CONEXÃO UNIFAMETRO 2025
XXI SEMANA ACADÊMICA
ISSN: 2357-8645

DATA WAREHOUSE, DATA LAKE E DATA LAKEHOUSE: FUNDAMENTOS E PERSPECTIVAS

Francisco Jonas Pereira Lino
Monitor Bolsista – Análise e Desenvolvimento de Sistemas
jonasplino@gmail.com
Eduardo Julião Máximo
Orientador – Professor Msc. Em Informática Aplicada
eduardo.maximo@professor.unifametro.edu.br
Marcondes Josino Alexandre
Orientador – Professor Msc. Em Computação Aplicada
marcondes.alexandre@professor.unifametro.edu.br

Área Temática: Engenharia de Software e Computação em Nuvem
Área de Conhecimento: Ciências Tecnológicas
Encontro Científico: XIII Encontro de Monitoria

RESUMO

Introdução: Data warehouses, data lakes e data lakehouses representam soluções distintas para armazenamento e análise de dados em organizações. **Objetivo:** Analisar os fundamentos, arquiteturas, benefícios, desafios e perspectivas dessas tecnologias, destacando a importância da governança e adequação à LGPD. **Metodologia:** Revisão bibliográfica de publicações e artigos recentes sobre o tema. **Resultados:** O estudo evidenciou que data warehouses são eficientes para análises estruturadas, com dados limpos e organizados, enquanto data lakes oferecem armazenamento flexível para dados estruturados, semiestruturados e não estruturados, mas exigem governança rigorosa para evitar pântanos de dados. Data lakehouses unem os benefícios de DW e DL, oferecendo suporte a consultas rápidas, BI e inteligência artificial, mas exigem maturidade tecnológica e política de governança. **Considerações finais:** A implementação bem-sucedida depende do levantamento de requisitos, da maturidade da equipe e da adequação às normas de proteção de dados, sendo fundamental documentação, políticas de ciclo de vida e governança para garantir qualidade e segurança da informação.

Palavras-chave: Data Warehouse; Data Lake; Data Lakehouse; LGPD.

INTRODUÇÃO

O crescimento exponencial de dados nas organizações exige soluções eficientes para armazenamento, integração e análise. Os DWs surgem como bancos de dados estruturados voltados para análises OLAP, consolidando informações de múltiplas fontes para suporte à decisão (RAMAKRISHNAN et al., 2008; DATE, 2004). Data marts, subconjuntos focados em

departamentos específicos, exemplificam uma segmentação de DW (ELMASRI; NAVATHE, 2010).

Com o aumento do *big data*[1], os DL emergiram como repositórios de dados em seu formato bruto, estruturados, semiestruturados ou não estruturados, usando abordagem schema-on-read. Entretanto, a flexibilidade exige governança rigorosa para evitar desorganização e garantir qualidade.

Os DLH de acordo com Harby e Zulkernine (2022) são um casamento, combinam a flexibilidade do Data Lake com a performance e organização do Data Warehouse, integrando e viabilizando a análise de BI, aprendizado de máquina e inteligência artificial.

O presente estudo objetiva analisar fundamentos, arquiteturas, benefícios e desafios dessas tecnologias, destacando a necessidade de políticas de governança, documentação e adequação à LGPD para garantir segurança, integridade e uso ético dos dados.

METODOLOGIA

Realizou-se revisão bibliográfica em livros, artigos e publicações especializadas sobre DW, DL e DLH, com ênfase em aspectos de arquitetura, benefícios, desafios e governança. A análise incluiu a verificação das melhores práticas de gestão de dados, ciclos de vida e conformidade com normas de proteção de dados.

RESULTADOS E DISCUSSÃO

Um DW é um tipo de banco de dados projetado especificamente para consultas e análises eficientes (OLAP[2]). Ele requer que os dados sejam limpos e organizados antes do armazenamento, resultando em dados do tipo estruturado. Para Ramakrishnan et al. (2008), os data warehouses contêm dados consolidados de muitas fontes, ampliados com informações de resumo e cobrindo um longo período. A ampliação com informações de resumo, retoma o conceito de enriquecimento, amplamente difundido em formações para profissionais de dados. Uma definição correlata que tem ganhado popularidade são os Data marts, que são DWs “voltados para um subconjunto da organização, como um departamento, e possuem um foco mais estreito” (ELMASRI; NAVATHE, 2010, p. 722).

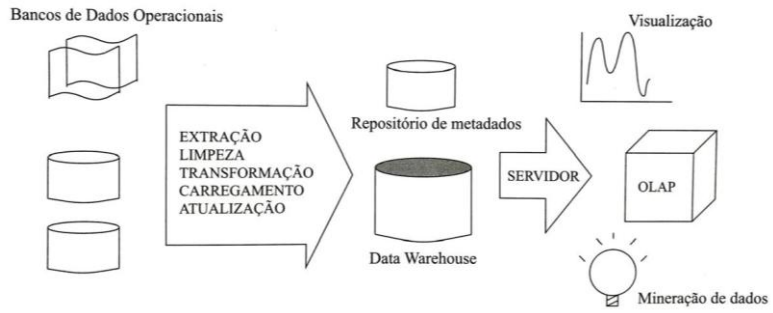
[1] Big Data: conjuntos de dados muito grandes, complexos e variados, que exigem tecnologias avançadas para armazenamento, processamento e análise, permitindo extrair insights estratégicos.
[2] OLTP vs OLAP: OLTP lida com transações do dia a dia, dados normalizados e operações rápidas; OLAP lida com análises, grandes volumes históricos e consultas complexas.

ISSN: 2357-8645

Alguns dos principais motivos que podem levar à implementação do DW em uma organização são: grande volume de dados; necessidade de desempenho de consulta analítica; exigência de integração de dados de várias fontes e o requisito de oferecer suporte à inteligência de negócio (BI[3]). Os sistemas de apoio à decisão são sistemas que ajudam na análise de informações do negócio. Sua meta é ajudar a administração a "definir tendências, apontar problemas e tomar […] decisões inteligentes" (DATE, 2004, p. 590). Affeldt e Silva Júnior (2013) apontam que existem duas causas principais para as dificuldades e limitações relacionadas a obtenção de indicadores de desempenho: sistemas que não contêm as informações necessárias e o perfil de uso de ferramentas analíticas. Observasse que informações relevantes para solucionar os problemas dos negócios ainda não foram levantadas e que os diretores, gestores ou executivos ainda possuem uma cultura orientada a dados. Como uma saída para essas dificuldades, possuir o entendimento dos requisitos do negócio deve ser a primeira etapa em um plano de ação para implementar um DW. Nessa etapa, serão levantados insumos suficientes para definir escopo, esquema de dados e identificar a real necessidade de implementação.

Para efeitos de ilustração, dentre as diversas arquiteturas analisadas, a apresentada na Figura 1, criada por Ramakrishnan et al. (2008, p. 723) destacou-se por contribuir significativamente para o entendimento lógico de um Data Warehouse.

Figura 1 – Arquitetura típica de um Data Warehouse



Fonte: RAMAKRISHNAN; GEHRKE; TANIWAKE; TORTELLO; SOUSA, 2008, p. 723.

Como mencionado no início, o DW requer que os dados sejam limpos e organizados antes do armazenamento e "as diferentes etapas envolvidas na obtenção de dados para um

[3] Business Intelligence (BI): práticas e tecnologias que permitem coletar, analisar e apresentar dados de uma empresa para apoiar decisões estratégicas.

ISSN: 2357-8645

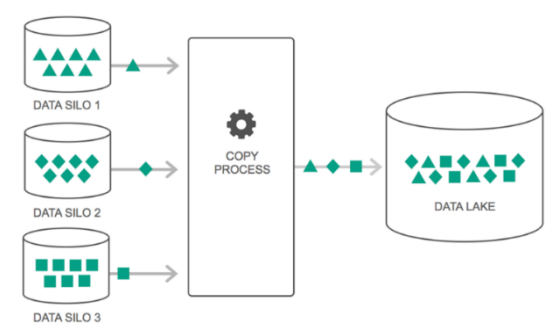
depósito de dados são chamadas de tarefas *extract, transform and load* (ETL)" (SILBERSCHATZ; KORTH; SUDARSHAN, 2012, p. 562). Essas etapas conduzem a uma base de OLAP íntegra, coesa e sem inconsistências graves, como por exemplo, os *outliers*, que são valores atípicos em um conjunto de dados, ou seja, pontos que se desviam significativamente do padrão ou da distribuição geral.

Conforme atenuado por Bentaib et al. (2024), a literatura sobre Data Lake é um pouco confusa e escassa, como também não oferece uma ideia consistente ou plano de implementação abrangente. De todo modo, por definição, um Data Lake permite o armazenamento de grandes quantidades de dados no seu formato bruto para posterior processamento e organização. Utilizam uma abordagem de *schema-on-read*, o que significa que não é aplicado um formato padrão aos dados quando são armazenados.

Pelo fato de os DL armazenarem os dados em seu formato nativo, isso permite que sejam armazenados dados estruturados (por exemplo, SQL), não estruturados (por exemplo, imagens, PDFs, texto) e semiestruturados (por exemplo, XML, JSON).

A arquitetura de um Data Lake pode ser concebida conforme a Figura 2.

Figura 2 – Arquitetura típica de um Data Lake



Fonte: DS Academy, 2024. Disponível em: https://blog.dsacademy.com.br/os-4-estagios-para-construir-um-data-lake-de-forma-eficiente/. Acesso em: 24 ago. 2025.

Essa tecnologia surgiu como forma de dar vazão às demandas organizacionais no que se refere ao desencadeamento de big data no final da década de 2000 e início de 2010.

Entre os principais usuários dessa tecnologia estão os cientistas de dados, engenheiros de dados e arquitetos de dados. As principais habilidades necessárias incluem gerenciamento de infraestrutura em nuvem, experiência em segurança e conformidade além de familiaridade com ferramentas de big data como Spark.

Quatro benefícios foram identificados no uso dos data lakes: a redução do esforço inicial através do armazenamento de dados; melhor aquisição de dados, acesso rápido a dados brutos e preservação de dados (LLAVE, 2018).

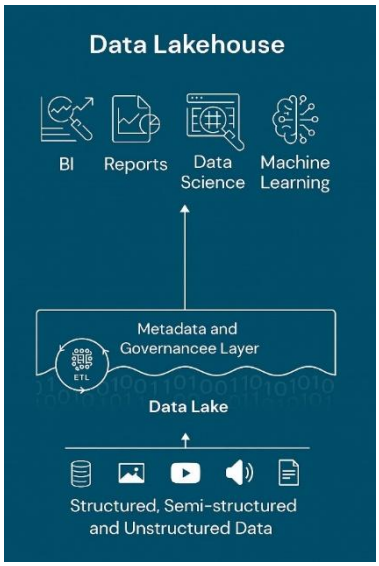
Mas também alguns desafios foram mencionados, tais como a problemas de qualidade dos dados, pois, como os dados são ingeridos na forma bruta, os dados de baixa qualidade ou incompletos podem não ser detectados até o final do processo. Um dos principais problemas identificados é quando um DL se transforma em um pântano de dados, isso ocorre, pois, o DL se torna desorganizado e inutilizável devido ao gerenciamento e à governança dos dados, o que remete ao conceito *ELT* (*Extract, Load, Transform*).

Existem atualmente boas práticas que dão vazão aos problemas oriundos da má utilização dos data lakes, como o uso de estruturas padronizadas de nomes e pastas, a automação de políticas de ciclo de vida e a exclusão automática dos dados obsoletos após um período de retenção definido.

Um Data Lakehouse, de acordo com Errami et al. (2023), fornece uma plataforma unificada para as cargas de trabalho de dados, sem comprometer a governança e o desempenho dos DW ou a flexibilidade e a economia dos DL. Assim como um DW, um DLH é compatível com consultas rápidas e análise de dados otimizada, sendo também uma fonte ideal para AI. A questão é que nem todas as empresas estão usando AI, nesse sentido, o levantamento de requisitos é essencial apontar a real necessidade de implementação de um DLH.

Em relação à arquitetura de um DLH, a apresentada na Figura 3, desenvolvida por Kaplan e Kara (2025), é adequada para fins de ilustração. A interpretação dessa arquitetura pode ser dividida em camadas: camada de ingestão; camada de metadados e governança e por fim camada de consumo.

Figura 3 – Arquitetura típica de um Data Lakehouse



Fonte: KAPLAN, Ari; KARA, Amit. *The Data LakehouseFor Dummies, Databricks Special Edition*. Hoboken:

John Wiley & Sons, 2025. Disponível em: https://www.databricks.com/resources/ebook/the-data-lakehouse-for-dummies.

Entre os desafios identificados, constatou-se que, embora os DLHs possam otimizar fluxos de trabalho de dados, pode ser complicado colocá-los em funcionamento. Os usuários podem ter que passar por uma curva de aprendizado, pois o seu uso pode ser diferente dos DWs com os quais estão acostumados. Além disso os DLHs também são uma tecnologia relativamente nova, cujo framework ainda está evoluindo.

A Lei Geral de Proteção de Dados (LGPD) regulamenta o tratamento de informações pessoais, identificadas, identificáveis ou sensíveis. Com isso, tanto organizações públicas e privadas quanto o âmbito acadêmico, científico ou de pesquisa devem estar, conforme elucidado por Azeroual, Schöpfel, Ivanovic e Nikiforova (2022), estrategicamente preocupados com os temas: governança de dados e *Data Quality*[4].

Dados identificados, como CPF e nome, ou identificáveis, como endereço e cargo, exigem cuidado especial, assim como informações sensíveis, como saúde, origem racial e opiniões políticas. O conceito de Minimização de Dados exige que as empresas coletem apenas os dados pessoais estritamente necessários para alcançar uma finalidade específica. Em relação aos dados, no que tange o titular é o dono da informação, enquanto o controlador define como os dados serão tratados, o operador executa o processamento, o DPO supervisiona a conformidade e a ANPD fiscaliza o cumprimento da lei.

Os princípios de governança de dados como transparência, necessidade, segurança, rastreabilidade e accountability (responsabilização) orientam a coleta, armazenamento e análise de dados (Data Mapping[5] e Data Lineage[6]). Para conformidade, recomenda-se mapear o ciclo de vida dos dados, auditar processos, revisar contratos e aplicar medidas técnicas e administrativas que garantam proteção e não discriminação.

CONSIDERAÇÕES FINAIS

Os DWs, DLs e DLHs atendem a diferentes necessidades de negócios. O sucesso da implementação depende do levantamento de requisitos, maturidade da equipe tecnológica, engajamento de stakeholders e políticas de governança. A documentação adequada,

[4] Data Quality: envolve um conjunto de técnicos que corroboram para a confiabilidade e utilidade dos dados, determinando sua precisão, consistência. Na maioria dos casos, se faz necessária uma alfabetização da organização como um todo.
[5] Data Mapping: prática que orienta a coleta e o armazenamento de dados, garantindo que informações de diferentes fontes sejam corretamente associadas e estruturadas.
[6] Data Lineage: prática que acompanha a trajetória dos dados desde a origem até o destino, permitindo rastreabilidade e compreensão de como os dados foram transformados e utilizados.

padronização de processos e conformidade com a LGPD são pilares fundamentais para garantir qualidade, segurança e eficiência no uso dos dados, abrindo oportunidades para análise avançada e inteligência artificial.

REFERÊNCIAS

AFFELDT, Fabrício Sobrosa; SILVA JÚNIOR, Sady Darcy da. Information Architecture Analysis Using Business Intelligence Tools Based on the Information Needs of Executives. **Journal of Information Systems and Technology Management**, v. 10, n. 2, p. 251-270, 2013. DOI: 10.4301/S1807-17752013000200002.

AZEROUAL, O.; SCHÖPFEL, J.; IVANOVIC, D.; NIKIFOROVA, A. Combining Data Lake and Data Wrangling for Ensuring Data Quality in CRIS. **Procedia Computer Science**, v. 211, p. 3-16, 2022. DOI: 10.1016/j.procs.2022.10.171.

BENTAIB, Mohssine et al. Storage structures in the era of big data: from data warehouse to lakehouse. **Journal of Theoretical and Applied Information Technology**, v. 102, n. 6, p. 31, mar. 2024.

DATE, C. J. **Introdução a sistemas de bancos de dados**. 8. ed. Rio de Janeiro: Elsevier, 2003.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados**. 6. ed. São Paulo: Pearson Addison Wesley, 2011.

ERRAMI, Soukaina Ait et al. Spatial big data architecture: from data warehouses and data lakes to the lakehouse. **Journal of Parallel and Distributed Computing**, v. 176, p. 70-79, jun. 2023. DOI: 10.1016/j.jpdc.2023.02.007.

HARBY, A.; ZULKERNINE, F. From Data Warehouse to Lakehouse: A Comparative Review. In: **IEEE International Conference on Big Data (Big Data)**, 2022. p. 389-395. DOI: 10.1109/BigData55660.2022.10020719.

LLAVE, M. R. Data lakes in business intelligence: reporting from the trenches. **Procedia Computer Science**, v. 138, p. 516-524, 2018. DOI: 10.1016/j.procs.2018.10.071.

RAMAKRISHNAN, Raghu; GEHRKE, Johannes. **Sistemas de gerenciamento de bancos de dados**. 3. ed. São Paulo: McGraw-Hill, 2008.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. **Sistema de banco de dados**. 6. ed. Rio de Janeiro: Elsevier, 2012.