Data Lakes Orientados por Metadados: um Estudo Exploratório sobre Governança e Proveniência

Márcio Ortolan[1], **Felipe Ivo da Silva**[1],**Marilde Terezinha Prado Santos**[1]

[1]Departamento de Ciencias da Computação – Universidade Federal de São Carlos (UFSCar)
São Carlos – Brazil

marcio.ortolan@gmail.com, felipe_ivodasilva@hotmail.com, marilde.santos@ufscar.br

Abstract. *The increasing centrality of data in organizational activities has driven the adoption of flexible architectures such as data lakes, which allow for the storage of raw data for multiple analytical purposes. However, the absence of systematic metadata strategies compromises interoperability and reuse, frequently leading to the degradation of these environments into costly and inefficient "data swamps". This paper investigates the role of metadata governance not merely as informational support, but as a fundamental FinOps driver for optimizing Total Cost of Ownership (TCO). By articulating foundations of Information Science and Data Engineering, the research analyzes the impact of optimized query engines, such as Photon, and elastic allocation frameworks, like BigOPERA, on reducing operational costs. The results suggest that approaches guided by semantic metadata and provenance enhance data transparency and reliability, allowing technical optimizations to reach their full savings potential. It is concluded that the integration of rigorous governance and processing efficiency is the only path for the evolution of data lakes into financially sustainable analytical infrastructures.*

Resumo. *A crescente centralidade dos dados nas atividades organizacionais impulsionou a adoção de arquiteturas flexíveis como os data lakes, que permitem o armazenamento de dados em estado bruto para múltiplos fins analíticos. Contudo, a ausência de estratégias sistemáticas de metadados compromete a interoperabilidade e o reuso, levando frequentemente à degradação desses ambientes em data swamps onerosos e ineficientes. Este artigo investiga o papel da governança de metadados não apenas como suporte informacional, mas como driver fundamental de FinOps para a otimização do Custo Total de Propriedade (TCO). Articulando fundamentos da Ciência da Informação e da Engenharia de Dados, a pesquisa analisa o impacto de motores de consulta otimizados, como o Photon, e frameworks de alocação elástica, como o BigOPERA, na redução de custos operacionais. Os resultados sugerem que abordagens orientadas por metadados semânticos e proveniência ampliam a transparência e a confiabilidade dos dados, permitindo que otimizações técnicas atinjam seu potencial máximo de economia. Conclui-se que a integração entre governança rigorosa e eficiência de processamento é o único caminho para a evolução de data lakes em infraestruturas analíticas financeiramente sustentáveis.*

1. Introdução

A sociedade contemporânea encontra-se profundamente marcada pela intensificação da produção, circulação e consumo de dados digitais. Processos de digitalização institucional, expansão da ciência orientada a dados e popularização de tecnologias como sensores, redes sociais e sistemas distribuídos resultaram em ecossistemas informacionais caracterizados pela heterogeneidade e pelo crescimento exponencial dos dados. Nesse cenário, infraestruturas tradicionais de armazenamento e análise, como os data warehouses [Yuan et al. 2021], passaram a apresentar limitações significativas, sobretudo no que se refere à flexibilidade, ao custo e à capacidade de lidar com dados não estruturados.

Como resposta a essas limitações, os data lakes emergiram como uma arquitetura capaz de armazenar dados em seu formato original, permitindo posterior exploração analítica conforme diferentes necessidades. Segundo [Yuan et al. 2021], essa abordagem favorece a experimentação analítica e o reuso de dados, especialmente em contextos de ciência de dados e aprendizado de máquina. No entanto, a literatura também aponta que a simples adoção de data lakes não garante benefícios automáticos, uma vez que a ausência de mecanismos de organização e descrição pode comprometer sua utilidade ao longo do tempo.

A problemática central reside no fato de que muitos data lakes são concebidos com ênfase exclusiva na camada de armazenamento, negligenciando aspectos informacionais fundamentais, como descrição, contexto, proveniência e qualidade dos dados. Estudos recentes indicam que iniciativas de catalogação, adoção de padrões de metadados e incorporação de camadas semânticas têm sido propostas para enfrentar essas limitações, embora ainda de forma fragmentada e sem consenso consolidado [Silva et al. 2024] e [Rolim et al. 2024].

Complementarmente aos desafios de organização informacional, emerge a necessidade de garantir a sustentabilidade econômica destas infraestruturas através de estratégias de FinOps, visto que o superprovisionamento e a ineficiência de processamento elevam drasticamente o TCO. A literatura técnica recente demonstra que a eficiência de um data lake não depende apenas da catalogação, mas da otimização da camada de execução e da alocação elástica de recursos. Frameworks como o BigOPERA, proposto por [Caderno et al. 2025], introduzem a gestão dinâmica de ativos computacionais para reduzir o desperdício em nuvem, alcançando melhorias expressívas desempenho através do agendamento oportunista.

Simultaneamente, o desenvolvimento de motores de consulta de alto desempenho, como o Photon, detalhado por [Behm et al. 2022], e de paradigmas que elevam estruturas de acesso ao estatuto de elementos centrais, como o LakeHarbor de [Yamada et al. 2024], revelam que a integração entre metadados estruturais e execução vetorizada é capaz de reduzir o tempo de processamento e a movimentação desnecessária de dados. Tais avanços corroboram as discussões de [Madera 2018] sobre a 'gravidade dos dados', evidenciando que a governança de metadados deve ser o pilar central para a mitigação de custos operacionais em arquiteturas massivas.

Diante desse contexto, o objetivo geral deste trabalho é analisar o papel dos metadados na estruturação, governança e interoperabilidade de data lakes, a partir de uma síntese teórica e conceitual da literatura. Para atingir esse objetivo, adota-se um método

exploratório e bibliográfico, com análise qualitativa de estudos recentes. Os resultados sugerem que a integração sistemática de metadados, especialmente aqueles relacionados à proveniência e à semântica, constitui um elemento central para a evolução dos data lakes em infraestruturas confiáveis e reutilizáveis.

Para a realização deste estudo de caráter exploratório, adotou-se o método de revisão narrativa da literatura. O levantamento bibliográfico foi conduzido em bases de dados científicas amplamente reconhecidas, incluindo Scopus, IEEE Xplore e Google Scholar, por meio da combinação de descritores como Data Lake, Metadata Governance, Provenance e FinOps. Foram priorizadas publicações disponibilizadas entre os anos de 2020 e 2025, com ênfase em estudos que abordassem, de forma integrada, os desafios relacionados à organização e governança da informação em ambientes de dados, bem como aspectos associados à eficiência computacional.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 contextualiza os metadados de proveniência em data lakes; a Seção 4 discute a relação entre proveniência e governança; a Seção 5 apresenta os resultados obtidos; a Seção 6 traz a discussão dos achados; e a Seção 7 apresenta as conclusões e perspectivas de trabalhos futuros.

2. Trabalhos Relacionados

A literatura sobre data lakes e metadados abrange contribuições oriundas de diferentes áreas, como Engenharia de Dados, Ciência da Informação e Sistemas de Informação. Trabalhos clássicos sobre data lakes enfatizam sua flexibilidade arquitetural e o paradigma de schema-on-read, no qual a estrutura dos dados é definida no momento da análise e não da ingestão [Yuan et al. 2021]. Essa característica é apontada como um diferencial em relação aos data warehouses, especialmente em cenários de dados heterogêneos.

Entretanto, diversos estudos ressaltam que a ausência de metadados adequados compromete a compreensão e o reuso dos dados armazenados. No contexto governamental, [Silva et al. 2024] propõem uma abordagem baseada no Dublin Core para a catalogação de fontes de dados, destacando a importância de padrões consolidados para garantir interoperabilidade e transparência. De forma complementar, [Rolim et al. 2024] exploram a construção de grafos de metadados como meio de representar relações semânticas e processos organizacionais, ampliando a capacidade de integração e descoberta de dados.

No domínio científico, pesquisas aplicadas à área da saúde demonstram que data lakes multicêntricos demandam modelos robustos de metadados para assegurar interoperabilidade e reprodutibilidade [Lima et al. 2021]. Já estudos voltados a dados geoespaciais indicam que arquiteturas híbridas, como os data lakehouses, combinam características de data lakes e data warehouses, incorporando camadas transacionais e de metadados mais estruturadas [Vasconcelos and Coutinho 2024].

Para além da organização lógica e semântica, uma vertente recente da literatura associa a gestão de metadados à eficiência de infraestrutura e à sustentabilidade econômica através do conceito de FinOps. [Caderno et al. 2025] introduzem o framework BigO-PERA, demonstrando que a alocação elástica e oportunista de recursos, baseada em contêineres, pode gerar melhorias expressivas no desempenho computacional ao aproveitar ativos ociosos, mitigando custos de superprovisionamento. Tais avanços corroboram

a tese de [Madera 2018], que argumenta que a ”gravidade dos dados”— definida pelo volume, custo e sensibilidade — exige que a governança atue como o regulador central entre o processamento e a movimentação de dados.

No que tange à desempenho de execução, [Behm et al. 2022] apresentam o motor Photon, que utiliza vetorização em C++ nativo para superar os gargalos das máquinas virtuais tradicionais, reduzindo o tempo de consulta e o custo de faturamento em nuvem. Essa otimização é potencializada pelo paradigma LakeHarbor, proposto por [Yamada et al. 2024], que eleva estruturas de acesso ao status de ”cidadãos de primeira classe”. Ao integrar metadados estruturais à execução de baixo nível, esse modelo permite que o sistema evite varreduras de dados (full scans) onerosas, tornando a infraestrutura do data lake economicamente sustentável.

Adicionalmente, trabalhos na área de aprendizado de máquina destacam a relevância dos metadados experimentais para a explicabilidade e a tomada de decisão humana, enfatizando a necessidade de registrar parâmetros, versões e proveniência dos dados utilizados [Silva and Mattoso 2023]. Em conjunto, esses estudos evidenciam uma convergência conceitual: os metadados não devem ser tratados como elementos acessórios, mas como componentes centrais das arquiteturas de dados em Data Lakes.

Embora a literatura apresente estudos robustos sobre governança [Silva et al. 2024] e otimização de custos [Caderno et al. 2025], a maioria dos trabalhos investiga esses temas de forma isolada. Existem mapeamentos sistemáticos na literatura voltados para a arquitetura de ingestão de data lakes, contudo, poucos se propõem a cruzar a dimensão informacional (metadados e proveniência) com a dimensão operacional (FinOps e motores vetorizados). A principal lacuna que este estudo exploratório visa preencher é justamente a síntese conceitual de como a governança de metadados atua como o habilitador primário para que otimizações técnicas de infraestrutura alcancem sua máxima eficiência econômica.

3. Contextualização de Metadados de Proveniência em Data Lakes

A crescente adoção de arquiteturas de Data Lakes como repositórios centrais para armazenamento de dados brutos e heterogêneos tem evidenciado a necessidade premente de mecanismos robustos de governança, especialmente aqueles relacionados à rastreabilidade e à transparência dos dados ao longo de seu ciclo de vida [Nargesian et al. 2019]. Conforme discutido por [Giebler et al. 2021], a ausência de estratégias sistemáticas de metadados compromete a interoperabilidade e o reuso, frequentemente levando à degradação desses ambientes onerosos e ineficientes. Nesse contexto, os metadados de proveniência emergem como um elemento estruturante fundamental, indo além de uma função meramente descritiva para assumir um papel central na governança e na sustentabilidade econômica dessas infraestruturas [Madera 2018].

Os dados de proveniência, conforme definido por [da Silva 2025], consistem no registro detalhado sobre a origem dos dados e as transformações pelas quais passaram ao longo de sua existência. Em ambientes de Data Lakes, caracterizados pela diversidade de formatos e pela ausência de esquemas rígidos, a proveniência torna-se essencial para garantir a transparência, a confiabilidade e a segurança das informações [Simmhan et al. 2005]. Esse registro histórico permite que usuários compreendam não apenas a origem dos dados, mas também o conjunto de operações aplicadas, os agentes

responsáveis por tais transformações e as dependências entre diferentes artefatos de dados [Zhao et al. 2009].

A padronização na representação de dados de proveniência é um aspecto crítico para viabilizar a interoperabilidade entre sistemas heterogêneos, especialmente em ambientes distribuídos como os Data Lakes [Magagna et al. 2020]. Nesse sentido, o padrão PROV, estabelecido pelo World Wide Web Consortium (W3C), constitui a principal referência internacional para a representação de proveniência em múltiplos contextos [Gil and Miles 2013]. O modelo de dados PROV (PROV-DM) estabelece três elementos fundamentais para a descrição da proveniência: (i) entidades, que correspondem a objetos ou dados com estado persistente; (ii) atividades, representando os processos ou ações que geram, utilizam ou modificam entidades; e (iii) agentes, responsáveis por iniciar, controlar ou supervisionar tais atividades [Moreau and Missier 2013].

As relações entre esses elementos, tais como wasGeneratedBy, used, wasDerivedFrom, wasAssociatedWith e wasInformedBy, compõem o arcabouço semântico que permite a construção de grafos de proveniência completos e navegáveis [Almeida et al. 2025]. No contexto de Data Lakes, a utilização de grafos de metadados tem se mostrado particularmente promissora, pois permite representar relações complexas entre dados, processos e agentes, ampliando a capacidade de análise e rastreabilidade [Rolim et al. 2024].

4. Proveniência e Governança em Data Lakes

A governança de dados em ambientes de Data Lakes transcende a simples catalogação de ativos, exigindo a integração de metadados descritivos, estruturais e de proveniência em um modelo coeso de gestão informacional [Lis and Otto 2020]. Conforme argumentado por [Almeida et al. 2025], a rastreabilidade dos dados ao longo de todo o seu ciclo de vida é fundamental para a garantia de requisitos regulatórios e de negócios, especialmente em domínios críticos e sensíveis. Essa visão alinha-se aos princípios da data governance moderna, que enfatiza a necessidade de mecanismos que permitam não apenas a compreensão do estado atual dos dados, mas também a reconstituição histórica de sua evolução.

A estruturação de dados de proveniência em bundles, conforme previsto no padrão PROV, oferece um mecanismo natural para o encapsulamento de transações e a preservação do contexto semântico de cada operação [Moreau and Lebo 2013]. Como demonstrado por [Almeida et al. 2025], os bundles permitem agrupar entidades, atividades e agentes relacionados a uma determinada transação, possibilitando a composição histórica das interações entre diferentes componentes de um ecossistema de dados. O relacionamento mentionOf, especificamente, viabiliza a continuidade semântica entre transações distintas, estabelecendo ligações entre entidades pertencentes a diferentes bundles e permitindo o encadeamento temporal da proveniência.

No contexto de arquiteturas orientadas a Data Lakes, como a proposta por [Bezerra et al. 2025], a proveniência desempenha garantia de reprodutibilidade e auditabilidade dos processos de transformação. A segmentação em zonas (raw, processed e safe) requer mecanismos de rastreabilidade que permitam reconstituir o caminho percorrido pelos dados desde sua ingestão até sua disponibilização para consumo [Rodrigues and Mello 2022]. O versionamento sistemático dos dados em cada zona, ali-

ado ao registro de metadados de proveniência, assegura que cada transformação seja documentada e passível de verificação, promovendo a transparência e a confiabilidade necessárias para a tomada de decisão baseada em dados.

A relação entre governança de metadados e eficiência operacional em Data Lakes tem sido objeto de crescente atenção na literatura recente. Conforme argumentado por [Madera 2018] em sua Teoria da Gravidade dos Dados, a localização e o fluxo dos dados dentro de uma infraestrutura analítica têm impactos diretos no TCO. Ao avaliar os vetores de volume (massa), custo e sensibilidade, a teoria orienta decisões críticas sobre replicação física versus processamento federado, com implicações significativas para a sustentabilidade econômica do Data Lake. A integração entre metadados de proveniência e mecanismos de otimização de recursos computacionais potencializa os benefícios de eficiência obtidos por tecnologias como motores de consulta vetorizados e frameworks de alocação elástica. O motor Photon, por exemplo, explora estruturas de metadados para evitar varreduras de dados desnecessárias (full scans), alcançando acelerações significativas no processamento de consultas [Behm et al. 2022]. Da mesma forma, o framework BigOPERA demonstra que a alocação oportunista de recursos computacionais, quando combinada com metadados de execução, pode gerar melhorias expressivas no desempenho e redução do consumo energético [Caderno et al. 2025].

A complexidade inerente aos ambientes de Data Lakes torna evidente que os dados não permanecem estáticos, mas percorrem múltiplos fluxos de processamento ao longo de seu ciclo de vida. De acordo com [Suriarachchi and Plale 2016], esses fluxos são frequentemente executados por diferentes frameworks e paradigmas computacionais, selecionados conforme a natureza das cargas de trabalho, como processamento em lote, streaming e execução de scripts legados. Nesse cenário heterogêneo, a compreensão das transformações realizadas sobre os dados exige uma visão integrada das interações entre esses diversos componentes. A Figura a seguir ilustra essa dinâmica, evidenciando como múltiplos sistemas coexistem em um ecossistema analítico e como suas interações demandam mecanismos robustos de rastreabilidade baseados em metadados de proveniência.

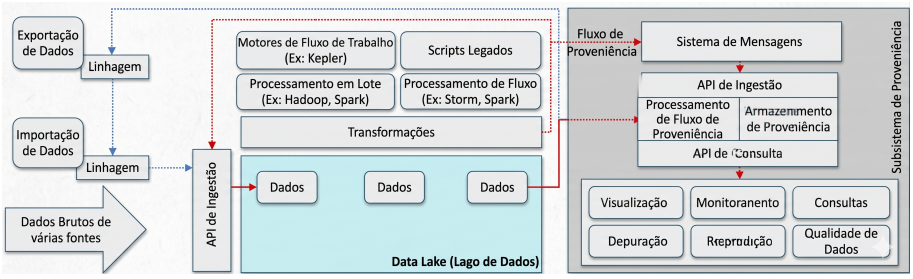


Figura 1. Arquitetura de proveniência em Data Lakes [Suriarachchi and Plale 2016]

Conforme representado, a diversidade de ferramentas e frameworks envolvidos no processamento de dados introduz desafios significativos para a captura e integração da proveniência. Embora existam abordagens específicas para coletar metadados de execução em plataformas como sistemas de processamento distribuído e motores de streaming, essas soluções tendem a ser fortemente acopladas a tecnologias individuais, dificultando a construção de uma visão unificada. A integração desses rastros em um modelo comum, embora necessária para garantir rastreabilidade completa, pode implicar perdas

semânticas e custos computacionais elevados, especialmente diante da escala dos grafos de proveniência gerados em ambientes de Big Data.

A consolidação dessas abordagens aponta para um modelo de governança ativa, no qual os metadados de proveniência não apenas documentam o histórico dos dados, mas também informam decisões operacionais em tempo real. Conforme proposto por [Jordão et al. 2025] no contexto de anonimização de dados, a integração entre metadados de privacidade e fluxos de processamento permite a aplicação adaptativa de técnicas de proteção, otimizando o trade-off entre utilidade e conformidade regulatória. Essa visão integrada entre governança informacional e eficiência computacional constitui a base para a evolução dos Data Lakes em infraestruturas analíticas financeiramente sustentáveis.

5. Resultados

Tabela 1. Sumarização das principais abordagens analisadas.

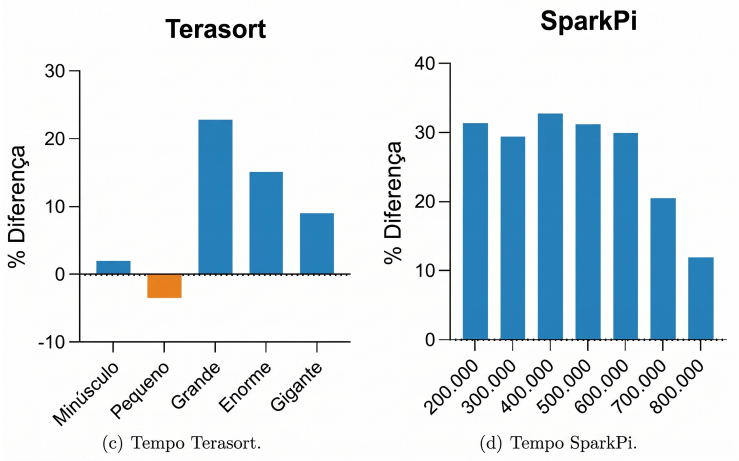
Trabalho/Tecnologia	**Foco Principal**	**Impacto Governança e TCO**
[Silva et al. 2024]; [Rolim et al. 2024]	Padronização semântica e grafos de metadados.	Evita *lago de dados ineficiente*, melhora a interoperabilidade e rastreabilidade.
Photon [Behm et al. 2022]	Motor de consulta nativo e vetorizado.	Reduz tempo de CPU e faturamento em nuvem, evitando varreduras desnecessárias.
BigOPERA [Caderno et al. 2025]	Alocação elástica e oportunista de recursos.	Reduz o TCO e o consumo de energia utilizando nós ociosos.
PROV-W3C	Padronização de dados de proveniência em grafos.	Garante auditoria, segurança e reconstrução histórica dos dados.

A análise dos estudos selecionados evidência que data lakes concebidos sem uma estratégia explícita de metadados tendem a apresentar problemas recorrentes de descoberta, interpretação e reuso dos dados. Esses ambientes, frequentemente denominados data swamps, caracterizam-se pela perda gradual de valor informacional, à medida que o volume e a complexidade dos dados aumentam [Yuan et al. 2021].

Por outro lado, abordagens que incorporam catálogos de dados, padrões de metadados e modelos semânticos demonstram avanços significativos em termos de governança e interoperabilidade. A utilização de padrões como o Dublin Core, conforme discutido por [Silva et al. 2024], contribui para a padronização descritiva e para a integração entre diferentes domínios. Além disso, grafos de metadados permitem representar relações complexas entre dados, processos e agentes, ampliando a capacidade de análise e rastreabilidade [Rolim et al. 2024].

Os resultados também indicam que arquiteturas data lakehouse introduzem mecanismos adicionais de controle, como versionamento e propriedades transacionais, baseados em uma camada formal de metadados. Essa característica aproxima tais arquiteturas de requisitos historicamente associados a data warehouses, sem eliminar a flexibilidade dos data lakes [Vasconcelos and Coutinho 2024], e demonstram que a otimização

econômica de ecossistemas analíticos é alcançada através da sinergia entre o framework BigOPERA, o motor de consulta Photon e os preceitos da Teoria da Gravidade dos Dados.



(c) Tempo Terasort. (d) Tempo SparkPi.

Figura 2. Benchmark SparkPi com 800.000 partições. [Caderno et al. 2025]

Conforme demonstrado por [Caderno et al. 2025] e ilustrado, o framework Big OPERA valida que a alocação elástica de recursos gera uma melhoria de até 35% no desempenho. O teste de tolerância a falhas utilizando o benchmark SparkPi evidenciou que, mesmo com a interrupção de nós oportunistas, o cluster híbrido manteve-se consistentemente mais rápido que o dedicado, provando sua robustez e capacidade de reduzir o TCO e mitigar o superprovisionamento.

Figura 3. Desempenho do sistema BigOPERA Spark em um cenário de tolerância a falhas. [Caderno et al. 2025]

O teste utilizou o benchmark SparkPi com 800.000 partições. O objetivo foi verificar como o BigOPERA se comporta quando os nós ”oportunistas”(não dedicados) são interrompidos aleatoriamente, simulando o que acontece na vida real quando o dono de um computador retoma o uso da máquina.

A barras azuis representam o tempo de execução no cluster híbrido do BigOPERA, onde os nós foram desligados e reiniciados aleatoriamente durante o processo em contrapartida as barras laranjas representa o tempo de execução em um cluster padrão, composto apenas por nós dedicados (estáveis).

Mesmo com as interrupções forçadas e a variabilidade nos tempos de execução, o cluster híbrido (barras azuis) foi consistentemente mais rápido que o cluster dedicado (barra laranja). O resultado demonstra que o ganho de desempenho obtido ao usar recursos extras (mesmo que instáveis) supera o tempo perdido com as falhas e reinicializações de tarefas. O gráfico prova que o framework é capaz de manter a funcionalidade e recuperar o trabalho sem falhar a aplicação final, validando sua robustez em ambientes dinâmicos.

No que tange à desempenho de execução, [Behm et al. 2022] detalham como o motor Photon alcança acelerações médias de 3x, atingindo picos de até 23x em consultas complexas e avaliações de aritmética decimal. Ao utilizar uma arquitetura em C++ nativo e execução vetorizada, o motor supera as limitações de desempenho e de gerenciamento de memória das máquinas virtuais Java (JVM) tradicionais. Como o faturamento em nuvem pública é frequentemente atrelado ao tempo de ocupação da CPU, a eficiência do Photon traduz-se em uma redução direta nos custos operacionais ao processar exabytes de dados diários de forma mais célere e eficiente.

6. Discussão

Os achados reforçam a compreensão de que os metadados desempenham um papel estruturante nos data lakes, indo além de uma função meramente descritiva. Metadados de proveniência, por exemplo, são essenciais para a transparência dos processos analíticos e para a confiabilidade dos resultados, especialmente em contextos científicos e de aprendizado de máquina [Silva and Mattoso 2023]. Essa relevância torna-se ainda mais evidente quando se observa o fluxo operacional de um data lake moderno, no qual dados são continuamente ingeridos, transformados e reprocessados por diferentes frameworks.

Conforme ilustrado na Figura 4, pipelines compostos por múltiplos estágios, desde a ingestão até a agregação analítica, geram sucessivas camadas de dados (raw, intermediárias e consolidadas), cuja interpretação depende diretamente da capacidade de rastrear suas transformações. Nesse cenário, a proveniência atua como um elo integrador entre essas etapas, registrando relações entre entidades, atividades e agentes, e permitindo a reconstrução completa do ciclo de vida dos dados.

Além disso, a captura sistemática de eventos de proveniência ao longo desses pipelines possibilita a construção de grafos ricos e navegáveis, nos quais cada transformação deixa um rastro explícito. Essa abordagem não apenas favorece auditorias e reprodutibilidade, mas também viabiliza análises mais sofisticadas sobre o próprio comportamento do sistema, como identificação de gargalos, redundâncias e padrões de uso. Em ambientes heterogêneos, nos quais ferramentas distintas coexistem, como mecanismos de ingestão, processamento distribuído e motores analíticos, a padronização da proveniência, especialmente com base em modelos como o PROV, torna-se fundamental para garantir interoperabilidade semântica e integração entre diferentes camadas tecnológicas.

A sustentabilidade financeira deste modelo integrado é reforçada pela aplicação da Teoria da Gravidade dos Dados, proposta por [Madera 2018], que atua como o alicerce estratégico para a tomada de decisão sobre o processamento e a movimentação da informação. Ao avaliar os vetores de volume (massa), custo e sensibilidade, a teoria orienta se o dado deve ser duplicado fisicamente no lago ou se o processamento deve ser federado na fonte, evitando a degradação do sistema em um "pântano de dados" (data

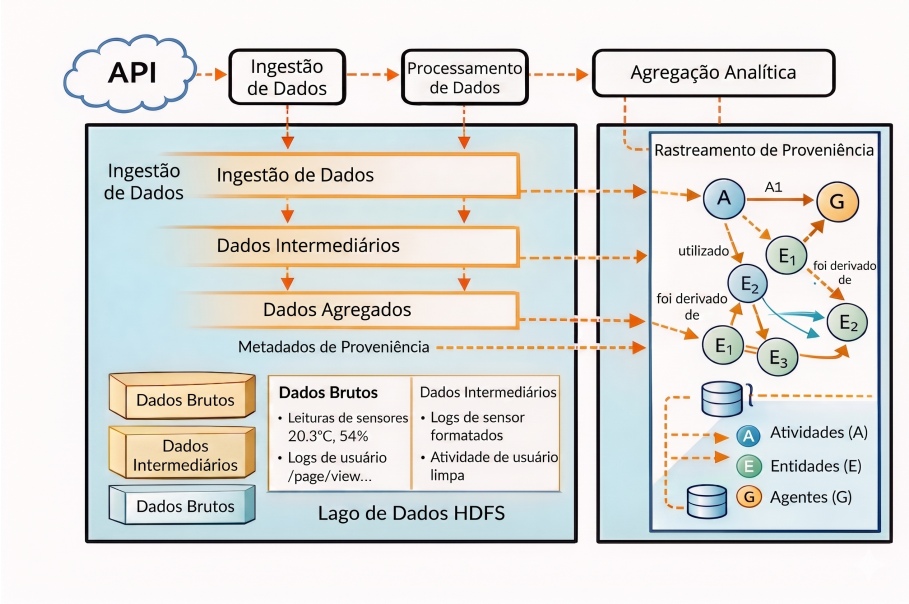


Figura 4. Modelo conceitual de pipeline analítico em Data Lake com integração de metadados de proveniência. [Fonte: elaborado pelos autores]

swamp). Quando articulada com os fluxos de processamento observados na arquitetura ilustrada, essa teoria evidencia que decisões sobre onde e como processar dados não podem ser dissociadas dos metadados disponíveis. Informações de proveniência e de uso permitem identificar quais dados são mais acessados, quais transformações são mais custosas e quais caminhos analíticos podem ser otimizados.

Estudos de caso reais discutidos por [Madera 2018] comprovam que a replicação indiscriminada pode ser financeiramente proibitiva, com custos para mover dados de um mainframe estimados em aproximadamente 22 milhões de dólares anuais. Nesse sentido, a integração entre metadados e mecanismos de execução avançados torna-se um diferencial competitivo. A execução vetorizada de alto desempenho, exemplificada pelo Photon, aliada à elasticidade computacional promovida por abordagens como o BigOPERA, permite que o sistema explore dinamicamente os recursos disponíveis, reduzindo tempos de processamento e custos operacionais. Essa sinergia é potencializada quando guiada por metadados de execução e proveniência, que fornecem subsídios para decisões automatizadas de escalonamento, alocação de recursos e otimização de consultas.

Entretanto, desafios persistem, como a padronização entre diferentes domínios, o custo de implementação e a necessidade de alinhamento organizacional. A instrumentação de pipelines para captura de proveniência, por exemplo, pode introduzir sobrecarga técnica e exigir adaptações nos sistemas existentes, especialmente em ambientes legados. Adicionalmente, a integração de metadados provenientes de múltiplas ferramentas pode gerar inconsistências semânticas, exigindo esforços adicionais de harmonização e modelagem conceitual.

Observa-se ainda que a dependência de soluções proprietárias pode comprometer a interoperabilidade a longo prazo, reforçando a importância de padrões abertos e modelos conceituais bem definidos. Nesse contexto, a adoção de abordagens desacopladas, baseadas em mensageria e coleta assíncrona de eventos de proveniência, surge como

uma alternativa promissora para mitigar tais limitações, permitindo a integração progressiva de diferentes componentes sem comprometer a escalabilidade do sistema. Dessa forma, a consolidação de uma governança orientada por metadados, aliada a estratégias de otimização computacional e sustentabilidade financeira, configura-se como um caminho essencial para a evolução dos data lakes em infraestruturas analíticas robustas, interoperáveis e economicamente viáveis.

7. Conclusão

Frente ao objetivo estabelecido de analisar o papel dos metadados na estruturação, governança e interoperabilidade de data lakes, os resultados obtidos neste trabalho demonstram que a adoção de abordagens orientadas por metadados — com ênfase especial em proveniência e governança — é condição fundamental para a consolidação desses ambientes como infraestruturas analíticas sustentáveis, governáveis e interoperáveis. A investigação realizada demonstra que a integração entre governança rigorosa e eficiência de processamento, materializada por meio de motores otimizados como o Photon e frameworks de alocação elástica como o BigOPERA, permite transformar data lakes em ecossistemas financeiramente viáveis, alinhados aos princípios de FinOps e à otimização do TCO. Nesse sentido, a articulação entre a camada informacional e a camada computacional revela-se não apenas benéfica, mas estruturante para a evolução dessas arquiteturas.

Cumpre reconhecer, contudo, as ameaças à validade do presente estudo, as quais residem primordialmente em sua dependência de literatura secundária e na ausência de validação empírica direta. Embora a síntese conceitual desenvolvida tenha permitido avanços significativos na compreensão do papel estratégico dos metadados, a ausência de experimentação em ambientes produtivos de larga escala impõe limites à generalização das conclusões, indicando a necessidade premente de investigações complementares que combinem análise teórica com validação prática.

No que se refere à relação com o estado da arte, este estudo consolida evidências de que modelos híbridos, como o data lakehouse, e o emprego de grafos de metadados representam tendências promissoras, alinhando-se a um movimento mais amplo de convergência entre flexibilidade arquitetural e requisitos de governança tradicionalmente associados a data warehouses. A literatura recente aponta, adicionalmente, para a necessidade de integração de padrões abertos, como o PROV-W3C, e de tecnologias de persistência orientadas a grafos, a exemplo do Neo4j, como meios para representar a proveniência de forma fiel, navegável e escalável. Tais abordagens superam limitações históricas de interoperabilidade e abrem caminho para uma governança ativa, na qual os metadados não apenas documentam o histórico dos dados, mas informam decisões operacionais em tempo real.

Do ponto de vista social e científico, os resultados obtidos podem contribuir para a construção de infraestruturas de dados mais transparentes, reutilizáveis e alinhadas aos princípios da ciência aberta, especialmente em domínios sensíveis como saúde pública e pesquisas multicêntricas. A utilização sistemática de metadados de proveniência fortalece a confiabilidade de processos analíticos, auxilia na conformidade com legislações como a Lei Geral de Proteção de Dados (LGPD) e fomenta práticas de computação sustentável ao reduzir desperdícios computacionais e energéticos. Ademais, a abordagem aqui delineada oferece subsídios para organizações que buscam conciliar inovação analítica com

responsabilidade fiscal e ambiental.

Apesar dos avanços consolidados na literatura, a implementação efetiva de estratégias de proveniência em data lakes ainda enfrenta desafios significativos. A diversidade de formatos de dados, a heterogeneidade dos sistemas envolvidos e a complexidade dos fluxos de processamento impõem obstáculos técnicos à captura sistemática e à representação unificada da proveniência. Questões relacionadas à escalabilidade do armazenamento de metadados, à granularidade da captura e ao overhead de desempenho introduzido pelos mecanismos de registro demandam soluções arquiteturais cuidadosamente projetadas, que conciliem riqueza semântica e eficiência operacional.

Nesse contexto, como desdobramento natural do presente estudo, vislumbra-se a necessidade de realização de estudos de caso empíricos em ambientes produtivos de larga escala, que permitam validar os benefícios quantitativos das estratégias de proveniência sobre a sustentabilidade econômica dos data lakes. Igualmente relevante é o desenvolvimento de modelos conceituais que integrem padrões de metadados, proveniência e governança, ampliando a interoperabilidade entre domínios e reduzindo a fragmentação observada nas práticas atuais. Adicionalmente, a exploração de técnicas de aprendizado de máquina para automatizar a análise de padrões de proveniência, detectar anomalias e otimizar, em tempo real, a alocação de recursos computacionais com base em metadados de similaridade constitui frente promissora de investigação. .

Em síntese, os achados deste estudo exploratório levantam a forte hipótese de que a coordenação entre elasticidade de infraestrutura, eficiência de software e governança fundamentada em metadados e gravidade constitui um modelo de FinOps ativo, capaz de transformar o data lake de um repositório passivo de custos em um ecossistema analítico economicamente viável, escalável e alinhado aos mais elevados padrões de transparência e confiabilidade.

Referências

Almeida, J. M., Braganholo, V., and Oliveira, D. (2025). Governança de dados em sistemas-de-sistemas: Uma abordagem orientada à dados de proveniência. In *Anais do 40º Simpósio Brasileiro de Bancos de Dados*, Fortaleza, CE, Brazil.

Behm, A., Palkar, S., Agarwal, U., Armstrong, T., Cashman, D., Dave, A., et al. (2022). Photon: A fast query engine for lakehouse systems. In *Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22)*, pages 2326–2339, New York, NY, USA. ACM.

Bezerra, A. L. d. S., Batalha, I. S., Filho, L. R. A., Almeida, C. M., Dantas, M. I. N., and Gouvêa, N. A. (2025). Rpas e data lakes para a indústria 4.0: Um estudo de caso de ecossistema de dados integrados. In *Anais do 40º Simpósio Brasileiro de Bancos de Dados*, Fortaleza, CE, Brazil.

Caderno, P. V., Awaysheh, F., Cabaleiro, J. C., and Pena, T. F. (2025). Bigopera: An opportunistic and elastic resource allocation for big data frameworks. *Cluster Computing*, 28(383):1–16.

da Silva, F. I. (2025). Proveniência de dados e metadados em repositórios de dados de pesquisa. Dissertação (mestrado), Universidade Federal de São Carlos, São Carlos, SP, Brasil.

Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., and Mitschang, B. (2021). The data lake architecture framework: A foundation for building a comprehensive data lake architecture.

Gil, Y. and Miles, S. (2013). Prov model primer. `https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/`. Acesso em: 25 mar. 2026.

Jordão, T., Bedo, M., and Oliveira, D. (2025). Arani: Uma abordagem baseada em linha de experimento para preservação de privacidade em data lakes. In *Anais do 40º Simpósio Brasileiro de Bancos de Dados*, Fortaleza, CE, Brazil.

Lima, D. M. et al. (2021). Uma proposta de data lake para pesquisa em saúde a partir de data pools multicêntricos interoperáveis. In *Anais do Simpósio Brasileiro de Bancos de Dados*, Brasil.

Lis, D. and Otto, B. (2020). Data governance in data ecosystems- insights from organizations. In *Americas Conference on Information Systems (AMCIS)*.

Madera, C. (2018). *L'évolution des systèmes et architectures d'information sous l'influence des données massives: les lacs de données*. Tese (doutorado em informática), LIRMM, Université de Montpellier, Montpellier, France.

Magagna, B., Goldfarb, D., Martin, P., Atkinson, M., Koulouzis, S., and Zhao, Z. (2020). Data provenance. In *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences: A Reference Model Guided Approach for Common Challenges*, pages 208–225. Springer.

Moreau, L. and Lebo, T. (2013). Prov-links. `https://www.w3.org/TR/2013/NOTE-prov-links-20130430/`. Acesso em: 25 mar. 2026.

Moreau, L. and Missier, P. (2013). Prov-dm: The prov data model.

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., and Arocena, P. C. (2019). Data lake management: Challenges and opportunities. *Proceedings of the VLDB Endowment*, 12(12):1986–1989.

Rodrigues, J. and Mello, R. (2022). Um estudo sobre arquiteturas e metadados em data lakes. In *Anais da XVII Escola Regional de Banco de Dados*, pages 131–134, Porto Alegre, RS, Brazil. SBC.

Rolim, T. V. et al. (2024). An approach to construct a metadata graph for specifying enterprise knowledge graphs. In *Anais do Simpósio Brasileiro de Bancos de Dados*, Brasil.

Silva, F. and Mattoso, M. (2023). Integrador de metadados modular para aprendizado de máquina com visualização em tempo de execução. In *Anais do Simpósio Brasileiro de Bancos de Dados*, Brasil.

Silva, H. A. B. et al. (2024). Uma proposta baseada no dublin core para catalogação de metadados de fontes de dados governamentais. In *Anais do Simpósio Brasileiro de Bancos de Dados*, Brasil.

Simmhan, Y. L., Plale, B., and Gannon, D. (2005). A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36.

Suriarachchi, I. and Plale, B. (2016). Crossing analytics systems: A case for integrated provenance in data lakes. In *2016 IEEE 12th International Conference on e-Science (e-Science)*, pages 349–354. IEEE.

Vasconcelos, F. F. and Coutinho, F. J. (2024). Data lakehouses para a análise de dados geoespaciais em larga escala. In *Anais do Simpósio Brasileiro de Bancos de Dados*, Brasil.

Yamada, H., Kitsuregawa, M., and Goda, K. (2024). Lakeharbor: Making structures first-class citizens in data lakes. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 5583–5592. IEEE.

Yuan, J. et al. (2021). Data lakehouse: A new generation of data architecture. *Proceedings of the VLDB Endowment*, 14(12):xxxx–xxxx.

Zhao, J., Miles, A., Klyne, G., and Shotton, D. (2009). Linked data and provenance in biological data webs. *Briefings in bioinformatics*, 10(2):139–152.