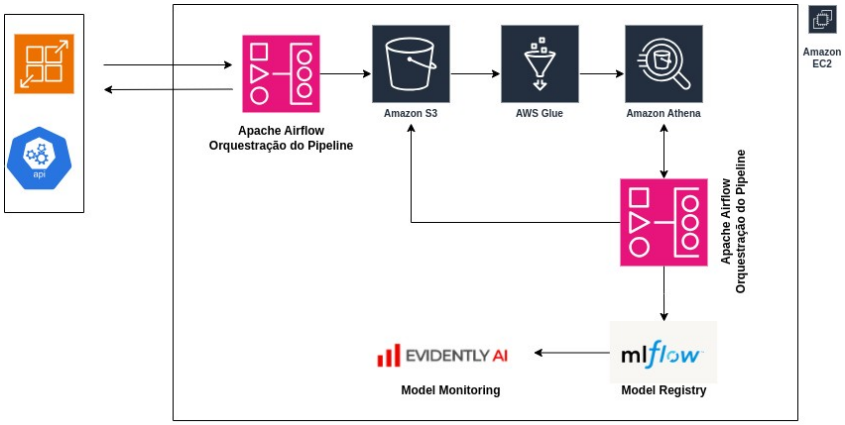
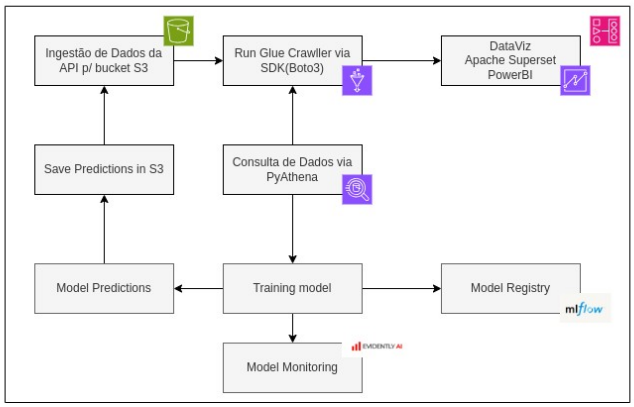
Documentação do Pipeline de Ingestão de Dados

Visão Geral

Este pipeline foi projetado para a ingestão, processamento e monitoramento de dados, utilizando uma série de tecnologias de AWS e ferramentas de orquestração e visualização. O fluxo de dados começa com a ingestão de dados de uma da API construída para consumir os dados de vitivinicultura, segue com a construção de um modelo de série temporal que por sua vez segue uma esteira de monitoramento.

Componentes e Fluxo de Trabalho





1. Ingestão de Dados da API para o Amazon S3

- **Descrição:** Os dados são ingeridos da API construída e armazenados em um bucket do Amazon S3. Neste ponto, podemos tanto consumir pela API quanto conectar no banco de dados que construímos na estrutura do projeto para fazer ingestão dos dados.
-
- **Tecnologias Utilizadas:**
    - API (Fonte dos dados)

- Amazon S3 (Armazenamento de dados)

2. AWS Glue

- **Execução do Glue Crawler via SDK (Boto3):**
  - **Descrição:** Um Crawler do AWS Glue é executado para catalogar os dados armazenados no S3. Isso cria ou atualiza o catálogo de dados, permitindo a consulta dos dados.
  - **Tecnologias Utilizadas:**
    - AWS Glue
    - Boto3 (SDK para Python)

3. Amazon Athena

- **Consulta de Dados via PyAthena:**
  - **Descrição:** Utilizando o PyAthena, uma biblioteca para Python, são realizadas consultas nos dados catalogados pelo AWS Glue.
  - **Tecnologias Utilizadas:**
    - Amazon Athena
    - PyAthena

4. Orquestração do Pipeline

- **Apache Airflow:**
  - **Descrição:** Apache Airflow é usado para orquestrar o fluxo de dados em todo o pipeline. Ele coordena as etapas de ingestão, processamento, e consultas de dados, além de iniciar o pipeline treinamento de modelos.
  - **Tecnologias Utilizadas:**
    - Apache Airflow

5. Treinamento de Modelos

- **Descrição:** Os dados processados são utilizados para o treinamento de modelos de machine learning.
- **Tecnologias Utilizadas:**
  - Mlflow
  - Scikit-learn
  - Statsmodels

6. Previsões e Armazenamento

- **Model Predictions:**
  - **Descrição:** O modelo treinado é usado para realizar previsões, que são então armazenadas no Amazon S3. Os detalhes do modelo que pode ser construído está no arquivo Modelo_Series_Temporais.pdf.
  - **Tecnologias Utilizadas:**
    - Amazon S3
    - MLFLOW

7. Model Registry

- **Descrição:** Os modelos treinados são registrados utilizando o MLflow, que gerencia as versões dos modelos.
- **Tecnologias Utilizadas:**
    - MLflow

8. Model Monitoring

- **Descrição:** O monitoramento dos modelos é realizado utilizando o Evidently AI, que verifica a performance e outros aspectos do modelo.
- **Tecnologias Utilizadas:**
    - Evidently AI

9. Visualização de Dados

- **Descrição:** A visualização de dados é feita utilizando ferramentas como Apache Superset ou Power BI, permitindo uma análise e exploração dos dados processados.
- **Tecnologias Utilizadas:**
    - Apache Superset
    - Power BI

Considerações Finais

Este pipeline permite uma ingestão de dados automatizada e escalável, com monitoramento contínuo e gerenciamento de modelos. A combinação das ferramentas de AWS e outras tecnologias permite uma solução robusta para o processamento dos dados de vitivinicultura e a construção de um modelo de série temporal.