

BleedWatch

“Antecipando Problemas,
Salvando Recursos.”

Grupo 06 - AeroGuardians

Gustavo Ferreira, Gustavo Pereira, Henrique Marlon, Jackson Wellington, Kil Teixeira, Luca Giberti, Lyorrei Shono.



Sumário

1. Recapitulação da Sprint 1
2. Exploração dos dados - Fase de ETL
 - a. Extração dos dados
 - b. Transformações dos dados
 - c. Carregamento dos dados
3. Demonstração
4. Próximos passos

Sprint 1

Objetivo

Entendimento do negócio, Ciclo de Produção, Requisitos e Viabilidade Técnica e Arquitetura dos dados

Como?



Personas, User Stories e Jornada do Usuário



Matriz de Risco e Canvas Value Proposition



Exploração dos dados e Pré-análise das features



Arquitetura da solução e Viabilidade Técnica

E na sprint 2?



Exploração dos Dados

O que foi realizado até agora?

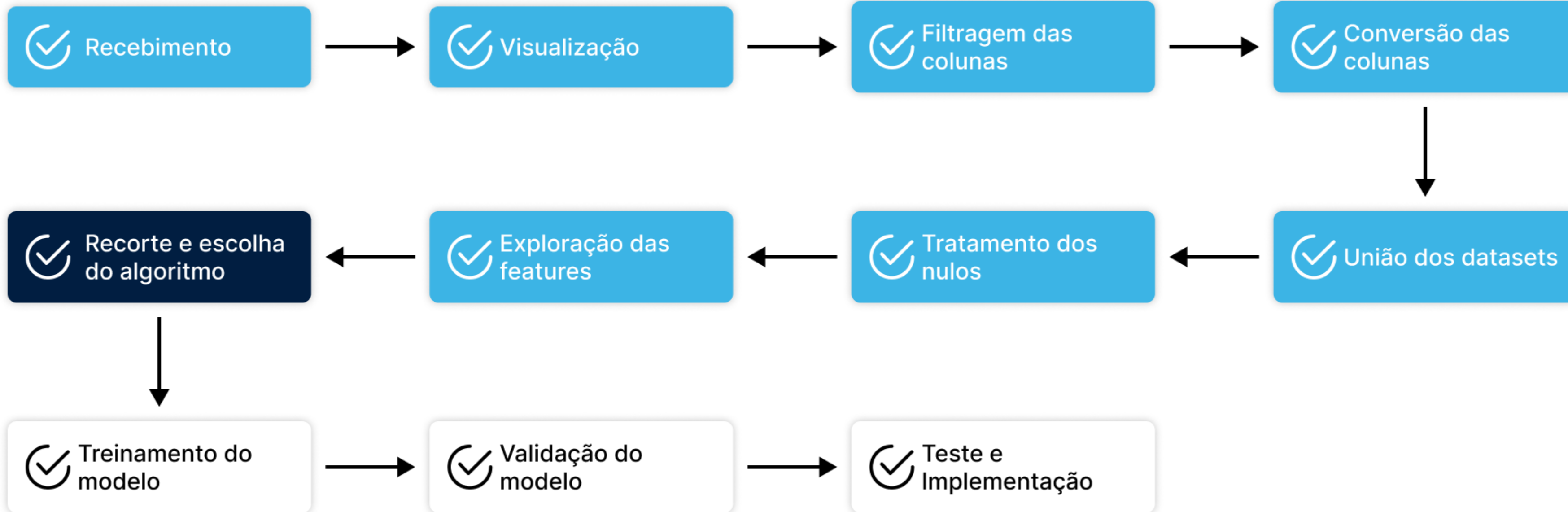
Concluído



Em progresso



Não iniciado



Nota: Utilização de pesquisas científicas para exploração das features

Sprint 1

Dados

Processo de ETL

Proximos passos

Processo de ETL

Objetivo

Qual é o objetivo principal do projeto?



Coletar os dados de várias fontes, transformá-los em um formato adequado e carregá-los em um lugar de armazenamento de dados.

Padronização dos dados de diferentes fontes

Maior qualidade dos dados do que da fonte inicial



Processo de ETL - Extração

Extração

Nesta fase, os dados são coletados de diversas fontes, como bancos de dados, arquivos CSV, APIs ou sistemas externos. A extração envolve a recuperação dos dados brutos.

1. Extrair dos arquivos zips os datasets
2. Visualização dos arquivos "PARQUET"

Download dos Dados

- FASE 4: Realização do download dos dados da base para esse notebook do Colab, para a visualização e manipulação.

```
[ ] 1 # Download dos dados para esta máquina virtual
    2 downloaded = drive.CreateFile({'id': '1nYYM_5xH6bWL5gPqBTx_mt0SCzm4EiuU'})
    3 downloaded.GetContentFile("06120018.zip")
    4 shutil.unpack_archive("06120018.zip", "./dataset")
```

Nota: Todos os notebooks Jupyter possuem autenticação de usuário.



Processo de ETL - Transformação

Transformação

Os dados extraídos geralmente estão em formatos e estruturas variadas. Nesta etapa, os dados são limpos, filtrados, enriquecidos e transformados em um formato padronizado e adequado para análise.

1. Transformação dos nulos em zeros
2. Remoção de dados duplicados
3. Agrupamento dos datasets em uma linha por meio da somatória, mediana, mínimo e máximo de algumas colunas.



Processo de ETL - Transformação

```
1  try:
2      for subdir, _, files in os.walk(directory):
3          for file in files:
4              file_path = os.path.join(subdir, file)
5              df = pd.read_parquet(file_path)
6              df = df[filtered_cols]
7              aggregated_data = df.agg(agg_methods)
8              aggregated_data.to_frame().T.to_csv(output_file_path, mode='a', header=False, index=False)
9  except:
10     print('New parquet not found')
```



Processo de ETL - Carregamento (load)

Carregamento (load)

Depois de transformados, os dados são carregados no destino final, que pode ser um armazém de dados, um data lake ou outro sistema de armazenamento. Esse processo garante que os dados estejam prontos para serem utilizados por equipes de análise, geração de relatórios e tomada de decisões.

1. Através de uma função em python, os dados são lidos.
2. Por meio do orm prisma os dados são coletados e enviado para um banco de dados.
3. O banco dados desenvolvido em postgres, salva todas as informações.



Processo de ETL - Carregamento (load)

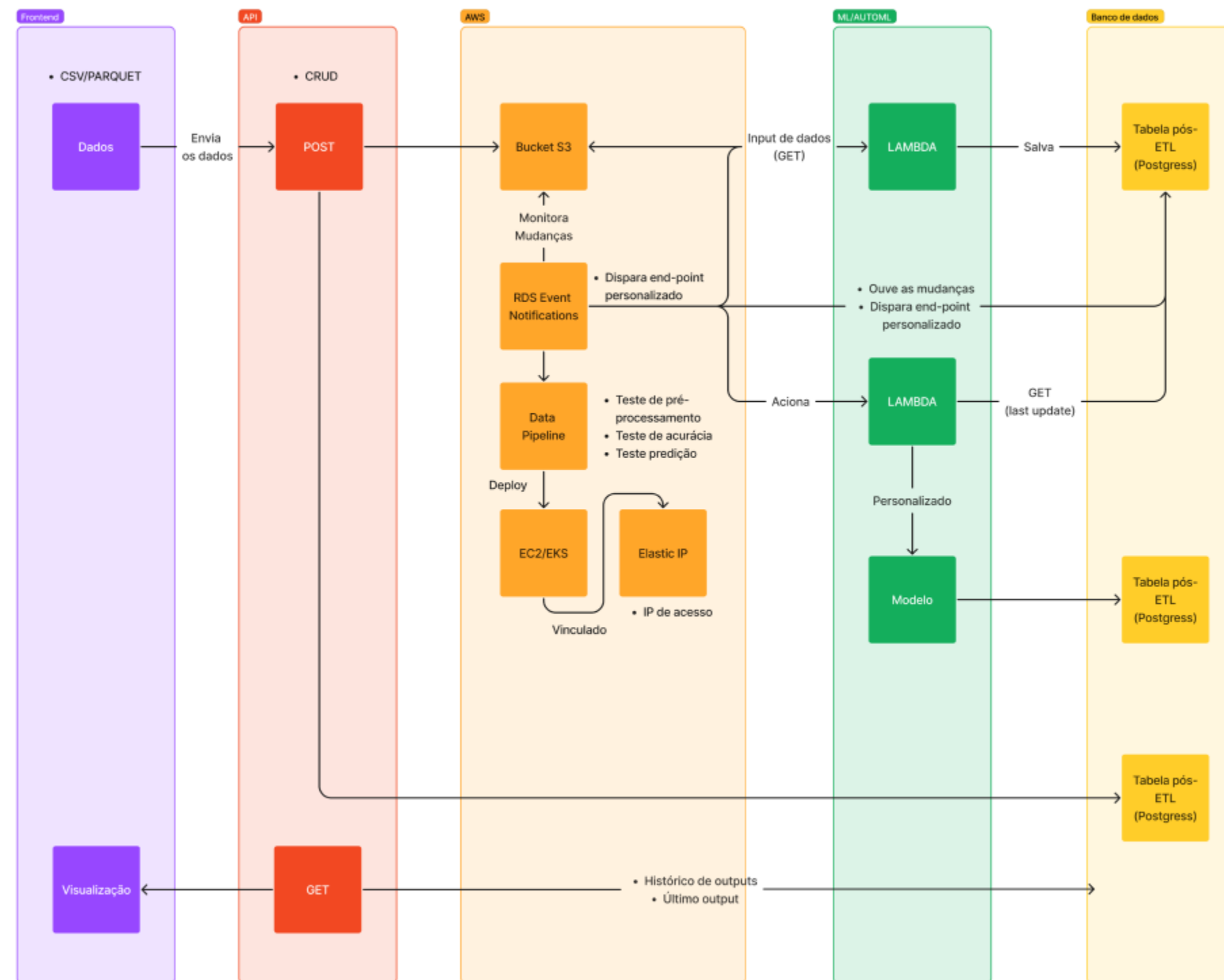
```
1 import pandas as pd
2 from datetime import datetime
3 from prisma import PrismaClient
4
5 prisma = PrismaClient()
6 df = pd.read_csv('C:\Users\Inteli\Documents\GitHub\grupo6\src\database\output.csv')
7 new_data = df.to_dict(orient='records')
8
9 def load_csv_to_db(data_list):
10     for data in data_list:
11         data['createdAt'] = datetime.now()
12         prisma.airplaneDataPredicted.create(**data)
13
14 if __name__ == '__main__':
15     # Chamar a função de carregamento de dados
16     load_csv_to_db(new_data)
17
18
```

Hora da demonstração



Arquitetura de Dados

Como funciona o fluxo?



Sprint 1

Dados

Processo de ETL

Proximos passos



Próximos passos



Recorte e escolha dos novos datasets disponibilizados



Configuração inicial da AWS



Treinamento do modelo com os dados pré-processados



Desenvolvimento de novas features no Dashboard



Agradecemos pela atenção!

“Antecipando Problemas,
Salvando Recursos.”