Instituto Nacional de Pesquisas Espaciais — INPE — São José dos Campos, 2026

Programa de Pós-Graduação em Computação Aplicada

Projeto da Disciplina Introdução à Geoinformática (CAP 395-3)

Avaliação de uma Arquitetura Lakehouse para Integração e Análise de Dados Geoespaciais: estudo de caso com PRODES, DETER e TerraClass

Discente: Sandro de Sena Machado

Orientador: Gilberto Ribeiro Queiroz

Co-orientadora: Karine Reis Ferreira

RESUMO

Este trabalho apresenta o desenvolvimento e avaliação de um protótipo de arquitetura Lakehouse para integração e análise exploratória de dados geoespaciais. O projeto integra dados do PRODES (desmatamento consolidado), DETER (alertas de desmatamento) e TerraClass (uso e cobertura da Terra) dos biomas Amazônia e Cerrado no período de 2018 à 2022. A arquitetura implementa a estrutura medalhão (Bronze-Prata-Ouro) utilizando Apache Spark, Apache Sedona e Delta Lake, com armazenamento local no MinIO (*Object Storage*). Os resultados sugerem que o protótipo atende de modo satisfatório à maior parte dos requisitos associados ao paradigma Lakehouse, sobretudo em armazenamento unificado, governança, rastreabilidade e processamento incremental. Entretanto, a experiência também revelou limitações relevantes: dificuldade de paralelização em algumas etapas, gargalos no processamento conjunto de vetores e rasters, restrições de reprojeção e maturidade ainda desigual das ferramentas geoespaciais no ecossistema distribuído. Os cruzamentos temáticos realizados com PRODES, DETER e TerraClass funcionam principalmente como demonstração de capacidade analítica.

Palavras-chave: lakehouse; geoprocessamento; desmatamento; uso e ocupação do solo.

1. Introdução

O desmatamento e as mudanças no uso e cobertura da terra configuram-se como alguns dos maiores desafios ambientais e climáticos contemporâneos. De acordo com o Painel Intergovernamental de Mudanças Climáticas, aproximadamente metade de todas as emissões globais de gases de efeito estufa decorrentes dos setores de agricultura, florestas e outros usos do solo está diretamente associada ao gás carbônico (CO_2) liberado por processos de desmatamento (IPCC, 2023).

Diante desse cenário, consolidaram-se diversas iniciativas ao redor do mundo voltadas à produção de dados geoespaciais sistemáticos para subsidiar a governança ambiental, a conservação da biodiversidade e a mitigação dos impactos antrópicos negativos na natureza. No cenário brasileiro, o Instituto Nacional de Pesquisas Espaciais (INPE) mantém três sistemas complementares de monitoramento da cobertura terrestre: o PRODES, encarregado de computar as taxas anuais consolidadas de desmatamento; o DETER, concebido como um sistema de alertas em tempo quase real para subsidiar a fiscalização ambiental; e o TerraClass, que mapeia de forma bienal o uso e a ocupação do solo (ALMEIDA, C. et al, 2025).

Contudo, por terem sido concebidos de forma independente, existem desafios de integração em virtude da alta variedade semântica, sintática e volumétrica. Historicamente, a harmonização dessas bases para a extração de trajetórias de uso e ocupação do solo — sequências cronológicas de uso do solo em uma mesma localização — tem sido solucionada pelo acoplamento entre Sistemas de Gerenciamento de Banco de Dados (SGBD), com distribuição via servidores de mapas e serviços web como o *Web Land Trajectory System* (ZIOTI, F. et al. 2022).

Este modelo arquitetural baseia-se no paradigma de *Data Warehouses*, que oferece vantagens como suporte à evolução de esquemas de tabelas e garantia de propriedades transacionais, mas apresenta desvantagens relacionadas ao processamento distribuído e ao tratamento de dados não estruturados ou semiestruturados (JANSSEN, N. et al. 2024).

Nesse contexto, este trabalho de caráter exploratório tem como objetivo investigar a viabilidade técnica do paradigma emergente Data Lakehouse em um ambiente local voltado ao processamento de dados geoespaciais. Por meio da integração das bases do PRODES, DETER e TerraClass como estudo de caso prático, busca-se avaliar as facilidades e as barreiras impostas por essa tecnologia para a extração de trajetórias de uso e cobertura da terra.

2. Revisão Bibliográfica

O panorama das arquiteturas de armazenamento e processamento de dados evoluiu em três gerações principais: *Data Warehouses*, *Data Lakes* e, mais recentemente, *Data Lakehouses*. Cada paradigma apresenta vantagens e limitações específicas que variam conforme o caso de uso e a natureza da aplicação (JANSSEN, N. et al. 2024).

2.1. Data Warehouses

Os *Data Warehouses* (DW) consolidam dados estruturados e pré-processados em esquemas modelados antes ou durante a ingestão (*schema-on-write*), garantindo as propriedades transacionais de Atomicidade, Consistência, Isolamento e Durabilidade (ACID), mas enfrentando limitações em escalabilidade distribuída e flexibilidade de formatos (NAMBIAR, A. et al. 2022).

No domínio geoespacial, a materialização mais comum dos Data Warehouses ocorre por meio dos SGBDs com extensões espaciais, que fornecem suporte nativo a tipos geométricos e geográficos. Uma das grandes vantagens dessa abordagem reside na robustez do mecanismo de indexação espacial e no cumprimento das garantias ACID, assegurando que operações de edição vetorial e consultas topológicas ocorram sem corrupção de dados (CASANOVA, M. et al 2005).

Contudo, quando submetidos à escala do Big Data, esses sistemas tradicionais enfrentam alguns gargalos. Por operarem majoritariamente em arquiteturas de nó único, a escalabilidade horizontal é limitada (HERDEN, O. 2020)

Ainda que *Cloud Data Warehouses* consigam lidar com questões de escalabilidade, algumas limitações permanecem, como o alto custo financeiro de ter armazenamento e processamento acoplados, a falta de suporte a *streaming* de dados e a inflexibilidade de armazenar e analisar dados semi-estruturados e não-estruturados (AIT ERRAMI, S. et al. 2023).

2.2. Data Lakes

Os *Data Lakes* (DL) por sua vez, possibilitam armazenamento de baixo custo, e maior flexibilidade devido ao suporte a dados semi e não-estruturados, e esquemas lidos sob demanda (AZZABI, S. et al. 2024).

Contudo, Schneider et al. (2024) identificam cinco desafios estruturais que arquiteturas *data lake* não conseguem endereçar adequadamente: a complexidade arquitetural decorrente da integração de múltiplos sistemas (C1); a dificuldade de unificar processamento batch e streaming (C2); as limitações no suporte a atualizações incrementais (C3); a ausência de garantias de consistência transacional (C4); e a falta de atomicidade em cenários concorrentes (C5).

Para superar as limitações dos *Data Warehouses* e dos *Data Lakes*, algumas organizações adotaram a arquitetura de dois níveis (*two-tier*). Nesse modelo híbrido, o *Data Lake* funciona como um repositório de baixo custo para o armazenamento flexível de dados brutos e heterogêneos, enquanto pipelines de ETL subsequentes limpam, transformam e movem uma fração estruturada dessas informações para um *Data Warehouse* proprietário, onde são disponibilizadas para consultas analíticas de alta performance (AIT ERRAMI, S, et al. 2023).

Contudo, essa abordagem introduz algumas fragilidades na manutenção dos sistemas. Ela exige o gerenciamento de ecossistemas tecnológicos heterogêneos e fragmentados, o que eleva a probabilidade de falhas e o custo operacional de manter múltiplos pipelines de sincronização entre as duas plataformas. Além disso, a governança e a conformidade regulatória tornam-se complexas e redundantes, pois as regras de controle de acesso e auditoria precisam ser replicadas de forma síncrona em dois sistemas de naturezas distintas. A partir dessa problemática emerge o paradigma

do *Data Lakehouse*, que se propõe a combinar as garantias transacionais do *Data Warehouse* com a escalabilidade do *Data Lake* em um mesmo ecossistema (ARMBRUST, M. et al. 2021).

2.3. Data Lakehouses

O paradigma do Data Lakehouse, proposto por Armbrust et al. (2021), é apresentado como uma plataforma que sintetiza o baixo custo de armazenamento, a flexibilidade e o suporte a dados diversos do Data Lake com as capacidades de governança e confiabilidade dos Data Warehouses.

Conforme destacado por Ait Errami et al. (2023), essa convergência é viabilizada pelo desacoplamento entre os recursos de computação e armazenamento, assentando-se sobre uma camada transacional de metadados que opera diretamente em formatos de tabela abertos.

Schneider et al. (2024) apresenta a formalização dessa arquitetura por meio de oito requisitos técnicos independentes de fornecedor, os quais funcionam como balizadores para a avaliação e validação de protótipos, conforme descritos abaixo:

- **R1 - Armazenamento e formato unificados:** Consolida dados em formatos abertos e escaláveis, mitigando a fragmentação e facilitando o acesso.
- **R2 - Operações CRUD:** Permite manipulação granular de registros (inserção, leitura, atualização e exclusão) para processamento incremental e governança.
- **R3 - Representação relacional:** Abstrai os arquivos físicos em tabelas lógicas (linhas e colunas), unificando o consumo para ferramentas analíticas.
- **R4 - Suporte a SQL:** Centraliza a exploração de dados massivos sob uma linguagem declarativa padronizada.
- **R5 - Garantias de consistência:** Impõe restrições estruturais (*schema enforcement*) na origem, rejeitando anomalias e garantindo a confiabilidade.
- **R6 - Atomicidade e isolamento:** Fornece garantias transacionais (ACID) que evitam a leitura de dados parciais ou corrompidos sob alta concorrência.
- **R7 - Acesso direto ao dado:** Permite que motores de Ciência de Dados leiam massas brutas diretamente no armazenamento, evitando gargalos de desempenho.
- **R8 - Integração Batch/Streaming:** Habilita tabelas para servirem nativamente a processos históricos em lote e ingestões em tempo real.

2.4. Formatos de Tabela Abertos (*Open Table Formats*)

A concretização prática do paradigma *Data Lakehouse* e o atendimento de seus requisitos operacionais dependem diretamente da adoção de formatos de tabela abertos (*Open Table Formats*). Essas tecnologias funcionam como uma camada de abstração de software que gerencia coleções de arquivos de dados colunares e computa metadados relacionais. Em vez de gerenciar diretórios físicos rígidos, os formatos de tabela gerenciam o estado lógico das tabelas, abstraindo a complexidade do armazenamento distribuído. Assim, esses formatos de tabela abertos (e.g. Delta

Lake, Apache Iceberg)possibilitam implementar garantias transacionais (ACID) sob sistemas de armazenamento de objetos (SIDDHARTHA, P. 2025).

2.4.1. Delta Lake

Armbrust et al. (2020) apresentaram o Delta Lake como um sistema que gerencia o estado das tabelas convertendo o armazenamento de objetos em um ambiente transacional robusto por meio da articulação de três componentes essenciais:

I. Camada de Armazenamento Delta (Delta Storage Layer)

II. Tabela Delta (Delta Table)

III. Motor Delta (Delta Engine)

A camada de armazenamento persiste os dados fisicamente em sistemas de arquivos distribuídos de baixo custo, utilizando engines distribuídas de alta performance (predominantemente o Apache Spark) para coordenar operações de leitura e escrita massiva.

Já a tabela delta organiza os dados lógicos em partições físicas compostas por arquivos Apache Parquet. O núcleo invariante desse componente é o Log de Transações Delta (Delta Log), um registro cronológico imutável de todos os comandos (commits) executados na tabela.

O DeltaLog funciona como a única fonte da verdade analítica: qualquer motor computacional que consulte a tabela lê primeiramente o log para identificar quais arquivos Parquet são válidos para aquele instante temporal, garantindo isolamento transacional completo e suporte nativo a auditorias históricas (Time Travel).

O motor delta fornece algoritmos avançados de otimização de layout de dados para acelerar o processamento paralelo. Destaca-se a técnica de agrupamento multidimensional Z-Order, uma curva espacialmente preenchedora que reorganiza as linhas dos arquivos agrupando variáveis correlacionadas na mesma proximidade física.

No processamento de grandes volumes, esse arranjo maximiza a eficiência de algoritmos de salto de arquivos (*data skipping*), permitindo ignorar partições inteiras irrelevantes durante consultas de seleção SQL.

3. Objetivos

3.1. Objetivo Geral

Desenvolver e avaliar um protótipo local de arquitetura Lakehouse para integração e análise exploratória de dados geoespaciais, validando a viabilidade do paradigma frente aos dados dos biomas Amazônia e Cerrado no recorte temporal 2018-2022.

3.2. Objetivos Específicos

I. Construir uma arquitetura lakehouse baseada na integração entre motores de processamento geoespacial, formatos de tabela aberto e sistemas de armazenamento de objetos.

II. Avaliar o protótipo à luz dos requisitos técnicos independentes de fornecedor estabelecidos por Schneider et al (2024).

III. Extrair e analisar trajetórias de mudança de uso da terra em áreas de desmatamento consolidado.

IV. Executar análises exploratórias agregadas por unidades federativas, municípios, terras indígenas e unidades de conservação.

4. Material e Métodos

4.1. Conjuntos de dados

Para viabilizar a análise das trajetórias de uso e cobertura da terra em áreas convertidas por desmatamento, este projeto baseia-se na integração de três conjuntos de dados geoespaciais estratégicos produzidos e mantidos pelo Instituto Nacional de Pesquisas Espaciais (INPE): o PRODES, o DETER e o TerraClass (ZIOTI, F. et al. 2022)

A justificativa para a escolha combinada deste conjunto de dados reside na complementaridade de suas características temporais e temáticas. Enquanto o PRODES delimita os polígonos de desmatamento consolidado sob uma escala anual, o DETER mantém alta frequência temporal, capturando a dinâmica e a velocidade do avanço da supressão da vegetação. Posteriormente, o TerraClass fornece o mapeamento temático pós-desmatamento, identificando as classes de uso da terra, como pastagem, agricultura ou vegetação secundária (ALMEIDA, C. et al. 2025)

Adicionalmente, os dados de desmatamento foram cruzados com dados complementares, como limites políticos-administrativos e áreas de interesse socioambiental. Foram incorporadas as malhas digitais de Estados e Municípios para regionalização das trajetórias, além das camadas de Terras Indígenas, Unidades de Conservação e Biomas, permitindo correlacionar as trajetórias de degradação com categorias de proteção territorial.

A Tabela 1 sintetiza as características estruturais e as estimativas de volume de todos os conjuntos de dados. Apenas os dados do TerraClass não foram baixados em sua totalidade, restringindo apenas para os anos de 2018, 2020 e 2022.

Tabela 1: Características dos conjuntos de dados

Conjunto de Dados	Formato Original	Escala Temporal	Volume
PRODES	Vetorial (WFS)	Anual	~ 1,5 GB
DETER	Vetorial (WFS)	Diária	~ 228 MB
TerraClass	Raster (GeoTIFF)	Bienal	~ 6,0 GB
Limites IBGE	Vetorial (Shapefile)	Estático	~ 322 MB
Terras Indígenas	Vetorial (WFS)	Dinâmico	~ 28 MB
Unidades de Conservação	Vetorial (WFS)	Dinâmico	~ 84 MB

Biomas	Vetorial (Shapefile)	Estático	~ 12 MB

4.2. Construção do ambiente e infraestrutura computacional

A arquitetura computacional do protótipo foi implementada em ambiente local (*standalone*) recorrendo à tecnologia de conteinerização. Utilizou-se o Docker como ferramenta de orquestração de microsserviços sob uma plataforma Linux virtualizada de arquitetura linux/amd64.

A adoção desse paradigma justifica-se pela necessidade de garantir a reprodutibilidade experimental da pesquisa, isolando as dependências de software em unidades leves (contêineres) e eliminando incompatibilidades entre os ambientes de execução. O ambiente foi separado em três módulos lógicos principais através do arquivo de configuração do Docker Compose:

I. **Módulo de Armazenamento:** Gerenciado por um contêiner MinIO, como servidor de armazenamento de objetos de código aberto que emula o comportamento e as APIs do Amazon S3, atuando como o repositório centralizado (data lake) do sistema.

II. **Módulo de Ingestão**: Baseado em uma imagem contendo as bibliotecas GDAL/OGR e Python, responsável pelo consumo programático das APIs WFS do INPE e pela conversão inicial dos formatos brutos.

III. **Módulo de Transformação:** Centralizado em um contêiner customizado contendo o motor de computação distribuída Apache Spark, estendido nativamente com o framework Apache Sedona 1.9 (para o processamento escalável de geometrias e matrizes) e a biblioteca Delta Lake 4.2 (como formato de tabela aberto transacional).

A infraestrutura de hardware utilizada no desenvolvimento e execução desse ecossistema baseia-se em uma máquina hospedeira equipada com processador Intel(R) Core(TM) i7-1165G7 @ 2.80GHz (4 núcleos e 8 threads), 16,0 GB de memória RAM (com 15,7 GB utilizáveis), placa gráfica dedicada NVIDIA GeForce MX350 (2 GB VRAM) acoplada à interface integrada Intel(R) Iris(R) Xe Graphics, e armazenamento físico em SSD NVMe de 477 GB.

4.3. Organização dos Dados: Arquitetura Medalhão

Para estruturar o fluxo de processamento dos dados dentro do MinIO, adotou-se o padrão de *design* conhecido como Arquitetura Medalhão (MOHNA, H, et al. 2022). Essa abordagem consiste na organização lógica e física dos dados em três camadas sequenciais de maturidade, cada uma isolada em um *bucket* exclusivo dentro do armazenamento de objetos:

- **Camada Bronze:** Armazena os dados exatamente como foram extraídos das fontes originais, sem qualquer modificação estrutural. Os arquivos são gravados no bucket bronze preservando seu estado bruto, servindo como o histórico imutável do sistema. A existência desta camada garante a linhagem de dados e permite reprocessar todo o ecossistema do zero.

- **Camada Prata:** Recebe os dados da camada Bronze após passarem por rotinas de limpeza, padronização e enriquecimento geoespacial. Nesta etapa, os dados são convertidos para o formato Delta Lake, as geometrias inválidas são corrigidas, os Sistemas de Referência de

Coordenadas (CRS) são padronizados e técnicas de indexação espacial (como o Z-Order) são aplicadas.

- **Camada Ouro:** É a camada final onde os dados da Prata são agregados, sumarizados e estruturados para responder a perguntas específicas do projeto. No contexto desta pesquisa, a camada Ouro armazena o produto final da integração: as tabelas de Trajetórias de Uso da Terra, cruzando espacialmente os polígonos de desmatamento históricos com suas respectivas evoluções temáticas ao longo dos anos.

A justificativa para o uso da arquitetura medalhão reside na robustez operacional que ela confere. Em vez de transformar o dado bruto diretamente em um produto final analítico, a segregação por níveis de tratamento mitiga riscos de corrupção, simplifica a manutenção dos esquemas, isola falhas de ingestão e garante que múltiplos usuários acessem bases confiáveis, performáticas e auditáveis temporalmente (MOHNA, H, et al. 2022).

4.4. Rotinas de processamentos

Visando automatizar a coleta, transformação e carga dos dados e garantir a reprodutibilidade do fluxo de processamento, foram desenvolvidas rotinas para preencher cada camada da arquitetura de acordo com seu propósito.

4.4.1. Camada Bronze

Na camada bronze, as rotinas de processamento (*scripts*) foram escritas em *shell script* operando de forma encapsulada no módulo de ingestão.

No caso das variáveis vetoriais, os dados brutos são consumidos diretamente das APIs oficiais de serviços de mapeamento web (OGC Web Feature Service - WFS) mantidas pelas instituições de origem. O componente de automação dispara rotinas baseadas na ferramenta utilitária *ogr2ogr* da biblioteca GDAL, mapeando os *endpoints* remotos, extraindo os atributos e as geometrias e persistindo-os imediatamente no *bucket* bronze do MinIO.

Optou-se por salvar esses dados diretamente no formato GeoParquet. O GeoParquet estende o formato de armazenamento colunar Apache Parquet para o domínio geoespacial, codificando as geometrias no formato binário padronizado Well-Known Binary (WKB) e injetando metadados estruturados sobre o Sistema de Referência de Coordenadas (CRS) no cabeçalho do arquivo.

Por outro lado, o conjunto de dados raster do TerraClass exigiu uma abordagem diferente. Os arquivos originais foram obtidos manualmente em formato GeoTIFF padrão. Para viabilizar seu consumo eficiente dentro da arquitetura baseada em arquivos, utilizou-se a biblioteca GDAL (*gdal_translate*) para convertê-los no formato COG (Cloud Optimized GeoTIFF) antes da carga para a camada bronze.

O COG baseia-se em uma organização interna em blocos (*tiling*) combinada com imagens reduzidas pré-calculadas (*overviews* ou pirâmides de resolução). A lógica por trás da adoção do COG reside na capacidade de permitir leituras parciais do arquivo por meio de requisições baseadas em intervalos de bytes (HTTP Range Requests no ecossistema S3).

4.4.2. Camada Prata

A transição dos dados da camada Bronze para a camada Prata (Silver) representa o núcleo de refinamento, padronização e otimização geoespacial do Lakehouse. Nesta etapa, as rotinas de processamento consomem os arquivos da camada bronze para aplicar regras de refinamento dos dados, tornando-os padronizados.

O fluxo de processamento e enriquecimento da camada Prata consistiu no tratamento e padronização de dados geoespaciais considerando: (a) garantia de qualidade topológica; (b) padronização cartográfica; (c) indexação espacial e (d) modelagem relacional.

Para economizar etapas de processamento, optou-se pelo formato de tabela aberto Delta Lake que suporta nativamente os tipos geoespaciais (*geometry*), ao contrário do Apache Iceberg que mesmo na sua versão mais recente, ainda necessita de conversões para WKB ou WKT. Assim, as feições espaciais são lidas, indexadas e mantidas na memória do Spark em seu formato nativo, reduzindo a sobrecarga de decodificação e acelerando os predicados de intersecção espacial.

Considerando aspectos de qualidade topológica, os polígonos de desmatamento originais apresentaram inconsistências como auto-intersecções e anéis abertos. Para assegurar a integridade geométrica e evitar erros fatais em cálculos geográficos subsequentes, a rotina executa sistematicamente a função espacial ST_MakeValid. Esse operador reconstrói a topologia dos polígonos inválidos, gerando feições geometricamente coerentes sem alterar a delimitação espacial do desmatamento original.

A rotina também realiza a reprojeção de todos os datasets vetoriais para o código EPSG:10857 (SIRGAS 2000 / Albers Equal Area para a América do Sul). Sendo uma projeção cônica equivalente, ela preserva a fidelidade das áreas mapeadas, permitindo o cálculo padronizado do atributo de área em hectares (area_ha) de cada polígono, eliminando distorções métricas comuns de projeções geográficas ou cilíndricas.

Para acelerar as consultas analíticas e os cruzamentos espaciais massivos, aplicou-se a otimização de layout baseada na curva de preenchimento espacial Z-Order. A rotina computa o código GeoHash a partir do centróide de cada geometria vetorial. Posteriormente, o comando OPTIMIZE do Delta Lake utiliza essa *string* geocodificada como chave para o algoritmo de agrupamento multidimensional, reorganizando fisicamente as linhas dentro dos arquivos Parquet de modo que polígonos que estão geograficamente próximos no mundo real sejam gravados na mesma proximidade física no MinIO. Esse arranjo eleva a eficiência dos algoritmos de salto de arquivos.

Visando unificar e centralizar as informações históricas, os dados originais do PRODES (historicamente segregados por biomas e subcamadas como anual, acumulado e residual) e do DETER (divididos originalmente entre os biomas Amazônia e Cerrado) foram consolidados em tabelas únicas para cada respectivo sistema de monitoramento.

Para assegurar a integridade referencial e a unicidade de cada registro nas bases consolidadas, gerou-se uma chave substituta (*surrogate key* — id_silver) para cada feição. Essa chave foi calculada por meio da aplicação do algoritmo de criptografía analítica MD5 (hashmd5) sobre a

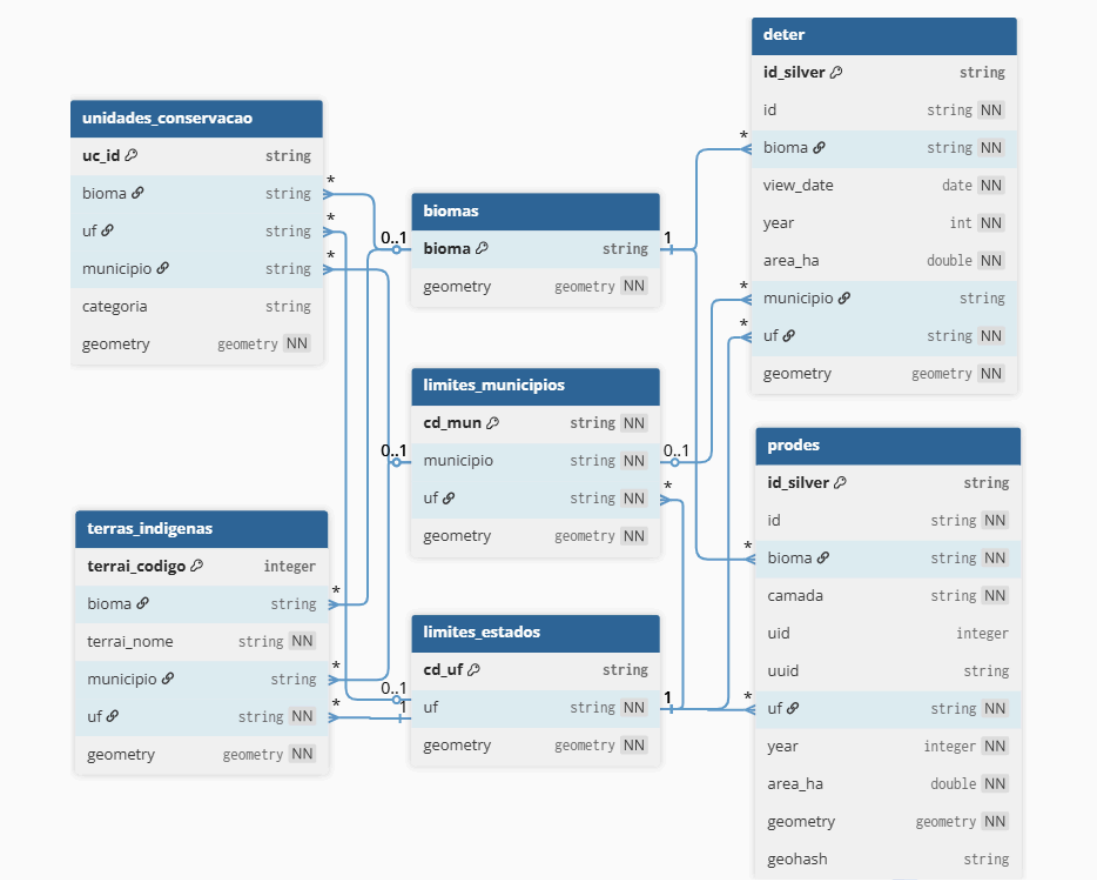
concatenação de atributos estruturais de origem, blindando a tabela contra registros duplicados oriundos de diferentes cargas de dados.

A capacidade de fornecer uma representação relacional rigorosa (R3) e um suporte robusto a consultas estruturadas via linguagem SQL (R4) constituem pilares fundamentais das vantagens dos formatos de tabela abertos. Alinhado a essas diretrizes, o esquema físico e conceitual da camada Prata foi concebido previamente à persistência dos dados. Foram declaradas restrições de não-nulidade para campos críticos, além da criação e definição de chaves primárias e chaves estrangeiras lógicas, baseadas nos atributos de município, unidade federativa e bioma. Por fim, as tabelas do PRODES e DETER foram particionadas fisicamente pelo atributo ano.

A partir dessa padronização, foi desenvolvida uma modelagem relacional estruturada para fins analíticos, conforme apresentado no diagrama da Figura 1. Esse arranjo lógico permite que as tabelas de monitoramento ambiental conectem-se de forma direta e padronizada às dimensões territoriais e político-administrativas através de chaves estrangeiras.

Assim, essa estrutura viabiliza a execução de filtros alfanuméricos e junções na memória do cluster, otimizando o processamento de dados textuais e numéricos antes que o motor computacional precise executar os predicados de intersecção espacial, que são computacionalmente mais onerosos.

Figura 1: Modelagem relacional da camada Prata



Buscando atender às exigências mais rígidas de governança e auditoria ambiental, ativou-se o recurso *Change Data Feed* (CDF) nas tabelas Delta Lake da camada Prata. O CDF registra e expõe os metadados de todas as mutações sofridas pelos dados ao longo do tempo (identificando precisamente quais linhas sofreram inserções, atualizações ou deleções em cada ciclo de processamento). Essa capacidade confere total rastreabilidade retroativa (*Time Travel*), permitindo auditar o histórico de modificações do monitoramento sem gerar duplicatas.

Sob a perspectiva do processamento contínuo, o fluxo de sincronização entre as camadas Bronze e Prata apoia-se nativamente no comando MERGE do Delta Lake para operacionalizar a lógica de ingestão incremental. Em arquiteturas como os data lakes, a chegada de novos registros de monitoramento exigiria a leitura completa da tabela de destino, a união dos dados em memória e a reescrita total dos arquivos em disco.

4.4.3. Camada Ouro

A camada Ouro constitui a etapa final de maturação e especialização dos dados dentro da arquitetura Lakehouse. Para evitar cálculos em tempo real, a lógica nessa camada é cruzar as tabelas de desmatamento contra as tabelas de limites municipais e estaduais, Terras Indígenas e Unidades de Conservação, e adicionar os atributos resultantes do cruzamento em uma nova tabela.

Para o desenho experimental, foram isolados os 5 maiores polígonos de desmatamento mapeados pelo PRODES para cada um dos biomas avaliados (Amazônia e Cerrado). Esses polígonos foram submetidos a um processo de interseção espacial e fatiamento temporal contra a série histórica do TerraClass.

O motor computacional Apache Sedona calculou e discriminou, linha a linha, a área em hectares (area_classe_ha) ocupada por cada classe temática do TerraClass (como pastagem, agricultura ou vegetação secundária) dentro do perímetro daquele desmatamento específico, para cada ano da série histórica.

Para atualizações futuras, todo esse fluxo pode ser viabilizado por meio de processamento incremental. Com o CDF ativado, o ecossistema consome apenas as mutações e novos logs de transação capturados pelo Delta Lake, garantindo que o cálculo de trajetórias e o enriquecimento de novos polígonos ocorram de forma eficiente e otimizada.

5. Resultados

Como o escopo central deste estudo reside na validação infraestrutural da arquitetura *Lakehouse*, o cruzamento entre os dados de perda de vegetação (PRODES e DETER) e as malhas de limites administrativos e áreas protegidas operou, primordialmente, como um teste de estresse e prova de conceito das operações de junção espacial do ecossistema.

As informações extraídas, cumprem o papel de ilustrar a capacidade da plataforma em agregar e sumarizar dados, não devendo ser interpretados como um censo estatístico definitivo da dinâmica ambiental.

Os resultados foram gerados por meio de operações de relacionamento espacial do Apache Sedona — ST_Intersects, ST_Intersection e ST_Touches — combinadas a cálculos métricos de área (ST_Area), agregações por divisões territoriais (GROUP BY), filtros temporais restritos ao período de 2018 a 2022 (WHERE) e ranqueamentos (ORDER BY).

Além do cruzamento entre os dados vetoriais, também foram calculadas as estatísticas relativas à trajetória das classes de uso e ocupação da Terra em áreas de desmatamento através da sobreposição de polígonos do PRODES com os dados matriciais do TerraClass.

Abaixo, a Tabela 2 sintetiza os principais resultados obtidos através do cruzamentos entre as os dados vetoriais do protótipo, onde são apresentados alguns índices de desmatamento agrupados por sistema de detecção e bioma.

Tabela 2: Desmatamento por recorte territorial

Fonte	Bioma	Área Total (km²)	UF mais afetada	Município mais afetado	Sobreposição com UCs (km²)	Sobreposição com TIs (km²)
PRODES	Amazônia	53.600,08	PA (40,33%)	Altamira - PA (6,01%)	7.057,99 (13,17%)	2.007,08 (3,74%)
PRODES	Cerrado	42.918,92	MA (24,59%)	Balsas - MA (2,56%)	3.458,93 (8,06%)	490,15 (1,14%)
DETER	Amazônia	112.995,98	MT (44,28%)	São Félix do Xingu - PA (5,13%)	10.878,67 (9,63%)	15.142,24 (13,40%)
DETER	Cerrado	25.748,91	TO (21,72%)	Formosa do Rio Preto - BA (3,24%)	2.335,73 (9,07%)	263,97 (1,03%)

5.1. Desmatamento por limites administrativos

As consultas analíticas em SQL espacial demonstraram a eficácia do motor de processamento no agrupamento de dados em múltiplas escalas geográficas. Ao processar o período de 2018 a 2022, o sistema evidenciou capacidades de agregação consistentes.

Na Amazônia, foi identificado uma área total de 112.995,98 km² em alertas (DETER) face a 53.600,08 km² de área consolidada de desmatamento (PRODES). No Cerrado, a proporção calculada inverteu-se, com o PRODES a somar 42.918,92 km² contra os 25.748,91 km² do DETER.

A nível estadual e municipal, as agregações validaram a extração de métricas em zonas de pressão histórica. O motor SQL identificou que, na Amazônia, o processamento espacial concentrou o maior volume de dados nos estados do Pará (PA) e de Mato Grosso (MT), destacando-se agrupamentos expressivos em municípios como Altamira e São Félix do Xingu.

No Cerrado, destacam-se a proeminência do Maranhão (MA) e do Tocantins (TO), com convergência espacial para municípios como Balsas (MA) e Formosa do Rio Preto (BA). A execução destas agregações confirmou a viabilidade da transição dos dados da camada Prata para a geração de sumários executivos na camada Ouro.

5.2. Desmatamento em áreas protegidas

5.2.1. Terras Indígenas

Ao processar os dados de desmatamento consolidado (PRODES) de 2018 a 2022, o motor analítico quantificou uma supressão de 2.007,08 km² no interior de TIs na Amazônia (representando 3,74% do desmatamento global do bioma) e de 490,15 km² no Cerrado (1,14%). As consultas SQL de ordenamento demonstraram fluidez ao hierarquizar os territórios mais afetados, isolando a TI Apyterewa como o alvo de maior incidência na Amazônia, com 340,26 km² (16,95% do total na categoria). No Cerrado, a agregação apontou a TI Areões na liderança isolada, respondendo por mais de um terço da área suprimida em territórios indígenas no bioma (174,69 km² ou 35,64%).

Ao submeter os alertas rápidos (DETER) às mesmas rotinas de álgebra espacial, o sistema evidenciou sua capacidade de escalar as agregações frente a volumes e padrões de dados substancialmente distintos. Na Amazônia, o somatório de alertas em TIs processado pelo Lakehouse atingiu 15.142,24 km² (13,40% dos alertas do bioma), com o algoritmo detectando uma concentração massiva de 5.640,81 km² apenas na TI Parque do Xingu. Em contrapartida, no Cerrado, as consultas retornaram um volume acumulado de 263,97 km² (1,03%), agrupando a maior área na TI Paresi (52,02 km²).

5.2.2. Unidades de Conservação

O cruzamento espacial e a mensuração das áreas mapeadas pelo PRODES e DETER no interior de Unidades de Conservação (UCs), entre os anos de 2018 e 2022, indicam os totais absolutos e relativos de supressão vegetal nesses territórios protegidos.

No bioma Amazônia, o sistema PRODES registrou um acumulado de 7.057,99 km² de desmatamento em Unidades de Conservação, o que equivale a 13,17% de toda a área desmatada no bioma no período. No bioma Cerrado, o desmatamento acumulado em Unidades de Conservação registrou um total de 3.458,93 km², representando 8,06% do desmatamento global mapeado para o bioma no intervalo pelo sistema PRODES.

No bioma Amazônia, o sistema DETER registrou um acumulado de 10.878,67 km² de alertas no interior de Unidades de Conservação, o que representa 9,63% de toda a área de alertas contabilizada para o bioma no período. No bioma Cerrado, os alertas acumulados em Unidades de Conservação perfizeram um total de 2.335,73 km², representando 9,07% do total geral de alertas mapeados para o bioma no intervalo avaliado.

5.3. Sobreposições entre PRODES e DETER

Os resultados discriminam a área total do polígono PRODES, a quantidade absoluta de alertas do DETER interceptados e a extensão em hectares da sobreposição interna entre as duas bases de dados para o ano de 2018, conforme apresentado na Tabela 5.

Tabela 3: Correspondência espacial entre os maiores polígonos desmatados (PRODES) e os alertas emitidos (DETER) em 2018

ID PRODES	Bioma	Área PRODES (ha)	Qtd. Alertas DETER	Área DETER Interna (ha)
amazonia_1	Amazônia	5.376,76	49	5.202,35
amazonia_2	Amazônia	4.613,83	62	4.055,53
amazonia_3	Amazônia	1.873,35	17	1.611,45
amazonia_4	Amazônia	1.835,65	1	63,84
amazonia_5	Amazônia	1.826,02	28	1.228,03
cerrado_1	Cerrado	6.868,10	6	6.832,05
cerrado_2	Cerrado	3.680,99	14	2.458,46
cerrado_3	Cerrado	3.609,35	24	3.541,81
cerrado_4	Cerrado	2.817,92	1	167,48
cerrado_5	Cerrado	2.744,78	25	2.558,14

5.4. Trajetória de uso e ocupação da terra

Devido às limitações de hardware do protótipo, não foi possível fazer o cálculo de trajetória de uso e ocupação da Terra para todos os polígonos do PRODES. Para dar continuidade à análise exploratória, foram selecionados os maiores polígonos PRODES de cada bioma para o cruzamento com os dados matriciais do TerraClass, conforme apresentado na Tabela 4:

Tabela 4 : Trajetória de Uso e Ocupação do Solo em áreas de desmatamento no Cerrado

ID do Polígono	Área Total (ha)	Cenário em 2018	Cenário em 2020	Cenário em 2022
cerrado_1	6.868,10	Desmatamento (99,6%)	Pastagem (96,0%)	Pastagem (83,3%) Agric. 1 Ciclo (13,0%)
cerrado_2	3.680,99	Desmatamento (99,8%)	Pastagem (99,8%)	Pastagem (86,5%) Agric. Vários Ciclos (8,3%)
cerrado_3	3.609,35	Desmatamento (99,7%)	Pastagem (51,8%) Agric. 1 Ciclo (35,4%)	Agricultura 1 Ciclo (99,9%)
cerrado_4	2.817,92	Desmatamento (99,7%)	Pastagem (99,7%)	Pastagem (99,7%)
cerrado_5	2.744,78	Desmatamento (99,3%)	Pastagem (51,5%) Veg. Secundária (48,3%)	Agricultura 1 Ciclo (96,1%)

No bioma Cerrado, observou-se uma rápida conversão das áreas desmatadas para atividades agropecuárias, com predomínio inicial da pecuária e posterior expansão da agricultura mecanizada em parte dos polígonos analisados. Em todos os casos, o desmatamento registrado

em 2018 atingiu praticamente a totalidade das áreas estudadas, evidenciando a substituição quase completa da cobertura vegetal original.

As trajetórias identificadas podem ser agrupadas em três padrões principais. O primeiro corresponde à consolidação da atividade pecuária, observada nos polígonos cerrado_1, cerrado_2 e cerrado_4, onde a pastagem passou a ocupar a maior parte da área após o desmatamento. O segundo padrão caracteriza-se pela intensificação agrícola, observada nos polígonos cerrado_3 e cerrado_5, nos quais áreas inicialmente ocupadas por pastagens ou vegetação secundária foram posteriormente convertidas para agricultura mecanizada, culminando na predominância de lavouras temporárias em 2022. O terceiro padrão refere-se à diversificação produtiva, observada em menor intensidade nos polígonos cerrado_1 e cerrado_2, onde a atividade pecuária passou a coexistir com diferentes modalidades agrícolas.

De forma geral, os resultados indicam que o desmatamento no Cerrado está fortemente associado à expansão das atividades agropecuárias, corroborando a literatura que aponta a conversão de áreas naturais para sistemas de produção agrícola e pecuária como um dos principais vetores de transformação da paisagem no bioma.

Tabela 5: Trajetória de Uso e Ocupação do Solo em áreas de desmatamento na Amazônia

ID do Polígono	Área Total (ha)	Cenário em 2018	Cenário em 2020	Cenário em 2022
amazonia_1.	5.376,76	Desmatamento (99,6%)	Vegetação Florestal Sec. (64,3,%) Pastagem Arbustiva (35,0%)	Pastagem Arbustiva (99,7%)
amazonia_2	4.613,83	Desmatamento (99,8%)	Vegetação Florestal Sec. (61,5%) Pastagem Arbustiva (21,8%) Pastagem Herbácea (16,6%)	Vegetação Florestal Sec. (37,7,%) Pastagem Arbustiva (31,7%) Pastagem Herbácea (30,5%)
amazonia_3	1.873,35	Desmatamento (99,7%)	Vegetação Florestal Sec. (85,1%) Pastagem Arbustiva (14,4%)	Vegetação Florestal Sec. (79,2%) Pastagem Arbustiva (12%) Pastagem Herbácea (8,4%)
amazonia_4	1.835,65	Desmatamento (99,7%)	Vegetação Florestal Sec. (3,3%) Pastagem Arbustiva (96,3%)	Vegetação Florestal Sec. (7,6%) Pastagem Arbustiva (30,5%) Pastagem Herbácea (61,8%)
amazonia_5	1.826,02	Desmatamento (99,3%)	Vegetação Florestal Sec. (5,5%) Pastagem Arbustiva (5,1%) Pastagem Herbácea (88,9%)	Vegetação Florestal Sec. (6,2%) Pastagem Arbustiva (7,3%) Pastagem Herbácea (86,2%)

No bioma Amazônia, as trajetórias de uso e ocupação do solo apresentaram maior heterogeneidade quando comparadas às observadas no Cerrado, evidenciando a coexistência de processos de regeneração secundária e expansão das atividades pecuárias.

Os resultados permitiram identificar três padrões principais de transição. O primeiro corresponde à consolidação da pecuária após um período inicial de regeneração florestal secundária, observado nos polígonos amazonia_1 e amazonia_2. Nesses casos, áreas inicialmente ocupadas por vegetação secundária foram gradualmente substituídas por diferentes

modalidades de pastagem, indicando a incorporação progressiva dessas áreas aos sistemas produtivos.

O segundo padrão caracteriza-se pela persistência da regeneração natural, representada pelo polígono amazonia_3. Nesse caso, a vegetação florestal secundária permaneceu predominante ao longo do período analisado, sugerindo processos de pousio ou recuperação da cobertura vegetal com baixa intensidade de uso antrópico.

Por fim, os polígonos amazonia_4 e amazonia_5 ilustram uma conversão mais direta e estável para atividades pecuárias, com predominância de pastagens manejadas desde os primeiros anos após o desmatamento. Embora tenham sido observadas alterações na composição dos tipos de pastagem, o uso predominante permaneceu associado à atividade pecuária durante todo o período analisado.

De maneira geral, os resultados indicam que, diferentemente do Cerrado — onde predominou a rápida expansão agrícola —, a dinâmica observada na Amazônia foi marcada pela maior relevância da pecuária e pela ocorrência de trajetórias intermediárias de regeneração florestal secundária.

6. Discussão

A presente seção de Discussão propõe uma reflexão aprofundada sobre os resultados obtidos, estruturando-se a partir de três questões norteadoras fundamentais. Sob a perspectiva da Análise Arquitetural, investiga-se (Q1) quais são os pontos fortes e fracos de uma arquitetura Lakehouse local (single-node) ao processar e unificar dados geoespaciais heterogêneos, em contraponto aos paradigmas tradicionais.

No que tange à Dinâmica do Fenômeno, busca-se compreender (Q2) como os polígonos de desmatamento alteraram seu padrão de uso e cobertura da terra ao longo do tempo (2018, 2020 e 2022) nos biomas Amazônia e Cerrado.

6.1. Avaliação da arquitetura Lakehouse

Para responder à primeira questão norteadora (Q1), a implementação do protótipo permitiu avaliar, em um cenário real de integração entre dados vetoriais e matriciais, até que ponto a arquitetura Lakehouse consegue atender aos requisitos propostos por Schneider et al. (2024) para ambientes analíticos geoespaciais.

Em comparação com arquiteturas tradicionais de dois níveis (two-tier), baseadas na separação entre Data Lake e Data Warehouse, o principal benefício observado foi a consolidação do armazenamento e do processamento em uma única infraestrutura. Em vez de replicar dados entre múltiplos sistemas especializados, os diferentes conjuntos de dados puderam ser mantidos em um repositório comum, acessado diretamente pelo mesmo mecanismo de processamento distribuído. Essa abordagem reduziu a complexidade dos fluxos de ingestão e eliminou a necessidade de processos intermediários de sincronização, um dos principais problemas apontados na literatura para ambientes geoespaciais de grande escala.

Outro aspecto relevante foi a possibilidade de acesso direto aos dados armazenados no sistema de arquivos, sem dependência de camadas intermediárias de publicação de serviços geográficos. Tradicionalmente, a disseminação de dados espaciais exige servidores dedicados, como GeoServer ou MapServer, responsáveis por disponibilizar informações via protocolos OGC. No protótipo desenvolvido, entretanto, os dados puderam ser acessados diretamente pelo mecanismo analítico através de consultas SQL enriquecidas com operadores espaciais do Apache Sedona, aproximando o domínio geoespacial dos paradigmas modernos de engenharia de dados.

Sob a perspectiva de gerenciamento dos dados, o Delta Lake desempenhou papel fundamental para atender requisitos relacionados à consistência, processamento incremental e governança operacional. Recursos como MERGE, OPTIMIZE, VACUUM, Time Travel e Change Data Feed permitiram implementar funcionalidades normalmente associadas a sistemas transacionais, mas raramente disponíveis em Data Lakes convencionais. Essas capacidades facilitaram tanto a manutenção das tabelas quanto a reconstrução de estados históricos dos dados, característica particularmente relevante em aplicações ambientais, nas quais a rastreabilidade temporal constitui requisito fundamental para auditoria e reprodutibilidade científica.

Apesar desses avanços, a implementação também evidenciou limitações importantes. O requisito de formato unificado foi apenas parcialmente alcançado. Enquanto os dados vetoriais puderam ser armazenados e processados diretamente em tabelas Delta, os dados matriciais permaneceram em arquivos Cloud Optimized GeoTIFF (COG). A integração entre esses domínios ocorreu por meio de metadados e referências cruzadas, mas não através de um modelo verdadeiramente unificado de armazenamento. Esse resultado reforça observações recentes da literatura de que o ecossistema de Big Data geoespacial ainda apresenta níveis distintos de maturidade para dados vetoriais e raster.

Além disso, verificou-se que a simples adoção de um formato de tabela aberto não resolve integralmente os desafios de governança. Embora o Delta Lake forneça garantias transacionais locais, ambientes colaborativos e distribuídos demandam mecanismos adicionais para gerenciamento de metadados, controle de acesso e coordenação entre múltiplos usuários. Nesse contexto, torna-se necessária a incorporação de catálogos especializados, como Unity Catalog, Apache Polaris ou soluções equivalentes, capazes de atuar como camada centralizadora de governança e descoberta dos dados.

Os resultados também permitiram observar, na prática, diversos desafios apontados por Errami et al. (2023) para Lakehouses geoespaciais. Entre eles destacam-se a necessidade de estratégias adequadas de particionamento espacial, a gestão padronizada de metadados geográficos, a integração entre modelos raster e vetorial e o elevado custo computacional associado a operações geométricas complexas.

Entretanto, a principal contribuição do experimento talvez tenha sido revelar que muitos dos desafios atuais não estão associados ao conceito Lakehouse em si, mas ao grau de maturidade das ferramentas geoespaciais disponíveis no ecossistema Big Data.

O processamento vetorial apresentou desempenho satisfatório e boa integração com o modelo distribuído do Apache Spark. Em contrapartida, o processamento matricial mostrou-se significativamente mais problemático. Embora o Spark tenha sido concebido para trabalhar com estruturas tabulares particionadas, dados raster representam superfícies contínuas e exigem operações matemáticas distintas das tradicionalmente utilizadas em DataFrames. Como consequência, tarefas relativamente simples em softwares SIG convencionais tornaram-se computacionalmente custosas quando executadas no ambiente distribuído.

As limitações foram particularmente evidentes durante as etapas de reprojeção cartográfica e cruzamento entre dados vetoriais e matriciais. A ausência de suporte robusto para reprojeção distribuída de rasters e as dificuldades encontradas com determinadas projeções cartográficas comprometeram parte da flexibilidade originalmente esperada da arquitetura. Em diversos momentos foi necessário adaptar o fluxo metodológico para contornar restrições das bibliotecas disponíveis, evidenciando que a escalabilidade oferecida pelos frameworks de Big Data ainda não substitui completamente as capacidades consolidadas dos sistemas tradicionais de geoprocessamento.

Dessa forma, os resultados sugerem que o paradigma Lakehouse representa um avanço significativo para integração, governança e análise de grandes volumes de dados geoespaciais, mas ainda convive com limitações importantes relacionadas ao processamento raster.

6.2. Análise da Trajetória de Uso e Ocupação da Terra

Para responder à segunda questão norteadora (Q2), a integração entre os dados do PRODES e do TerraClass permitiu reconstruir trajetórias espaço-temporais de uso e cobertura da terra entre 2018 e 2022, evidenciando diferenças marcantes entre os processos de ocupação observados no Cerrado e na Amazônia.

No Cerrado predominou um padrão de conversão rápida e altamente direcionada para atividades produtivas. Os polígonos analisados demonstraram que o desmatamento foi frequentemente seguido por uma breve fase de ocupação pecuária, posteriormente substituída pela agricultura mecanizada. Em diversos casos, áreas originalmente convertidas para pastagem passaram a ser dominadas por lavouras temporárias em um intervalo inferior a quatro anos. Essa dinâmica confirma o papel do Cerrado como principal fronteira contemporânea de expansão agropecuária do país e reflete as condições favoráveis de relevo, infraestrutura e aptidão agrícola presentes no bioma.

Na Amazônia, por outro lado, as trajetórias apresentaram maior heterogeneidade. Diversos polígonos exibiram fases intermediárias de regeneração secundária após o desmatamento inicial, indicando tentativas temporárias de recuperação da cobertura vegetal. Contudo, em parte significativa dos casos, essa regeneração mostrou-se transitória, sendo posteriormente substituída pela expansão de diferentes modalidades de pastagem.

Esse comportamento sugere a existência de um processo gradual de degradação da paisagem, no qual o corte raso inicial é seguido por períodos de regeneração parcial, posteriormente interrompidos pela consolidação de atividades pecuárias. Tal padrão é compatível

com a literatura que descreve mecanismos de degradação progressiva e savanização da floresta amazônica, especialmente em áreas submetidas à recorrência de fogo e ao manejo extensivo do gado.

A análise também evidenciou diferenças importantes entre os instrumentos de monitoramento utilizados. Enquanto o PRODES registra predominantemente eventos consolidados de supressão florestal, o DETER foi capaz de capturar estágios intermediários de degradação e perturbação da vegetação, sobretudo na Amazônia. Entretanto, observou-se uma capacidade significativamente menor de detecção no Cerrado, possivelmente associada à própria estrutura fisionômica do bioma e às limitações de resolução espacial dos sensores empregados.

Esse resultado reforça que a interpretação das trajetórias de uso da terra depende não apenas dos processos ambientais observados, mas também das características técnicas dos sistemas de monitoramento utilizados para sua detecção.

Finalmente, é importante destacar que os resultados apresentados não possuem pretensão de representar estatisticamente a totalidade dos biomas analisados. Os polígonos selecionados constituem um estudo de caso voltado à validação da arquitetura proposta. Ainda assim, os experimentos demonstraram que a integração de múltiplas bases geoespaciais em um ambiente Lakehouse possibilita reconstruir trajetórias complexas de ocupação territorial, evidenciando o potencial da abordagem para aplicações futuras em monitoramento ambiental, planejamento territorial e apoio à formulação de políticas públicas.

7. Considerações finais

O estudo indica que uma arquitetura Lakehouse pode ser uma base coerente para integração e análise exploratória de dados geoespaciais. O protótipo demonstrou vantagens concretas em armazenamento unificado, representação relacional, suporte a SQL, rastreabilidade temporal e processamento incremental.

Ao mesmo tempo, a avaliação mostrou que a adoção do Lakehouse não elimina desafios centrais do geoprocessamento. Persistem gargalos de desempenho, dificuldades de paralelização, limitações de reprojeção e lacunas de maturidade no ecossistema geoespacial. Em especial, os dados raster continuam mais sensíveis a essas limitações do que o componente vetorial.

Como conclusão geral, os dados ambientais usados neste trabalho corroboram, em termos exploratórios, padrões já descritos na literatura para Amazônia e Cerrado. Entretanto, a contribuição principal do artigo está na avaliação da arquitetura e não em inferências territoriais amplas. O protótipo sugere viabilidade técnica do Lakehouse em análises geoespaciais, mas ainda como solução em amadurecimento, que depende de avanços de governança, otimização e integração distribuída para atingir uso mais robusto em produção.

Referências

AIT ERRAMI, Soukaina. Hajji H, Ait El Kadi K, Badir H. Spatial big data architecture: from data warehouses and data lakes to the lake-house. Journal of Parallel Distributed Computing. 2023;176:70–9. https://doi.org/10.1016/j.jpdc.2023.02.007.

ALMEIDA, Cláudio Aparecido de et al. Monitoramento oficial da vegetação nativa brasileira por imagens de satélite: o programa BiomasBR e os sistemas Prodes, Deter e TerraClass. Cadernos de Astronomia, Vitória-ES, Brasil, v. 6, n. 1, p. 23–38, 2025. DOI: 10.47456/Cad.Astro.v6n1.47411.

ARMBRUST, M., Ghodsi, A., Xin, R., & Zaharia, M. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. 2021. In Proceedings of CIDR (Vol. 8, No. 1, p. 28).

ARMBRUST, M. et al. Delta lake: high-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, v. 13, n. 12, p. 3411-3424, 2020.

AZZABI, S., Alfughi, Z., & Ouda, A. (2024). Data Lakes: A Survey of Concepts and Architectures. Computers, 13(7), 183. https://doi.org/10.3390/computers13070183

CASANOVA, M., Camara, G., Davis Jr, C., Vinhas, L., & Queiroz, G. (2005). Bancos de Dados Geográficos. Zenodo. https://doi.org/10.5281/zenodo.20163033

HERDEN, O. (2020). Architectural Patterns for Integrating Data Lakes into Data Warehouse Architectures. Big Data Analytics. BDA 2020. Lecture Notes in Computer Science(), vol 12581. Springer, Cham. https://doi.org/10.1007/978-3-030-66665-1_2

IPCC, 2023: Sections. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland, pp. 35-115, doi: 10.59327/IPCC/AR6-9789291691647

JANSSEN, N., Ilayperuma, T., Jayasinghe, J. et al. The evolution of data storage architectures: examining the secure value of the Data Lakehouse. J. of Data, Inf. and Manag. 6, 309–334 (2024). https://doi.org/10.1007/s42488-024-00132-1

MOHNA, Hosne Ara et al. AI-ready data engineering pipelines: a review of medallion architecture and cloud-based integration models. American Journal of Scholarly Research and Innovation, v. 1, n. 01, p. 319-350, 2022.

NAMBIAR, A, Mundra D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. Big Data and Cognitive Computing. 2022; 6(4):132. https://doi.org/10.3390/bdcc6040132

SIDDHARTHA, PARIMI. A Comparative Performance & Metadata Study of Open Table Formats: Iceberg vs Delta vs Hudi at Scale. Journal of Computer Science and Technology Studies, [S. l.], v. 7, n. 12, p. 513–520, 2025. DOI: 10.32996/jcsts.2025.7.12.56.

SCHNEIDER, J., Gröger, C., Lutsch, A. et al. The Lakehouse: State of the Art on Concepts and Technologies. SN COMPUT. SCI. 5, 449 (2024). https://doi.org/10.1007/s42979-024-02737-0.

ZIOTI, F.; Ferreira, K. R.; Queiroz, G. R.; Neves, A. K.; Carlos, F. M.; Souza, F. C.; Santos, L. A.; Simoes, R. E. O. A platform for land use and land cover data integration and trajectory analysis. International Journal of Applied Earth Observation and Geoinformation. V 106, P 102655, Feb 2022.