Data Lakehouses para a análise de dados geoespaciais em larga escala

Felipe F. Vasconcelos[1], **Fábio J. Coutinho**[2]

[1]Instituto de Ciências Matemáticas e Computação – Universidade de São Paulo

[2]Instituto de Computação – Universidade Federal de Alagoas

felipevsc@usp.br,fabio@ic.ufal.br

Abstract. *Data Warehouses and Data Lakes are architectures capable of handling complex analyses, however, the increase in geospatial data generation, driven by the Internet of Things, highlights the limitations of both architectures. Data Lakehouses emerge as the new state-of-the-art for Big Data storage, offering an integrated and cost-effective solution. This paper proposes the use of Data Lakehouses for a Big Geospatial Data storage and analysis environment. In addition, a case study with geolocation data of municipal buses was conducted to demonstrate the feasibility of the proposed environment.*

Resumo. *Data Warehouses e Data Lakes são arquiteturas capazes de lidar com análises complexas, entretanto, o aumento da geração de dados geoespaciais, impulsionado pela Internet das Coisas, evidencia limitações de ambas arquiteturas. Os Data Lakehouses surgem como o novo estado-da-arte para armazenamento de dados em larga escala, ofertando uma solução integrada de baixo custo. Este artigo propõe a utilização de Data Lakehouses para um ambiente de armazenamento e análise de dados geoespaciais em larga escala. Além disso, foi implementado um estudo de caso com dados de geolocalização de ônibus municipais para demonstrar a viabilidade do ambiente proposto.*

1. Introdução

Em 2025, estima-se que 175 *zettabytes*[1] de dados sejam gerados por dia e que cerca da metade desse montante seja produzida a partir de dispositivos IdC [Reinsel et al. 2018]. Lidar com esse volume de dados requer arquiteturas que sejam capazes de prover escalabilidade, flexibilidade e interoperabilidade, levando em consideração questões como a governança dos dados e o custo operacional.

Dados geoespaciais são dados que contêm informações acerca de um objeto, fato ou fenômeno associado a uma latitude e longitude, sendo representados por pontos ou polígonos. A manipulação desses dados em larga escala exige alguns requisitos que constituem um desafio comum para as principais arquiteturas de armazenamento existentes. A implementação de ambientes para esta finalidade deve considerar fatores como heterogeneidade dos dados, particionamento, armazenamento e consultas espaciais eficientes. Ao mesmo tempo, também deve ser capaz de prover interoperabilidade com diferentes ferramentas, como soluções para análise e visualização de dados geoespaciais.

[1]um *zettabyte* equivale a 10^6 *petabytes*

As arquiteturas de armazenamento Data Lakes (DL) e Data Warehouses (DW) apresentam limitações para o armazenamento e análise de grandes volumes de dados, sejam geoespaciais ou convencionais. Para superar essas limitações, os Data Lakehouses (DLH) emergem como o novo estado da arte em armazenamento, oferecendo uma solução unificada. Na literatura, são encontrados os trabalhos de [Errami et al. 2023] e [Mete 2023] cuja abordagem faz uso de DLH como solução para ambientes de dados geoespaciais. Em ambos os trabalhos, os autores discutem e propõem arquiteturas DLH, porém, não efetivam sua implementação.

Este trabalho propõe a utilização de Data Lakehouse como uma solução eficiente e econômica para a construção de um ambiente de armazenamento e análise de dados geoespaciais em larga escala. O ambiente proposto é constituído de quatro módulos, que são implementados a partir de um estudo de caso para a análise do comportamento de ônibus municipais, com o intuito de validar a utilização de DLH. O documento encontra-se organizado da seguinte forma: a seção 2 trata das arquiteturas de armazenamento; a seção 3 discute o uso de DLH para armazenar dados geoespaciais; a seção 4 apresenta a proposta do ambiente utilizando DLH; a seção 5 apresenta o estudo de caso implementado e a seção 6 discute as considerações finais e trabalhos futuros.

2. As Arquiteturas de Armazenamento e o contexto dos Dados Geoespaciais

A maior parte dos dados geoespaciais existentes são obtidos, em tempo real, a partir de diferentes dispositivos IdC [Errami et al. 2023]. A geração de dados a partir de inúmeros sensores e dispositivos de diferentes tipos confere aos dados geoespaciais grande volume e heterogeneidade. Os paradigmas de armazenamento mais conhecidos, como DW e DL, não são capazes de lidar de forma eficiente com a manipulação e análise de dados geoespaciais, apresentando limitações relacionadas à heterogeneidade de formatos, manutenção da espacialidade e uso de metadados espaciais [Hassan 2024].

Os DW Espaciais apresentam dificuldades relacionadas ao processamento de dados em larga escala, tais como falta de escalabilidade, flexibilidade e ausência de tratamento a fluxos de dados (*data streams*). Atender a essas demandas é de fundamental importância considerando o grande volume de dados gerados por dispositivos IdC. Ademais, os DW apresentam um ambiente de armazenamento e computação integrado, não sendo possível escalar de forma independente a etapa de computação, fator crítico para dados geoespaciais [Armbrust et al. 2021] [Errami et al. 2023].

Os DL solucionam problemas encontrados nos DW, entretanto, novos desafios surgem, como a falta de suporte nativo para armazenamento e consultas de dados geoespaciais, além da falta de estruturas nativas de indexação espacial que possam melhorar o desempenho das consultas. Além disso, fatores como a governança e a qualidade dos dados são temas recorrentes em DL e que devem ser levados em consideração.

A arquitetura DLH surge como fruto de uma intersecção de funcionalidades entre um Data Lake e um Data Warehouse (Lake + House). DLHs possuem uma arquitetura que alia os princípios de baixo custo e escalabilidade dos Data Lakes com características de Data Warehouse como gerenciamento, propriedades ACID, versionamento e outros. Desta forma, as plataformas DLH atendem demandas como volume intenso e heterogeneidade de dados com o objetivo de prover uma plataforma única para todos os dados existentes, entregando custo-benefício, desempenho, governança e formatos abertos.

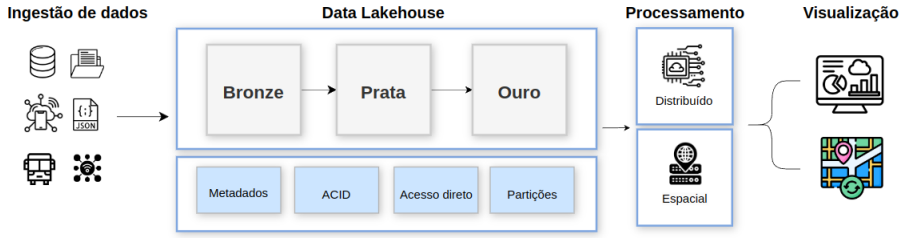

Figura 1. Proposta de ambiente de análise de dados geoespaciais usando DLH.

3. Data Lakehouse para Dados Geoespaciais

Os Data Lakehouses se apresentam como um ambiente integrado capaz de se adequar a natureza volátil e volumosa de diferentes contextos Big Data. A arquitetura clássica de Lakehouse proposta por Armbrust inclui 3 camadas: i) armazenamento em DL; ii) metadados, ofertando governança, ACID e versionamento; iii) APIs de interação [Armbrust et al. 2021].

A **camada I** dos DLH consegue suprir as demandas relacionadas à heterogeneidade e fluxos de dados gerados a partir de dispositivos IdC. Essa camada também permite o acesso direto e o uso de formatos abertos, o que possibilita tipos nativos de dados geoespaciais por meio de soluções como SpatialParquet e GeoParquet. A **camada II** fornece a governança necessária para trazer confiabilidade aos dados geoespaciais, ofertando metadados que podem ser utilizados para otimizações e o suporte a metadados espaciais. A **camada III** oferta a separação dos ambientes de armazenamento e computação, o que no contexto de computação intensiva de dados geoespaciais é de extrema importância. Essa camada também permite a implementação de técnicas de indexação e particionamento espaciais com eficiência similar à encontrada em DW.

Assim, conclui-se que os Data Lakehouses suprem as principais necessidades de ambientes unificados para armazenamento e análise de dados espaciais, respeitando a dimensão espacial e a característica de computação intensiva. Assim, a utilização de DLH em ambientes voltados para a análise de dados geoespaciais permitem melhorar as condições de funcionamento em diferentes casos de uso, por exemplo, analisar dados urbanos massivos no contexto de Cidades Inteligentes.

4. Proposta de Ambiente de Análise de Dados Geoespaciais usando DLH

Esta seção descreve uma proposta de ambiente para a análise de dados geoespaciais utilizando DLH como solução de armazenamento. Uma representação gráfica do ambiente pode ser visualizada na Figura 1, reunindo quatro módulos: Ingestão de Dados, Lakehouse, Processamento e Visualização. A representação de [Errami et al. 2023] adaptada neste trabalho visa incorporar soluções para o processamento e armazenamento de dados geoespaciais.

O **Módulo de Ingestão de Dados** é responsável por controlar a entrada de dados no ambiente. Esse processo não é trivial, visto que precisa considerar questões como volume, variedade e velocidade, as quais podem ser diferentes de acordo com o cenário abordado. Por exemplo, considerando o contexto de um ambiente para analisar dados de uma cidade inteligente, pode-se produzir uma gama de dados com características e deman-

das distintas mediante aplicações voltadas para monitoramento do trafégo de veículos, previsão de tempo, entre outras.

Em [Vasconcelos et al. 2023], são comparadas duas ferramentas para ingestão de dados: Apache Kafka e Apache Nifi. Os autores concluem ainda que o Apache Kafka é uma solução mais abrangente, pois permite a integração com uma maior diversidade de ferramentas e dispositivos. Esse fator é relevante para a análise de dados geoespaciais com DLH, visto que favorece a integração de diversas bases e fontes de dados. Portanto, a ferramenta Apache Kafka é indicada como solução para a implementação do módulo de ingestão de dados proposto neste trabalho.

O **Módulo de Armazenamento Lakehouse** é implementado por meio de um framework de Data Lakehouse, sendo o responsável por todo o armazenamento e governança do ambiente, implementando propriedades e operações como acesso direto, particionamento e transações ACID. O ambiente deve ser capaz de armazenar nativamente formatos de dados geoespaciais, para isso, propõe-se a utilização do GeoParquet, que adiciona o suporte a tipos geoespaciais como pontos, linhas e polígonos para o Parquet[2]. A utilização do GeoParquet permite uma melhor interoperabilidade entre sistemas, aumentando a possibilidade de uso de ferramentas de análise de dados espaciais. Ele também apresenta capacidade de distribuição, ao contrário do GeoJSON, além de permitir uma melhor compressão e a existência de arquivos maiores do que em formatos como o Shapefile [Medina et al. 2023].

O módulo de armazenamento também deve ser capaz de comportar soluções de metadados, incluindo espaciais, que requerem uma maior atenção, visto que armazenam atributos adicionais, relacionados à referência espacial e aos objetos geométricos [Errami et al. 2023]. Esses metadados podem ser incorporados no DLH por meio de catálogos como o STAC[3]. Ademais, a utilização das arquiteturas de dados disponíveis em DLH (e.g. Medallion em Delta Lakehouse) facilitam o gerenciamento dos dados e dos fluxos de transformação, permitindo a existência de diversos níveis de refinamentos nos dados[4]. Soluções voltadas para segurança e privacidade dos dados também devem ser discutidas e implementadas, entretanto, essas questões não serão abordadas neste trabalho.

O **Módulo de Processamento** é responsável por realizar o processamento eficiente das consultas espaciais e não espaciais. De acordo com as arquiteturas DLH, existe uma separação entre armazenamento e processamento. Esse princípio favorece a coexistência de diferentes tipos de operações no ambiente, como consultas espaciais e não espaciais, à medida que permite a criação de clusters específicos de acordo com a finalidade esperada. Em se tratando de processar dados geoespaciais volumosos, o módulo de processamento pode prover execução distribuída e paralela a partir de técnicas de indexação e particionamento adaptadas para o contexto espacial.

Segundo [de Carvalho Castro et al. 2020] e [Mete 2023], Apache Sedona atende a todas as necessidades de uma ferramenta de análise espacial, ofertando um melhor desempenho que seus concorrentes e facilitando a integração com o ambiente de DLH.

[2]https://geoparquet.org/

[3]https://stacspec.org/

[4]Camadas "Bronze", "Prata"e "Ouro"da Figura 1 representam uma arquitetura Medallion.

[Errami et al. 2022] apresentam um teste de desempenho para diferentes abordagens de indexação e particionamento em um ambiente DLH, demonstrando a capacidade dos DLH para prover técnicas de otimização de dados geoespaciais.

O **Módulo de Visualização** é responsável por tornar acessível para as partes interessadas as informações coletadas e analisadas. Os motores de processamento como Spark e Sedona conseguem entregar uma diversidade de consultas e análises, todavia, se limitam as APIs de *DataFrame* e SQL para demonstrar os resultados. O acesso direto aos dados permitidos pelo DLH, em conjunto com o uso de soluções de armazenamento como o GeoParquet, conferem uma interoperabilidade para o ambiente que permite a utilização de bibliotecas como Matplotlib e programas como QGIS[5] e ArcGIS[6], focados em dados geoespaciais e amplamente utilizados por especialistas.

5. Implementação do Ambiente a partir de um Estudo de caso

Esta seção apresenta um estudo de caso sobre a análise do comportamento de ônibus municipais de três cidades brasileiras, Brasília, Rio de Janeiro e São Paulo. Para tal, foi utilizado um conjunto de dados de geolocalização de ônibus de cidades brasileiras baseado no trabalho de [Melo et al. 2023]. O objetivo deste estudo de caso é discutir a capacidade analítica do ambiente, de modo que aspectos como governança, escalabilidade e desempenho não foram estudados. A utilização desse conjunto de dados possibilita a replicação de casos de uso reais, explorando a temática da análise de dados da movimentação de ônibus municipais, conforme abordado na literatura em trabalhos como [Queiroz et al. 2019]

A Figura 2 apresenta um esquema representativo das soluções de código aberto utilizadas para a implementação do ambiente, as quais foram escolhidas de acordo com as discussões apresentadas na seção anterior. Como solução de framework de Data Lakehouse, responsável por implementar as camadas e propriedades como metadados, armazenamento, transações ACID e particionamento, foi escolhida a plataforma Delta Lake. A sua escolha foi motivada por apresentar um desempenho superior a outras plataformas, uma grande aceitação e uso na literatura, além de uma boa integração com diferentes soluções possibilitada pelo uso do Apache Spark com seu motor de processamento [Jain et al. 2023] [Errami et al. 2023].

O fluxo de ferramentas utilizado para implementar este ambiente de análise de dados geoespaciais permitiu extrair dados analíticos acerca do estudo de caso considerado, alguns dos quais estão representados nas Figuras 3 e 4 e são explicados a seguir.

A Figura 3 apresenta uma análise da velocidade média nas cidades de Brasília e Rio de Janeiro, no dia 05/02/2024 (segunda-feira). O gráfico considera apenas os veículos com uma velocidade maior que zero. Percebe-se a variação da velocidade durante o dia, visto que o horário de pico leva à diminuição da velocidade média em ambas as cidades. Também percebe-se que, na maior parte do dia, o Rio de Janeiro apresenta uma velocidade média superior a Brasília.

A Figura 4 apresenta o mapa de calor da região central da cidade de São Paulo gerado por meio da biblioteca PyDeck[7], integrada ao Apache Sedona. O mapa de calor é

[5]https://qgis.org/
[6]https://www.arcgis.com/index.html
[7]https://deckgl.readthedocs.io/en/latest/

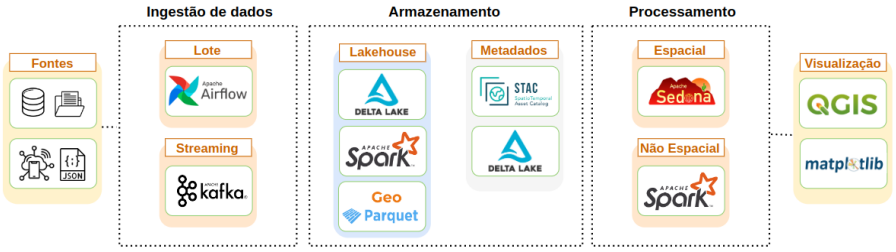

Figura 2. Fluxo de ferramentas utilizadas para para implementar o ambiente proposto.

referente ao dia 05/02/2024 (segunda-feira) às 12:00 horas e está centralizado na região da Estação da Sé. É possível perceber a região da Sé com a maior concentração de ônibus, sendo uma das principais localidades da região central da cidade. Também é possivel visualizar a diferença entre o fluxo de ônibus em avenidas principais e em ruas secundárias, indicando que os ônibus realizam mais trajetos em vias principais.

A partir das análises realizadas, conclui-se que o ambiente proposto, baseado na arquitetura DLH, é viável para ser utilizado para armazenamento e análise de dados geoespaciais. O ambiente DLH consegue reproduzir consultas e análises que seriam feitas normalmente em ambientes de DL ou DW. Também foi possível demonstrar a interoperabilidade do ambiente proposto e da arquitetura DLH com diversas bibliotecas e ferramentas voltadas para análise e visualização de dados geoespaciais e não espaciais.

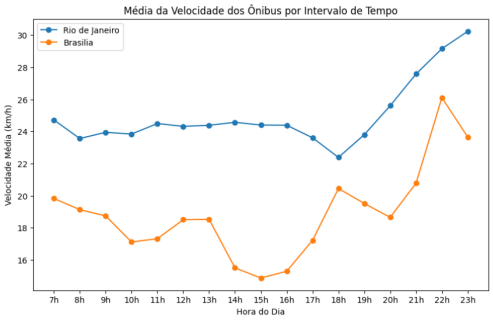

Figura 3. Média das velocidades médias de ônibus em Brasília e Rio de Janeiro.

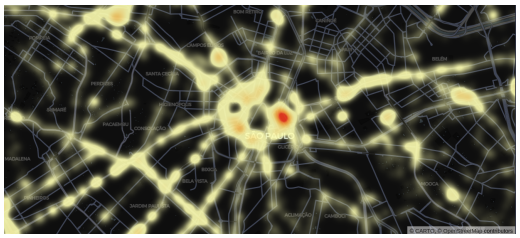

Figura 4. Mapa de calor dos ônibus na região central de São Paulo.

6. Considerações Finais

Este trabalho apresentou uma proposta de um ambiente para armazenamento e análise de dados geoespaciais baseado no conceito de Data Lakehouse, que busca ofertar um Data Lake com governança e funcionalidades de um Data Warehouse. O ambiente e fluxo propostos servem como linha de base para projetos de análise de dados envolvendo dados espaciais e não espaciais, possibilitando um armazenamento eficiente, de baixo custo e com capacidades analíticas semelhantes a Data Warehouses. Foi implementado um ambiente de análise com DLH a partir um estudo de caso utilizando dados de geolocalização de ônibus municipais, com o objetivo de reproduzir cenários reais e demonstrar a viabilidade do uso de DLH para a análise de dados geoespaciais.

Referências

[Armbrust et al. 2021] Armbrust, M., Ghodsi, A., Xin, R., and Zaharia, M. (2021). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR*, volume 8, page 28.

[de Carvalho Castro et al. 2020] de Carvalho Castro, J. P., Chaves Carniel, A., and Dutra de Aguiar Ciferri, C. (2020). Analyzing spatial analytics systems based on Hadoop and Spark: A user perspective. *Software: Practice and Experience*, 50(12):2121–2144.

[Errami et al. 2023] Errami, S. A., Hajji, H., El Kadi, K. A., and Badir, H. (2023). Spatial big data architecture: from data warehouses and data lakes to the Lakehouse. *Journal of Parallel and Distributed Computing*, 176:70–79.

[Errami et al. 2022] Errami, S. A., Hajji, H., Kadi, K. A. E., and Badir, H. (2022). Managing Spatial Big Data on the Data LakeHouse. In *International Conference on Networking, Intelligent Systems and Security*, pages 323–331. Springer.

[Hassan 2024] Hassan, I. (2024). Storage structures in the era of big data: From data warehouse to lakehouse. *Journal of Theoretical and Applied Information Technology*, 102(6).

[Jain et al. 2023] Jain, P., Kraft, P., Power, C., Das, T., Stoica, I., and Zaharia, M. (2023). Analyzing and Comparing Lakehouse Storage Systems. In *13th Conference on Innovative Data Systems Research, CIDR*.

[Medina et al. 2023] Medina, A., Mosquera, D., and Gallegos, F. A. (2023). A Methodological Approach for Data Collection and Geospatial Information of Healthy Public Spaces in Peripheral Neighborhoods—Case Studies: La Bota and Toctiuco, Quito, Ecuador. *Sustainability*, 15(21):15553.

[Melo et al. 2023] Melo, R. T., Vasconcelos, F. F., Silva, R. L. L., Santos, P. V., Ramos, V. T., and Coutinho, F. J. (2023). BRBus-construindo um dataset para monitoramento geoespacial dos ônibus de cidades brasileiras. In *Anais do V DSW*. SBC.

[Mete 2023] Mete, M. (2023). Geospatial Big Data Analytics for Sustainable Smart Cities. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:141–146.

[Queiroz et al. 2019] Queiroz, A. R. M., Santos, V. B., Nascimento, D. C., and Pires, C. E. S. (2019). Conformity analysis of GTFS routes and bus trajectories. In *Anais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 199–204. SBC.

[Reinsel et al. 2018] Reinsel, D., Gantz, J., and Rydning, J. (2018). The Digitization of the World, from Edge to Core. *Relatório Técnico. An IDC White Paper-US44413318, Sponsored by Seagate*.

[Vasconcelos et al. 2023] Vasconcelos, F. F., Ramos, V. T., and Coutinho, F. J. (2023). Os desafios e soluções para a implementação de Big Data Analytics em cidades inteligentes. In *Anais Estendidos do XXXVIII SBBD*. SBC.