**Relatório de desenvolvimento [Desafio Técnico - Engenheiro de Dados – IMAFLORA]**

**Responsável:** Maria Luiza Torres Pires

# OBJETIVO DO PROJETO

O objetivo deste teste é desenvolver um processo de Extração, Transformação e Carga (ETL) para dados de desmatamento do bioma Cerrado, utilizando os dados do *GeoServer* do *TerraBrasilis*. O objetivo final foi armazenar os dados processados em um banco de dados PostgreSQL [PostGIS], garantindo a integridade e eficiência das consultas geoespaciais e permitindo análises espaciais de alto desempenho.

# FLUXO DE TRABALHO

O projeto foi dividido em quatro etapas principais:

## EXTRAÇÃO DOS DADOS

* Implementei a classe *GeoServerDownloader* para acessar a API do *GeoServer* do *TerraBrasilis* via WFS;
* A classe permite parametrização do bioma, camada e ano do desmatamento, garantindo flexibilidade para diferentes análises;
* Implementei paginação para garantir que todos os dados fossem baixados corretamente, evitando limitações de resposta do servidor, conforme indicado no teste, e;
* O processo inclui tratamento de erros robusto, com tentativas automáticas em caso de falha de conexão ou *timeout.*

## PROCESSAMENTO DOS DADOS

* Utilizei o geopandas para manipulação e processamento dos dados geoespaciais;
* Garanti a integridade das geometrias:
  + Correção de *GeometryCollection* para *MultiPolygon*, removendo geometrias inválidas;
  + Padronização do SRID para EPSG:4326 (WGS 84) para compatibilidade com análises geoespaciais;
  + Remoção de registros nulos ou inconsistentes, e;
* Os dados foram convertidos para o formato *GeoJSON,* facilitando a integração com o banco.

## CARGA DOS DADOS NO POSTGRESQL

* Criei a estrutura do banco de dados com um *schema* específico (*raw\_data*) para organização dos dados;
* Criei a tabela ‘*deforestation’*, garantindo que as colunas *‘year’, ‘area\_km’ e ‘geometry’* estivessem corretamente tipadas;
* Utilizei o ‘*GeoPandas.to\_postgis()*’ para carregar os dados de forma eficiente no banco, e;
* Criei um índice espacial ‘*GIST’* na coluna *‘geometry’* para acelerar consultas geoespaciais.

## VALIDAÇÃO FINAL E CONSULTAS GEOESPACIAIS

* Confirmei a consistência dos dados após a carga:
  + Total de registros na tabela ‘*deforestation*’: 1000;
  + Total de áreas encontradas com geometrias válidas: 9047;
  + SRID da geometria: EPSG:4326 ‘(validado via ST\_SRID())’;
  + Índice espacial criado: ‘*deforestation\_geom\_idx’*.
* Executei consultas espaciais para identificar os estados mais afetados pelo desmatamento:

SELECT s."SIGLA\_UF", COUNT(\*) AS total\_desmatamento

FROM raw\_data.deforestation d

JOIN raw\_data.states s

ON ST\_Intersects(d.geometry, s.geometry)

GROUP BY s."SIGLA\_UF"

ORDER BY total\_desmatamento DESC;

# DESAFIOS E SOLUÇÕES

## PAGINAÇÃO NO DOWNLOAD DOS DADOS

* O *GeoServer* limita a quantidade de registros retornados por requisição;
* Implementei a paginação usando *‘maxFeatures’* e *‘startIndex’*, garantindo que todas as páginas de dados fossem recuperadas corretamente.

## CORREÇÃO DE GEOMETRIAS INVÁLIDAS

* Alguns registros continham *‘GeometryCollection’*, o que não é suportado diretamente pelo PostgreSQL;
* Dessa forma, criei uma função para converter *‘GeometryCollection’* para *‘MultiPolygon’*, descartando objetos sem relevância, e;
* Alternativamente, poderia utilizar *‘ST\_MakeValid()’* do PostGIS para corrigir automaticamente algumas geometrias inválidas.

## DIFERENTES SRIDS NAS TABELAS

* A tabela ‘*states’* estava com SRID 4674, enquanto os dados de desmatamento estavam em 4326, e;
* Converti os dados de *‘states’* para 4326 usando *‘ST\_Transform()’* antes de realizar o JOIN espacial, evitando erros de referência.

# OTIMIZAÇÕES APLICADAS

* Uso de índice espacial GIST para otimizar consultas geoespaciais;
* Conversão padronizada de SRID para garantir integridade nas análises;
* Implementação de *logging* para rastreamento e depuração do pipeline ETL;
* Armazenamento intermediário em GeoJSON para modularização e integração simplificada, e;
* Validação de integridade dos dados após a carga no PostgreSQL.

# SUGESTÕES PARA MELHORIAS FUTURAS

* Desenvolver um dashboard interativo para visualizar as áreas de desmatamento diretamente no mapa;
* Automatizar a atualização dos dados para incorporação de novos registros periodicamente;
* Implementar compressão e otimização de dados geoespaciais para redução do consumo de armazenamento, e;
* Explorar integração com sistemas de alerta e monitoramento para análise em tempo real.

# CONCLUSÃO

O desafio foi concluído com sucesso, garantindo que:

* Os dados do GeoServer foram extraídos corretamente, mesmo com limitação de requisições;
* As geometrias foram padronizadas e corrigidas, garantindo compatibilidade com o PostgreSQL/PostGIS;
* A carga dos dados foi otimizada, incluindo indexação geoespacial para eficiência nas consultas, e;
* As análises geoespaciais puderam ser realizadas sem inconsistências.