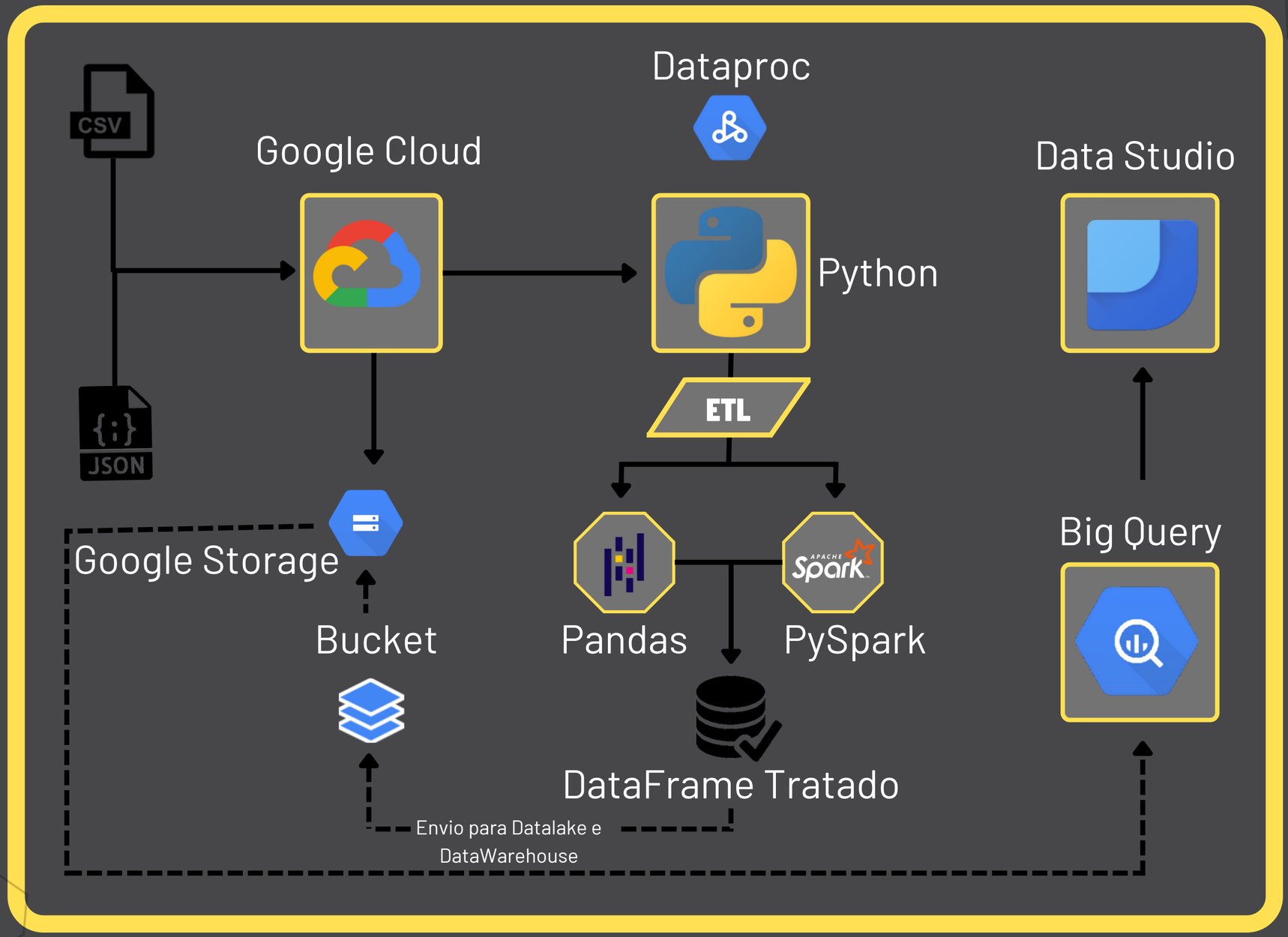
Carlos Dudas

ÍNDICE

DESCRIÇÃO DO PROJETO

A Soul's Traffic Company, uma empresa de sinalização de trânsito está estudando abrir uma nova filial na cidade de Charlotte (Estados Unidos), para auxiliar essa decisão, nossa equipe foi contratada para normalizar 2 DataSets (2016 até 2021) e gerar insights que contribuam com a equipe de dados para tomar essa decisão. Com o objetivo de normalizar os dados e gerar insights através de um processo de ETL.

WORKFLOW



Trello

TECNOLOGIAS

GCP

Dataproc

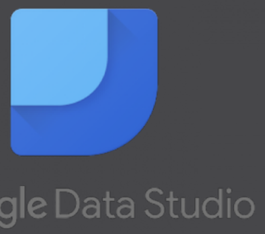
Jupyter
(Python)

Pandas

MySQL

PySpark/SparkSQL

MongoDB



Data Studio

ETAPAS DO PROCESSO DE ETL

Extract, Tranform and Load (ETL)

Extrair os dados de um banco de dados em cloud

Transformações, correções e padronizações

Carregar os dados tratados para uma base cloud

ETAPAS DO PROCESSO DE ETL

Extract — Transform — Load

- Extração do primeiro dataset (accidents_usa.csv)

```
df = pd.read_csv('path/accidents_usa.csv')
```

- Extração do segundo dataset (accidents_charlotte.json)

```
df =  pd.read_json('path/accidents.json', orient='records')
```

- Datasets originais salvos em banco MySQL

```
df = df.to_sql('tabela', conexão = engine, if_exists='replace', index = False)
```

ETAPAS DO PROCESSO DE ETL

Extract — Transform — Load

- Remoção de colunas que não condizem com o objetivo do projeto

```
df.drop(['Coluna'], axis=1, inplace=True)
```

- Tradução dos datasets para PT-BR

```
df.rename(columns={'Street': 'Rua'}, inplace=True)
```

- Procurando por inconsistências

```
pd.unique(df['Coluna'])
```

ETAPAS DO PROCESSO DE ETL

Extract — Transform — Load

- Correção de dados ausentes

```
df['Coluna'] = df['Coluna'].replace(np.nan, 'dado ausente')
```

- Validação dos dados com Pandera

```
schema.validate(df)
```

- Somando os dois datasets (Merge)

```
df = pd.merge(df1, df2, how="outer", on=["Coluna1", "Coluna2"...])
```

ETAPAS DO PROCESSO DE ETL

Extract — Transform — Load

- Montagem do DataFrame utilizando StructType

```
esquema = (StructType([StructField('ID', StringType(), False)...]))
```

- Normalização de colunas devido ao Merge

```
df = df.withColumn('Cidade', regexp_replace('Cidade', 'merge', 'Charlotte'))
```

- Salvando o DataFrame tratado no Bucket para dar inicio ao Load

```
df.repartition(1).write.format("parquet").option("header".....
```

ETAPAS DO PROCESSO DE ETL

Extract — Transform — Load

- Transformando o DataFrame em dicionário

```
df_dicio = df.to_dict("records")
```

- Inserindo o dicionário na coleção mongoDB atlas

```
collection.insert_many(df_dicio)
```

- Início aos insights

SparkSQL e Big Query em conjunto com Data Studio

PIPELINES

- Pipeline 01: Removendo colunas fora do escopo do projeto

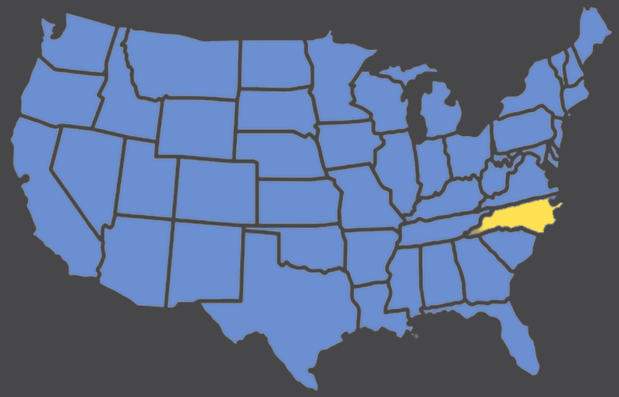
```
| 'Leitura do dataset'>> beam.io.ReadFromText('path', skip_header_lines=1)
| 'Indicando o separador do arquivo'>> beam.Map(record.split(','))
| 'Selecionando Colunas do projeto'>> beam.Map(Seleção das Colunas)
| 'Salvar resultado'>> beam.io.WriteToText('path', file_name_suffix='.csv')
```

DataFrame Inicial

DataFrame Após Pipeline

-Pipeline 02: DataFrame Filtratado por Estado (Carolina do Norte)

```
| 'Leitura do dataset'>> beam.io.ReadFromText('path', skip_header_lines=1)
| 'Indicando o separador do arquivo'>> beam.Map(record.split(','))
| 'Filtragem de colunas'>> beam.Filter( [9] == 'Carolina do Norte')
| 'Salvar resultado'>> beam.io.WriteToText('path', file_name_suffix='.csv')
```



DataFrame Filtrado pelo Estado

TRIGGERS/PROCEDURES

VALIDAÇÃO DE DADOS

TRIGGERS

SEMPRE QUE HOUVER INSERT -
SEMPRE QUE HOUVER UPDATE-
SEMPRE QUE HOUVER DELETE-

```
CREATE TRIGGER tr_auditoria_usa
BEFORE INSERT/UPDATE/DELETE ON acidentesusa
FOR EACH ROW
INSERT/UPDATE/DELETE INTO acidentes_auditoria
(auditoria)
VALUES (CONCAT('Foi inserido o ID:', new.ID));
```

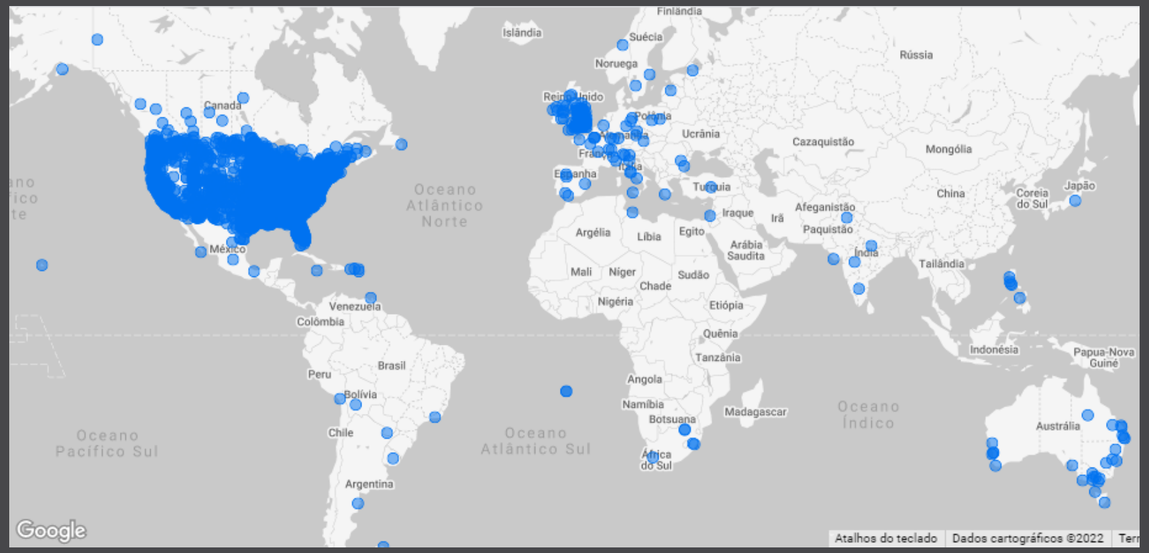
INTEGRIDADE DOS DADOS

PROCEDURE

CONTAGEM DE LINHAS
ATRAVES DO COUNT(*)

```
CREATE PROCEDURE Todos_os_acidentes()
SELECT COUNT(*) FROM acidentesusa;
```

PARTICULARIDADES



- Removendo dados de cidades não pertencentes ao nosso dataset

```
df_usa.drop(df_usa[df_usa.Cidade == 'Nome_da_Cidade'].index, inplace=True)
```

FORMATO DOS DATASETS

66 MIL LINHAS

6 COLUNAS

2.8 MILHÕES DE LINHAS

16 COLUNAS

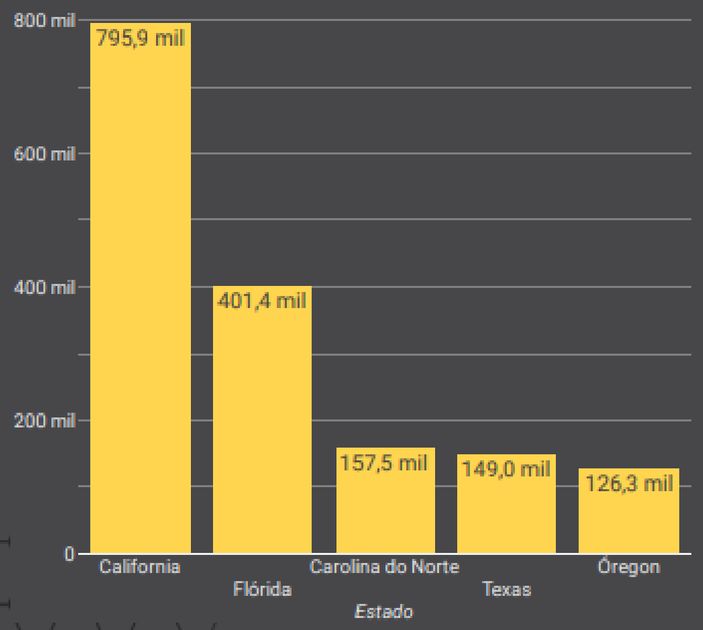
DF-MERGE: (2.8 MILHÕES, 16)

INSIGHTS

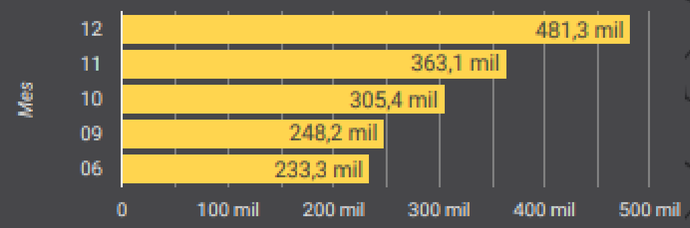
Google Data Studio

QUANTIDADE DE ACIDENTES POR REGIÃO

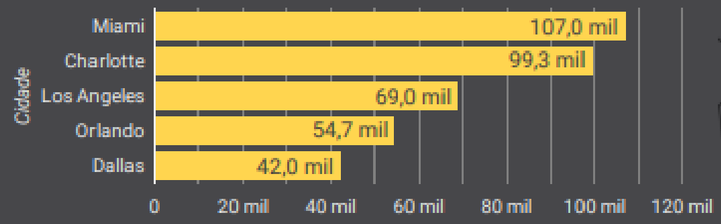
QUANTIDADE DE ACIDENTES POR ESTADO



QUANTIDADE DE ACIDENTES POR MÊS



QUANTIDADE DE ACIDENTES POR CIDADE



TOTAL DE ACIDENTES

2.911.468

CONCLUSÃO

Sugestões:

A redução da velocidade

Educação no trânsito

Estruturas adequadas

Campanhas educativas

AGRADECIMENTOS

CONTATO

https://www.linkedin.com/in/ana-paula-guimar%C3%A3es-ribeiro-36559a123/

https://www.linkedin.com/in/andreacgoulart/

https://www.linkedin.com/in/brunoyaporandy/

https://www.linkedin.com/in/carlos-dudas/

SOUL CODE

TNT