Instituto Nacional de Pesquisas Espaciais — INPE

Programa de Pós-Graduação em Computação Aplicada

Prototipagem de uma Arquitetura Lakehouse para Análise de Dados de Monitoramento Ambiental

Projeto de Pesquisa

Disciplina: Introdução à Geoinformática (CAP 395-3)

Discente: Sandro de Sena Machado

Orientador: Gilberto Ribeiro Queiroz

Co-orientadora: Karine Reis Ferreira

São José dos Campos

2026

1. Introdução

O Instituto Nacional de Pesquisas Espaciais (INPE) opera três sistemas complementares de monitoramento do desmatamento nos biomas Amazônia e Cerrado: o PRODES, que registra o desmatamento consolidado em ciclos anuais; o DETER, que emite alertas em tempo quase real; e o TerraClass, que classifica o uso e a cobertura da terra em levantamentos bienais. Embora compartilhem o mesmo domínio temático e operem sob a mesma instituição, esses sistemas foram concebidos de forma independente e não dispõem de infraestrutura analítica comum. A integração entre eles é realizada *ad hoc*, sem versionamento, sem rastreabilidade de transformações e sem mecanismos de reprodutibilidade.

A busca por integrar dados de monitoramento ambiental, revela desafios que vão além do volume de dados, como a variedade e velocidade dos dados, catalogação de metadados, otimização de armazenamento e consulta, entre outros. Schneider et al. (2024) identificam, por sua vez, cinco desafios estruturais que arquiteturas tradicionais de *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).

Diante desse contexto, este projeto investiga o paradigma *lakehouse* como alternativa arquitetural para integração dos sistemas de monitoramento ambiental do INPE.

2. Revisão Bibliográfica

2.1 Arquiteturas de dados e suas limitações no contexto geoespacial

Os *data warehouses* tradicionais baseiam-se no paradigma *schema-on-write*, o que garante forte governança e desempenho analítico, mas impõe rigidez estrutural. Errami et al. (2023) apontam limitações adicionais no contexto geoespacial: dificuldade de suportar diferentes representações geométricas para o mesmo fenômeno, incapacidade de lidar com dados de *streaming* e elevado custo de gerenciamento de metadados espaciais — que são semanticamente indispensáveis, pois um dado geoespacial sem sistema de referência de coordenadas pode produzir resultados analíticos incorretos. Os *data lakes*, por sua vez, oferecem flexibilidade via *schema-on-read*, mas introduzem os desafios C1–C5 descritos acima. A arquitetura dual (*two-tier*), que combina os dois paradigmas, atenua algumas limitações, mas gera duplicação de dados, *pipelines* complexos para manter ambas plataformas atualizadas e dificuldades de governança.

2.2 A arquitetura lakehouse

Armbrust et al. (2021) apresentam o *lakehouse* como plataforma unificada que combina a flexibilidade dos *data lakes* com a governança dos *data warehouses*, por meio de uma camada transacional de metadados sobre armazenamento de objetos de baixo custo. Schneider et al. (2024) formalizam o paradigma em oito requisitos técnicos independentes de fornecedor, que constituem os parâmetros de avaliação do protótipo deste projeto: armazenamento

e formato unificados (R1); suporte a operações CRUD completas sobre arquivos distribuídos (R2); representação relacional dos dados (R3); suporte a SQL com extensões espaciais (R4); garantias de consistência (R5); atomicidade e isolamento (R6); acesso direto ao dado pelo motor de processamento sem camadas intermediárias (R7); e integração nativa entre processamento *batch* e *streaming* (R8).

2.3 Lakehouse para dados geoespaciais

Errami et al. (2023) constituem a referência central deste projeto, sendo um dos únicos trabalhos encontrados na literatura que investiga sistematicamente a extensão do *lakehouse* para dados geoespaciais em periódico indexado. Os autores identificam quatro desafios que persistem mesmo no contexto *lakehouse*: (a) particionamento espacial, pois o particionamento alfanumérico convencional não respeita a proximidade geográfica, exigindo estratégias como Geohash ou XZ2 derivadas em tempo de ingestão; (b) gestão de metadados espaciais, para a qual recomendam o padrão ISO 19115 e o SpatioTemporal Asset Catalog (STAC); (c) heterogeneidade de integração entre fontes vetoriais e matriciais; e (d) computação sobre geometrias, na qual o uso de Minimum Bounding Rectangles (MBR) como filtro inicial é essencial para desempenho.

Uma distinção fundamental no contexto deste projeto diz respeito ao tratamento de dados matriciais. Diferentemente dos dados vetoriais — que podem ser armazenados nativamente em tabelas Iceberg, especialmente com os tipos

GEOMETRY e GEOGRAPHY introduzidos na especificação v3 —, dados matriciais como os do TerraClass não têm representação tabular viável.

Nesse caso a solução seria híbrida: os arquivos são armazenados como Cloud Optimized GeoTIFF (COG) no *object storage*, enquanto o Iceberg gerencia uma tabela de metadados com os *footprints* espaciais e os caminhos para os arquivos COG. Isso permite que o Sedona coordene a leitura eficiente dos arquivos em consultas que cruzam dados vetoriais e matriciais, sem duplicação e com rastreabilidade via Iceberg.

3. Justificativa

A justificativa deste projeto articula-se em três dimensões complementares. No plano da literatura, Schneider et al. (2024) optaram explicitamente por não considerar o trabalho de Zhang et al. (2023) em sua revisão, deixando em aberto avaliações sobre o tema de *lakehouses* geoespaciais. Já Errami et al. (2023) indicam que um dos caminhos de pesquisa em aberto, é a exploração do sistema *lakehouse* na criação de uma fonte única de repositório de dados geoespaciais.

No plano institucional, a coerência entre as três fontes escolhidas é um argumento adicional: PRODES, DETER e TerraClass são todos produtos do INPE, tornando o projeto uma demonstração de integração intra-institucional com valor prático direto.

O recorte temporal a partir de 2018 — com o levantamento TerraClass desse ano como linha de base e os de 2020 e 2022 como ciclos de acompanhamento — coincide com o período posterior a agosto de 2019, quando entraram em vigor as novas regras de crédito rural (Resolução CMN nº 5.268 de 18/12/2025) vinculadas ao CAR (Cadastro Ambiental Rural), conferindo relevância regulatória à série temporal adotada.

4. Objetivos

4.1 Objetivo Geral

Desenvolver e avaliar, à luz dos requisitos formais do paradigma *lakehouse* definidos por Schneider et al. (2024), um protótipo local de arquitetura *lakehouse* para integração e análise de dados geoespaciais heterogêneos dos sistemas PRODES, DETER e TerraClass do INPE, cobrindo os biomas Amazônia e Cerrado no período 2018–2022.

4.2 Objetivos Específicos

I. Implementar um *pipeline* de ingestão dos dados vetoriais do PRODES e DETER na camada Bronze, preservando os metadados espaciais originais;

II. Converter os dados matriciais do TerraClass para COG e implementar a tabela de metadados correspondente em Iceberg;

III. Construir a camada *Silver* com reprojeção para EPSG:5641 (SIRGAS 2000 / South America Albers Equal Area Conic), normalização de atributos e particionamento por Geohash;

IV. Construir a camada *Gold* à partir do cruzamento entre diferentes bases de dados de monitoramento ambiental;

V. Avaliar o protótipo à luz dos requisitos R1–R8, documentando quais são atendidos, parcialmente atendidos ou não atendidos no contexto geoespacial.

5. Metodologia

5.1 Conjuntos de dados

Os dados compreendem três conjuntos: os polígonos anuais de desmatamento do PRODES e os alertas do DETER, disponíveis nos repositórios públicos do INPE para Amazônia e Cerrado a partir de 2018; e os levantamentos de uso e cobertura da terra do TerraClass para 2018, 2020 e 2022, em formato GeoTIFF. O TerraClass de 2018 funcionará como linha de base anterior ao marco regulatório de agosto de 2019, e os levantamentos de 2020 e 2022 como ciclos de acompanhamento.

5.2 Arquitetura do protótipo

A arquitetura é composta por quatro componentes integrados: (a) MinIO como *object storage* local, simulando a camada de armazenamento do lakehouse; (b) Apache Iceberg como *open table format*, responsável pelos requisitos R1–R7 via camada transacional de metadados; (c) Apache Spark com extensão Apache Sedona como motor de processamento analítico e SQL espacial (R4 e R8); e (d)

GDAL para ingestão dos formatos de origem e conversão dos arquivos vetoriais para Parquet e matriciais para COG.

5.3 Organização em camadas

Foi adotada uma arquitetura medalhão para organizar os dados em camadas, típica de lakehouses. Assim, a camada Bronze recebe os dados em formato Parquet (vetoriais) ou COG (raster), preservando o CRS original de cada fonte. Para o TerraClass, é criada uma tabela Iceberg de metadados com o *footprint* de cada coleção e o caminho para os arquivos COG no MinIO. A camada *Silver* aplica as transformações de padronização: reprojeção para EPSG:5641 — sistema projetado que preserva área, propriedade essencial para análise de desmatamento em escala regional —, normalização de atributos, tratamento de geometrias inválidas e particionamento por Geohash. A camada *Gold* disponibiliza os dados integrados entre as três fontes, com os cruzamentos espaciais pré-computados e as agregações territoriais necessárias para as consultas.

5.4 Consultas e avaliação

Serão implementadas duas consultas representativas: (a) sobreposição entre polígonos PRODES e classes TerraClass para identificação das coberturas convertidas em cada intervalo bienal; e (b) agregação de alertas DETER por unidade territorial — municípios, terras indígenas e unidades de conservação — no período 2018–2022. A avaliação seguirá dois critérios: verificação qualitativa dos requisitos R1–R8 com documentação do grau de atendimento de cada um, e

comparação de desempenho de consultas de range sobre dados com e sem particionamento espacial por Geohash.

6. Resultados esperados

No plano da literatura, o projeto contribui com evidências empíricas sobre a manifestação concreta dos desafios do *lakehouse* geoespacial em um *dataset* institucional real, e com a aplicação do requisitos R1–R8 de Schneider et al. (2024) a um contexto geoespacial. No plano técnico, experimenta a construção de um protótipo sob um novo paradigma de arquitetura de dados geoespaciais. No plano institucional, a demonstração de que PRODES, DETER e TerraClass podem ser integrados em um ambiente analítico reprodutível e versionado tem valor prático direto para o INPE e serve de base para expansão em trabalhos futuros.

Referências

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).

ERRAMI, A., Hajji H, Ait El Kadi K, Badir H. Spatial big data architecture: from data warehouses and data lakes to the lake-house. J Parall Distrib Comput. 2023;176:70–9. https://doi.org/10.1016/j.jpdc.2023.02.007.

INPE. PRODES: Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite. Disponível em:

https://data.inpe.br/biomasbr/prodes-monitoramento-anual-da-supressao-de-vegetacao-nativa/. Acesso em: 2026.

INPE. DETER: Detecção de Desmatamento em Tempo Real. Disponível em: http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/deter/deter. Acesso em: 2026.

INPE. TerraClass: mapeamento do uso e cobertura da terra na Amazônia e no Cerrado. Disponível em: https://www.terraclass.gov.br. Acesso em: 2026.

ISO. ISO 19115-1: Geographic information — Metadata. Geneva: International Organization for Standardization, 2014. Disponível em: https://www.iso.org/standard/53798.html. Acesso em: 2026.

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.)

ZHANG, Y., Peng B., Du Y., Su J. GeoLake: bringing geospatial support to lakehouses. IEEE Access. 2023;11:143037–49. https://doi.org/10.1109/ACCESS.2023.3343953.