Documento de Arquitetura

PROJETO DE PIPELINE DE DADOS DE QUEIMADAS

1. Diagrama do Pipeline de Dados Atual (Ingestão, Armazenamento e Transformação)

O *pipeline* de dados atual, evidenciado pelo conteúdo do *notebook*, representa um estado de **processamento avançado e acelerado** em um ambiente de desenvolvimento (Google Colab).

O fluxo inicia com a **Ingestão**, onde os dados de focos de queimadas (originalmente CSV) já foram coletados e submetidos a uma etapa prévia de limpeza, sendo persistidos no formato otimizado **Parquet**.

O **Armazenamento** é realizado com os dados já no formato Parquet, simulando uma transição da camada Bronze para a **Silver** dentro de um Data Lake. Este formato é fundamental para o próximo passo, que é a aceleração, permitindo que as bibliotecas de Big Data leiam grandes volumes de forma eficiente.

A **Transformação e Processamento Acelerado** é o foco atual e representa a fase mais complexa do *pipeline*. O processo utiliza o **Dask-cuDF** para ler e manipular o *dataset* de forma distribuída. As transformações e operações analíticas (como seleção de colunas, filtragem e cálculo de médias) são realizadas com o poder de aceleração da **GPU (NVIDIA Tesla T4)**, comprovando a viabilidade de processamento intensivo de Big Data.

2. Tecnologias Utilizadas e Sugestões de Refinamento

Tecnologias Atuais e Ambiente: O *pipeline* evoluiu de um mero processamento em memória com Pandas para uma arquitetura de aceleração computacional. As tecnologias atualmente em uso são: **Python** como linguagem principal; bibliotecas **cuDF** e **Dask-cuDF** da suíte RAPIDS para manipulação e processamento distribuído de dados com aceleração **GPU**; e o formato **Parquet** como padrão de armazenamento. O ambiente é o **Google Colab** configurado especificamente com *runtime* **GPU (Tesla T4)**.

Sugestões de Refinamento (Tecnologias Paga/Gerenciadas e de Escala) e Justificativa: A escalabilidade para produção requer a migração do *Proof of Concept* (PoC) para uma infraestrutura na nuvem **AWS**. O foco é manter a capacidade de processamento em larga escala e acelerada, utilizando serviços gerenciados e sem a complexidade do MapReduce tradicional:

1. **Armazenamento e Data Lakehouse (AWS S3 & Redshift/Athena):**
    - O armazenamento central (Data Lake) deve ser persistido e escalado no **Amazon S3 (Simple Storage Service)**, que oferece durabilidade e baixo custo para armazenar os arquivos Parquet.
    - Para consultas analíticas de alto desempenho e a implementação da arquitetura *Lakehouse*, o destino deve ser o **Amazon Redshift** (para Data Warehouse) ou

o **Amazon Athena** (para consultas *serverless* diretamente no S3, ideal para o formato Parquet).

2. **Orquestração e Automação (AWS Step Functions & Managed Airflow):**
    - A execução manual deve ser substituída por um orquestrador. A **AWS Step Functions** permite definir e gerenciar fluxos de trabalho complexos e *serverless*, integrando-se nativamente com outros serviços AWS.
    - Alternativamente, pode-se usar o **Amazon Managed Workflows for Apache Airflow (MWAA)** para manter a compatibilidade com a sintaxe Airflow e o agendamento robusto de *workflows*.
3. **Processamento Acelerado e em Escala (AWS EMR & SageMaker/ECS):**
    - Para processamento distribuído, o **Amazon EMR (Elastic MapReduce)** deve ser configurado para rodar *clusters* **Apache Spark** (desativando o uso de MapReduce). Isso garante a capacidade de processar grandes volumes de dados de forma escalável.
    - Para manter a aceleração por GPU em produção (Dask-cuDF/RAPIDS), a solução é utilizar o **Amazon SageMaker** (para notebooks gerenciados de alto desempenho) ou o **Amazon ECS (Elastic Container Service)**, executando *containers* em instâncias **EC2** configuradas com GPUs (como as instâncias P3/P4), fornecendo o ambiente necessário para o processamento de alto desempenho do cuDF.

3. Arquitetura Parcial Implementada (Ambiente Simulado)

A arquitetura parcial implementada é um **PoC de Aceleração Computacional** sobre a camada Silver do *Data Lake*.

Este PoC demonstra a capacidade de:

1. **Ambiente GPU Validado:** Configurar e validar a funcionalidade do ambiente com GPU (Tesla T4) e as bibliotecas RAPIDS (**cuDF/Dask-cuDF**).
2. **Processamento Otimizado:** Comprovar a viabilidade de realizar operações intensivas de transformação (como seleção e filtragem) diretamente em dados Parquet, obtendo um *throughput* muito superior ao do Pandas tradicional.
3. **Viabilidade de Big Data:** O sucesso na execução de operações distribuídas e aceleradas com Dask-cuDF valida o caminho técnico para processar volumes de dados significativamente maiores (Big Data) na próxima fase do projeto.

4. Equipe Responsável e Divisão de Tarefas

Julio Padilha atua como **Engenheiro de Dados (Otimização e Escalabilidade do Pipeline)**. Suas responsabilidades centrais focaram na expansão do escopo de dados, realizando a concatenação e integração de arquivos para expandir o *dataset* de 1 para 22 meses. Mais importante, ele foi o arquiteto da alta performance, implementando soluções avançadas de otimização de desempenho e memória com as bibliotecas **Dask e cuDF**, garantindo o aproveitamento do processamento paralelo e o uso da GPU do Colab para Big Data.

Roberto Arruda é o **Cientista de Dados (Ingestão e Modelagem)**. Sua contribuição inicial foi fundamental para estruturar a base do projeto, realizando a ingestão inicial dos dados e organizando o *pipeline* conceitualmente nas camadas Bronze, Silver e Gold. Ele é o responsável pela arquitetura inicial de pastas, pela padronização do fluxo de dados e pela criação das primeiras transformações essenciais entre as camadas.

Nicole Victory assume o papel de **Analista de Dados (Validação e Qualidade dos Dados)**. Seu foco está em garantir a confiabilidade do *dataset*. Ela criou e executa *scripts* rigorosos para verificação e limpeza dos *datasets* após a ingestão na camada Bronze, assegurando que os arquivos finais contenham as colunas esperadas e estejam livres de valores nulos críticos. Além disso, é a responsável por gerar relatórios automáticos de estatísticas e qualidade dos dados (*profiling*), essenciais para a documentação e a tomada de decisões analíticas.

Matheus Bione oferece **Suporte Técnico** essencial ao projeto. Suas tarefas envolvem a manutenção da organização e integridade do projeto, verificando continuamente se os dados transformados mantêm a fidelidade em relação à ingestão original. Ele também realizou a organização física de diretórios, a limpeza de arquivos duplicados e a padronização de nomes dentro da estrutura do projeto.