LUMEN ET VIRTUS
ISSN: 2177-2789

ARQUITETURAS DE DADOS PARA *ANALYTICS*: PADRÕES DE INTEGRAÇÃO, ETL/ELT E SUSTENTAÇÃO DE PIPELINES

DATA ARCHITECTURES FOR ANALYTICS: INTEGRATION PATTERNS, ETL/ELT, AND PIPELINE SUPPORT

ARQUITECTURAS DE DATOS PARA *ANALYTICS*: PATRONES DE INTEGRACIÓN, ETL/ELT Y SOPORTE PARA PIPELINES

https://doi.org/10.56238/levv13n31-081

Data de submissão: 10/10/2023 **Data de publicação:** 10/11/2023

Eduardo dos Santos Souza

RESUMO

O presente artigo analisa arquiteturas de dados voltadas para analytics, com ênfase nos padrões de integração, nas abordagens de ETL e ELT e nos mecanismos de sustentação de pipelines analíticos, considerando a crescente necessidade organizacional de estruturar fluxos informacionais capazes de integrar múltiplas fontes, sustentar processamento contínuo e disponibilizar dados com maior consistência para uso analítico. A pesquisa foi desenvolvida por meio de abordagem qualitativa, com caráter descritivo e explicativo, apoiada em pesquisa bibliográfica, a partir da qual foram examinadas contribuições teóricas relacionadas a data warehouse, data lake, lakehouse, integração de dados, governança e operação de fluxos analíticos. Os resultados indicam que a evolução das arquiteturas de dados está associada à busca por maior flexibilidade estrutural, maior capacidade de integração e melhor coordenação entre armazenamento, processamento e consumo analítico, revelando que a eficiência dos ambientes de analytics depende do alinhamento entre modelo arquitetural, estratégia de processamento e mecanismos de sustentação dos pipelines. Verificou-se, ainda, que ETL e ELT correspondem a lógicas operacionais distintas, cuja adoção varia conforme volume, diversidade, frequência de atualização e finalidade analítica dos dados, enquanto a continuidade dos pipelines exige monitoramento, rastreabilidade, metadados e governança articulada. Conclui-se que arquiteturas de dados voltadas para analytics precisam ser compreendidas como ecossistemas integrados, nos quais estrutura, fluxo e controle operacional atuam de forma interdependente na produção de ambientes analíticos confiáveis e escaláveis.

Palavras-chave: Arquiteturas de Dados. Integração de Dados. ETL. ELT. Pipelines Analíticos.

ABSTRACT

This article analyzes data architectures oriented toward analytics, with emphasis on integration patterns, ETL and ELT approaches, and mechanisms for sustaining analytical pipelines, considering the growing organizational need to structure information flows capable of integrating multiple sources, sustaining continuous processing, and making data available with greater consistency for analytical use. The research was developed through a qualitative approach, with descriptive and explanatory character, supported by bibliographic research, through which theoretical contributions related to data warehouse, data lake, lakehouse, data integration, governance, and the operation of analytical flows were examined. The results indicate that the evolution of data architectures is associated with the search for greater structural flexibility, greater integration capacity, and better coordination between

storage, processing, and analytical consumption, revealing that the efficiency of analytics environments depends on the alignment between architectural model, processing strategy, and mechanisms for sustaining pipelines. It was also found that ETL and ELT correspond to distinct operational logics, whose adoption varies according to volume, diversity, update frequency, and analytical purpose of the data, while pipeline continuity requires monitoring, traceability, metadata, and articulated governance. It is concluded that data architectures oriented toward analytics need to be understood as integrated ecosystems, in which structure, flow, and operational control act interdependently in the production of reliable and scalable analytical environments.

Keywords: Data Architectures. Analytics. Data Integration. ETL. ELT. Analytical Pipelines.

RESUMEN

Este artículo analiza arquitecturas de datos orientadas a la analítica, con énfasis en patrones de integración, enfoques ETL y ELT, y mecanismos para el soporte de pipelines analíticos, considerando la creciente necesidad organizacional de estructurar flujos de información capaces de integrar múltiples fuentes, soportar el procesamiento continuo y poner los datos a disposición con mayor consistencia para su uso analítico. La investigación se desarrolló mediante un enfoque cualitativo, de carácter descriptivo y explicativo, apoyado en una investigación bibliográfica, a partir de la cual se examinaron contribuciones teóricas relacionadas con data warehouse, data lake, lakehouse, integración de datos, gobernanza y operación de flujos analíticos. Los resultados indican que la evolución de las arquitecturas de datos está asociada con la búsqueda de mayor flexibilidad estructural, mayor capacidad de integración y mejor coordinación entre almacenamiento, procesamiento y consumo analítico, revelando que la eficiencia de los entornos analíticos depende de la alineación entre el modelo arquitectónico, la estrategia de procesamiento y los mecanismos de soporte de pipelines. También se encontró que ETL y ELT corresponden a lógicas operacionales distintas, cuya adopción varía según el volumen, la diversidad, la frecuencia de actualización y el propósito analítico de los datos, mientras que la continuidad de los pipelines requiere monitoreo, trazabilidad, metadatos y una gobernanza articulada. Se concluye que las arquitecturas de datos orientadas al análisis deben entenderse como ecosistemas integrados, donde la estructura, el flujo y el control operativo interactúan entre sí para la creación de entornos analíticos fiables y escalables.

Palabras clave: Arquitecturas de Datos. Integración de Datos. ETL. ELT. Pipelines Analíticos.

1 INTRODUÇÃO

A crescente circulação de dados em ambientes corporativos vem alterando a maneira como as organizações estruturam informação para análise, decisão e geração de valor, sobretudo em contextos nos quais múltiplas fontes precisam ser consolidadas com consistência, rastreabilidade e capacidade de processamento, nesse cenário, as arquiteturas de dados voltadas para *analytics* passaram a ocupar posição central nas estratégias de modernização informacional (Nambiar *et al*., 2022).

Ao acompanhar essa transformação, percebe-se que a evolução das arquiteturas analíticas deixou de se restringir ao modelo clássico de armazenamento centralizado, passando a incorporar estruturas mais flexíveis, capazes de atender demandas heterogêneas, com diferentes velocidades de ingestão e variados formatos de dados, essa mudança está associada à reconfiguração dos processos de carga, tratamento e disponibilização da informação para uso analítico (Dhaouadi *et al*., 2022).

Nesse movimento, a discussão sobre integração de dados ganha maior densidade, pois a utilidade analítica de qualquer arquitetura depende da forma como registros dispersos são combinados, padronizados e disponibilizados para consumo confiável, em ambientes nos quais coexistem dados estruturados, semiestruturados e não estruturados, a integração exige mecanismos mais adaptáveis e maior coordenação entre os componentes do ecossistema informacional (Sawadogo; Darmont, 2021).

Sob essa ótica, a sustentação contínua dos fluxos depende de rotinas organizadas de extração, transformação e carregamento, ou, em formulações mais recentes, de estratégias que deslocam parte do processamento para as camadas de destino, esse deslocamento altera a lógica de construção dos ecossistemas de dados e interfere diretamente na maneira como escalabilidade, latência e manutenção são tratadas nos ambientes corporativos (Machado *et al*., 2019).

Nessa trajetória, a comparação entre ETL e ELT ultrapassa uma distinção operacional, pois envolve decisões relacionadas à arquitetura de armazenamento, ao tipo de processamento disponível e à governança aplicada em cada etapa do fluxo, assim, a escolha entre tais estratégias depende da combinação entre volume, variedade, frequência de atualização e capacidade computacional do ambiente adotado (Harby; Zulkernine, 2022).

À medida que os volumes crescem e os fluxos se tornam mais dinâmicos, a sustentação dos pipelines passa a depender de mecanismos arquiteturais que reduzam fricções entre ingestão, processamento e consumo, favorecendo maior continuidade operacional e menor vulnerabilidade a interrupções, inconsistências e retrabalhos, em arquiteturas voltadas a *analytics*, isso implica tratar o pipeline como estrutura permanente de produção de dados confiáveis (Wieder; Nolte, 2022).

Nesse encadeamento, a manutenção dos fluxos requer acompanhamento de metadados, documentação dos processos e visibilidade sobre as transformações aplicadas, elementos que ampliam a capacidade de auditoria e fortalecem a interpretação correta dos ativos informacionais utilizados nas

análises, sob essa perspectiva, a arquitetura passa a ser observada como arranjo integrado entre tecnologia, organização da informação e controle do ciclo de vida dos dados (Nadal *et al*., 2022).

De forma associada, a presença de metadados bem estruturados favorece a inteligibilidade dos ambientes analíticos e contribui para reduzir ambiguidades na circulação dos dados, sobretudo em cenários nos quais diferentes equipes interagem com múltiplos conjuntos informacionais em ritmos distintos de atualização, tal condição reforça a compreensão de que a consistência dos pipelines depende da articulação entre automação, catalogação e monitoramento contínuo das estruturas presentes na arquitetura (Sawadogo; Darmont, 2021).

Em cenários organizacionais marcados por operações analíticas de maior dinamismo, o desempenho das arquiteturas depende da capacidade de responder a fluxos intensos sem comprometer estabilidade e utilidade analítica, estudos aplicados demonstram que a adaptação arquitetural precisa considerar simultaneamente velocidade de processamento, integridade dos dados e capacidade de resposta das soluções implementadas (Fikri *et al*., 2019).

Nessa direção, a articulação entre arquitetura, integração e desempenho torna o tema altamente relevante para estudos acadêmicos e aplicações organizacionais, pois a qualidade das análises produzidas depende da coerência entre a estrutura dos ambientes de dados e os métodos empregados para mover, transformar e disponibilizar a informação ao longo do pipeline, investigar essas relações permite compreender com maior precisão como diferentes modelos influenciam a eficiência analítica e a governança dos fluxos de dados (Nambiar *et al*., 2022).

Diante desse quadro, o objetivo deste artigo consiste em analisar arquiteturas de dados voltadas para *analytics*, com foco nos padrões de integração, nas abordagens ETL e ELT e nos mecanismos de sustentação de pipelines, buscando identificar como essas dimensões se articulam na construção de ambientes analíticos mais consistentes, escaláveis e aderentes às exigências contemporâneas de processamento de dados.

Nessa perspectiva, a justificativa para a realização deste estudo reside no fato de que organizações de distintos setores dependem cada vez mais de arquiteturas capazes de integrar dados dispersos, sustentar fluxos confiáveis e responder a demandas analíticas crescentes, ao reunir contribuições recentes sobre data *warehouse*, *data lake*, *lakehouse*, integração e governança de pipelines, este trabalho busca oferecer uma leitura articulada e atualizada de um campo que vem se consolidando como eixo estruturante da inteligência orientada por dados (Harby; Zulkernine, 2022).

2 REFERENCIAL TEÓRICO

2.1 FUNDAMENTOS DAS ARQUITETURAS DE DADOS PARA *ANALYTICS*

Inicialmente, as arquiteturas de dados para *analytics* podem ser compreendidas como arranjos técnicos voltados à coleta, organização, processamento e disponibilização de dados para fins analíticos,

sua configuração depende da forma como as organizações lidam com volume, variedade, velocidade e finalidade de uso da informação, nesse enquadramento, a arquitetura deixa de representar apenas infraestrutura de armazenamento e passa a envolver uma lógica integrada de fluxo, transformação e consumo analítico (Nambiar *et al*., 2022).

Nesse horizonte, a consolidação histórica dos ambientes analíticos esteve ligada ao *data warehouse*, estrutura concebida para reunir dados oriundos de diferentes sistemas em um repositório orientado à consulta e ao apoio gerencial, com ênfase em consistência, integração e organização temática, tal base permitiu o avanço das rotinas analíticas empresariais e estabeleceu fundamentos que ainda influenciam o desenho de soluções contemporâneas (Dhaouadi *et al*., 2022).

Sob essa leitura, a expansão de novas fontes e formatos informacionais impulsionou o surgimento de arquiteturas mais flexíveis, entre elas o data lake, concebido para armazenar grandes volumes de dados em diferentes formatos com menor rigidez estrutural na entrada, essa inflexão alterou a lógica tradicional do preparo dos dados e abriu espaço para fluxos analíticos mais adaptáveis às demandas de exploração e reuso (Sawadogo; Darmont, 2021).

Nessa evolução, a literatura passou a examinar os efeitos dessa ampliação arquitetural sobre o gerenciamento do ciclo analítico, uma vez que ambientes mais abertos oferecem maior plasticidade de uso e requerem mecanismos mais refinados de organização, rastreamento e interpretação dos ativos informacionais, por essa razão, o debate sobre arquitetura passou a incluir com maior intensidade os temas de metadados, catalogação e governança operacional (Wieder; Nolte, 2022).

De forma articulada, a comparação entre data warehouse e data lake evidenciou que cada modelo responde a necessidades distintas, um tende a privilegiar estruturação prévia e padronização analítica, o outro favorece maior amplitude de ingestão e elasticidade de armazenamento, essa distinção tornou mais nítida a necessidade de selecionar arquiteturas conforme a natureza dos fluxos, das consultas e dos objetivos analíticos envolvidos (Harby; Zulkernine, 2022).

Na continuidade dessa discussão, a evolução recente das plataformas analíticas conduziu à formulação de arquiteturas híbridas, entre elas o *lakehouse*, proposta que busca aproximar a flexibilidade do *data lake* da organização analítica típica do *data warehouse*, com isso, o campo passou a operar com desenhos mais integrados, orientados à redução de fragmentações entre armazenamento, processamento e consumo de dados (Harby; Zulkernine, 2022).

Paralelamente, a efetividade dessas arquiteturas depende dos mecanismos de integração que articulam fontes internas e externas, sistemas legados, aplicações transacionais e ambientes analíticos em uma cadeia contínua de movimentação de dados, nessa dinâmica, os fluxos de integração passam a constituir a base funcional que sustenta a confiabilidade do ecossistema analítico como um todo (Machado *et al*., 2019).

Em desdobramento a isso, a discussão sobre arquitetura para *analytics* não se limita à escolha do repositório central, ela envolve o desenho dos processos que garantem coerência entre captura, transformação, disponibilização e monitoramento dos dados ao longo do pipeline, tal entendimento amplia a análise arquitetural e aproxima sua leitura de uma visão processual, permanente e articulada do ambiente informacional (Nadal *et al*., 2022).

Nessa moldura, a presença de metadados estruturados contribui para descrever origem, transformação, relacionamento e uso dos dados, favorecendo maior inteligibilidade das estruturas analíticas e melhor acompanhamento das rotinas executadas no ambiente, esse recurso fortalece a sustentação dos pipelines ao oferecer maior visibilidade sobre o comportamento da arquitetura e de seus componentes informacionais (Sawadogo; Darmont, 2021).

Em ambientes marcados por atualização frequente, ingestão contínua e exigências crescentes de desempenho, o amadurecimento das arquiteturas analíticas está diretamente associado à capacidade de responder com elasticidade e melhor coordenação entre processamento e armazenamento, em aplicações com maior dinamismo operacional, essa característica tende a influenciar a qualidade analítica e a estabilidade dos fluxos mantidos em produção (Fikri *et al*., 2019).

Diante dessa conjuntura, compreender os fundamentos das arquiteturas de dados para *analytics* requer observar o modo como diferentes modelos estruturam a relação entre dados brutos, processamento intermediário e camadas de consumo analítico, considerando simultaneamente aspectos técnicos e organizacionais envolvidos em sua sustentação, essa leitura favorece uma interpretação mais abrangente das escolhas arquiteturais e de seus efeitos sobre integração, governança e uso analítico da informação (Dhaouadi *et al*., 2022).

Por essa via, a análise desses fundamentos fornece base conceitual para examinar, nas seções seguintes, os padrões de integração e as estratégias de ETL e ELT que viabilizam o funcionamento contínuo dessas arquiteturas, já que a consistência do pipeline depende diretamente da solidez do desenho estrutural que o suporta, assim, o estudo das arquiteturas analíticas constitui etapa necessária para compreender a dinâmica dos fluxos de dados nas organizações contemporâneas (Nambiar *et al*., 2022).

2.2 PADRÕES DE INTEGRAÇÃO DE DADOS EM AMBIENTES ANALÍTICOS

Em arquiteturas voltadas para *analytics*, os padrões de integração correspondem ao conjunto de estratégias que viabilizam a circulação dos dados entre sistemas de origem, camadas intermediárias e ambientes de consumo analítico, seu funcionamento depende da compatibilidade entre fontes, formatos, regras de transformação e objetivos informacionais, a integração, nesse sentido, constitui a base que permite converter dados dispersos em ativos utilizáveis para análise, monitoramento e suporte

à decisão, essa dinâmica interfere diretamente na consistência dos resultados analíticos e na capacidade de sustentação dos fluxos em produção (Machado *et al*., 2019).

Ao se observar a diversidade dos ecossistemas corporativos, torna-se evidente que a integração precisa contemplar bancos relacionais, aplicações transacionais, arquivos, APIs e fluxos contínuos, cenário que amplia a necessidade de padrões capazes de organizar entradas heterogêneas com estabilidade e coerência, essa heterogeneidade exige arranjos preparados para tratar dados com estruturas distintas sem comprometer o uso analítico posterior, nessa condição, a integração passa a ser compreendida como elemento estrutural da arquitetura e da governança dos ambientes de dados (Sawadogo; Darmont, 2021).

Sob tal enfoque, os padrões de integração podem ser entendidos como mecanismos que definem onde os dados serão combinados, em que momento ocorrerá a transformação e de que forma a informação seguirá para as camadas analíticas, tais escolhas afetam desempenho, rastreabilidade e flexibilidade operacional, por essa razão, a integração deixa de ser etapa isolada e passa a compor a lógica permanente de organização dos fluxos analíticos, com isso, a arquitetura passa a refletir decisões sobre tempo de processamento, granularidade dos dados e distribuição das tarefas ao longo do pipeline (Dhaouadi *et al*., 2022).

Historicamente, abordagens centralizadas ganharam força durante a consolidação dos data *warehouses*, pois permitiam reunir dados de múltiplos sistemas em um repositório único, com tratamento prévio e organização orientada à consulta, esse padrão favoreceu consistência, padronização e melhor controle sobre indicadores gerenciais, especialmente em ambientes com forte dependência de relatórios consolidados, essa configuração tornou-se referência para a construção de estruturas analíticas mais organizadas e voltadas ao suporte decisório (Nambiar *et al*., 2022).

Com a ampliação do volume e da variedade dos dados, passaram a ganhar espaço padrões mais flexíveis, nos quais a integração pode ocorrer de modo progressivo, distribuído ou orientado por diferentes zonas de processamento dentro de *data lakes* e arquiteturas híbridas, tal mudança ampliou a capacidade de ingestão e favoreceu fluxos menos rígidos, compatíveis com cenários de exploração analítica mais amplos, dessa maneira, os ambientes passaram a admitir maior diversidade de formatos e maior elasticidade no tratamento das informações (Wieder; Nolte, 2022).

A partir dessa mudança, tornou-se relevante distinguir a integração realizada antes do armazenamento analítico da integração conduzida ao longo do uso dos dados, visto que cada escolha produz efeitos distintos sobre qualidade, velocidade de disponibilização e esforço de manutenção, essa diferença influencia a forma como os pipelines são desenhados e sustentados em ambientes com elevada exigência analítica, assim, a definição do padrão de integração depende da relação entre arquitetura, capacidade computacional e finalidade de uso dos dados (Harby; Zulkernine, 2022).

Nessa sequência, a frequência de atualização dos dados também interfere na escolha dos padrões de integração, uma vez que ambientes com processamento periódico operam sob lógicas diferentes daqueles que dependem de fluxos contínuos ou quase em tempo real, essa distinção afeta a modelagem dos pipelines e a forma de coordenar captura, transformação e disponibilização analítica, em contextos mais dinâmicos, a integração precisa preservar estabilidade sem comprometer a velocidade de resposta do ambiente (Fikri *et al*., 2019).

Além disso, a eficiência dos padrões de integração está ligada à capacidade de documentar a origem dos dados, registrar as transformações aplicadas e manter visibilidade sobre os relacionamentos entre diferentes conjuntos informacionais, tais elementos fortalecem o controle do ciclo de vida dos dados e reduzem ambiguidades na interpretação analítica, essa condição aproxima a integração das práticas de governança e da sustentação contínua dos pipelines analíticos (Nadal *et al*., 2022).

Sob uma ótica mais ampla, a presença de metadados estruturados contribui para organizar a navegação pelos ativos informacionais e ampliar a inteligibilidade das arquiteturas de dados, esse recurso favorece o entendimento sobre procedência, uso e transformação dos registros, permitindo maior segurança na reutilização das informações ao longo dos fluxos analíticos, desse modo, a integração passa a depender também da qualidade com que o ambiente descreve seus próprios componentes (Sawadogo; Darmont, 2021).

Quando se examinam arquiteturas contemporâneas, percebe-se que os padrões de integração precisam acomodar diferentes escalas de processamento e distintas expectativas de consumo analítico, pois áreas de negócio, times técnicos e sistemas automatizados operam com demandas próprias de latência, granularidade e atualização, essa multiplicidade reforça a necessidade de estruturas integradoras capazes de sustentar usos diversos sem perda de consistência arquitetural (Dhaouadi *et al*., 2022).

Diante desse cenário, os padrões de integração assumem função determinante na articulação entre armazenamento, processamento e consumo analítico, uma vez que deles depende a fluidez com que os dados percorrem o ambiente informacional até se converterem em insumos interpretáveis, por esse motivo, a integração precisa ser planejada como componente permanente da arquitetura e não como procedimento acessório vinculado apenas à carga inicial dos dados (Machado *et al*., 2019).

Desse modo, compreender os padrões de integração em ambientes analíticos permite reconhecer como diferentes arranjos técnicos condicionam a qualidade dos fluxos, a estabilidade dos pipelines e a confiabilidade da informação utilizada nas análises, essa compreensão oferece base teórica para examinar, na sequência, as abordagens de ETL e ELT como expressões operacionais dessas escolhas arquiteturais, articulando estrutura, processamento e sustentação contínua dos ecossistemas de dados (Harby; Zulkernine, 2022).

2.3 ETL, ELT E SUSTENTAÇÃO DE *PIPELINES* ANALÍTICOS

No âmbito das arquiteturas de dados para *analytics*, as abordagens de ETL e ELT representam formas distintas de organizar o deslocamento e o tratamento dos dados ao longo dos ambientes analíticos, ambas buscam viabilizar integração, disponibilidade e aproveitamento informacional, embora se diferenciem quanto ao local de processamento, à sequência das operações e à lógica de sustentação adotada nos fluxos de dados (Dhaouadi *et al*., 2022).

Na formulação clássica, o ETL estrutura o fluxo a partir da extração dos dados nas fontes de origem, seguida da transformação em camadas intermediárias e do carregamento no ambiente de destino, esse modelo foi amplamente difundido em arquiteturas centradas em data *warehouse*, nas quais a padronização prévia dos dados favorecia maior controle sobre a consistência analítica e sobre a organização das estruturas de consulta (Nambiar *et al*., 2022).

Com a ampliação das capacidades de armazenamento e processamento dos ambientes analíticos contemporâneos, o ELT passou a ganhar espaço como abordagem mais compatível com ecossistemas flexíveis e escaláveis, nessa lógica, os dados são inicialmente carregados no ambiente de destino e transformados posteriormente, o que amplia a elasticidade operacional e redistribui o esforço computacional ao longo das plataformas analíticas (Harby; Zulkernine, 2022).

Essa distinção entre ETL e ELT repercute diretamente sobre a forma como os pipelines são desenhados, monitorados e mantidos, pois cada abordagem estabelece diferentes exigências para orquestração, controle de versões, rastreamento das transformações e administração de falhas, por essa razão, a escolha entre um modelo e outro precisa considerar o alinhamento entre arquitetura, objetivos analíticos e maturidade operacional do ambiente (Machado *et al*., 2019).

Em contextos nos quais a organização depende de dados com atualização frequente e baixa tolerância à latência, a sustentação dos pipelines exige mecanismos capazes de garantir continuidade, previsibilidade e rápida recuperação em situações de instabilidade, isso torna a engenharia dos fluxos analíticos uma atividade permanente, associada não apenas à construção inicial do pipeline, mas também à sua manutenção contínua em produção (Fikri *et al*., 2019).

Sob essa leitura, sustentar pipelines analíticos envolve assegurar que cada etapa do fluxo opere com coerência em relação às anteriores e às subsequentes, preservando integridade, sincronização e qualidade dos dados ao longo da cadeia de processamento, essa condição demanda visibilidade sobre dependências técnicas, eventos de execução e transformações aplicadas em diferentes camadas da arquitetura (Nadal *et al*., 2022).

Nessa dinâmica, os metadados assumem relevância elevada, pois descrevem procedência, estrutura, alterações e vínculos entre os ativos informacionais movimentados pelo pipeline, com isso, tornam-se instrumentos importantes para auditoria, rastreabilidade e inteligibilidade dos fluxos,

favorecendo maior segurança na interpretação analítica e maior estabilidade operacional em ambientes complexos (Sawadogo; Darmont, 2021).

Ao se considerar arquiteturas mais recentes, percebe-se que a sustentação dos pipelines está fortemente associada à integração entre armazenamento, processamento e governança, uma vez que fluxos analíticos distribuídos requerem coordenação permanente entre zonas de ingestão, transformação, curadoria e consumo, essa articulação fortalece a capacidade de adaptação do ambiente diante de mudanças de volume, formato e ritmo de atualização dos dados (Wieder; Nolte, 2022).

Em cenários com múltiplas fontes e regras de negócio diversificadas, a manutenção dos pipelines requer padronização de rotinas, documentação consistente e mecanismos de automação que reduzam retrabalho e variabilidade operacional, esses elementos contribuem para estabilizar os fluxos e ampliar a confiabilidade dos conjuntos de dados entregues às camadas analíticas, favorecendo melhor aproveitamento das arquiteturas de dados em contextos corporativos (Machado *et al*., 2019).

Além disso, a comparação entre ETL e ELT não deve ser compreendida como oposição rígida entre modelos excludentes, pois diferentes ambientes podem combinar estratégias distintas conforme a criticidade dos dados, a sensibilidade das transformações e a estrutura tecnológica disponível, nessa perspectiva, a escolha metodológica tende a responder mais ao desenho arquitetural e às exigências de operação do que a uma preferência abstrata por determinada abordagem (Dhaouadi *et al*., 2022).

Essa compreensão reforça a necessidade de observar os pipelines como estruturas vivas dentro dos ecossistemas de dados, sujeitas a ajustes, monitoramento e aperfeiçoamento conforme se alteram as demandas analíticas e as condições técnicas de processamento, por isso, a sustentação dos fluxos depende de decisões continuadas sobre arquitetura, integração, governança e desempenho ao longo do tempo (Nadal *et al*., 2022).

Desse modo, o exame de ETL, ELT e sustentação de pipelines analíticos permite compreender como as operações de dados se materializam dentro das arquiteturas voltadas para *analytics*, articulando procedimentos técnicos, escolhas estruturais e mecanismos de controle contínuo, essa leitura oferece base para interpretar a eficiência e a robustez dos ambientes analíticos à luz das exigências contemporâneas de integração, escalabilidade e confiabilidade informacional (Harby; Zulkernine, 2022).

3 METODOLOGIA

Este estudo foi desenvolvido por meio de abordagem qualitativa, com orientação analítica voltada à compreensão das arquiteturas de dados para *analytics*, dos padrões de integração, das abordagens de ETL e ELT e dos mecanismos de sustentação de pipelines, a escolha desse percurso decorre da natureza conceitual do tema, que exige interpretação articulada de contribuições teóricas e técnicas relacionadas ao funcionamento dos ambientes analíticos contemporâneos.

Quanto aos objetivos, a pesquisa possui caráter descritivo e explicativo, descritivo porque organiza e apresenta os principais elementos constitutivos das arquiteturas de dados e de seus fluxos operacionais, explicativo porque busca compreender as relações estabelecidas entre estrutura arquitetural, integração de dados, processamento e continuidade dos pipelines, permitindo examinar como essas dimensões se influenciam mutuamente na consolidação dos ecossistemas analíticos (Gil, 2008).

No que se refere aos procedimentos, adotou-se a pesquisa bibliográfica como estratégia central de construção do estudo, tendo em vista que a análise foi fundamentada em produções científicas pertinentes ao tema investigado, esse tipo de procedimento possibilita reunir interpretações já consolidadas, identificar aproximações conceituais e construir uma leitura integrada sobre o objeto em exame (Lakatos; Marconi, 2003).

A composição do corpus teórico foi orientada por referências acadêmicas alinhadas ao eixo temático do trabalho, priorizando estudos voltados às arquiteturas de dados, à integração em ambientes analíticos, às lógicas de processamento em ETL e ELT e à sustentação operacional dos pipelines, essa seleção permitiu reunir bases conceituais convergentes e complementaridades relevantes para o desenvolvimento da argumentação.

A organização metodológica da análise foi estruturada em etapas articuladas, inicialmente realizou-se a leitura integral das referências selecionadas, em seguida foram identificados os conceitos centrais, os pontos de aproximação entre os estudos e as distinções interpretativas mais significativas, posteriormente o conteúdo foi agrupado em núcleos temáticos capazes de sustentar a construção progressiva das seções do artigo.

A análise do material ocorreu por meio de leitura interpretativa e comparativa, buscando relacionar os diferentes aportes teóricos a partir de eixos comuns, entre eles os fundamentos das arquiteturas analíticas, os padrões de integração de dados e as formas de operacionalização dos fluxos em ambientes corporativos, esse procedimento favoreceu uma compreensão conectada do tema e contribuiu para a elaboração de um texto coeso entre suas partes.

A opção por uma abordagem não experimental decorre do fato de que o objetivo do trabalho não consiste em testar variáveis em ambiente controlado, mas em examinar criticamente construções teóricas e técnicas presentes na literatura especializada, por essa razão, o estudo se concentra na interpretação de conhecimentos já produzidos, com ênfase na articulação conceitual e na sistematização analítica.

No tratamento do conteúdo, buscou-se preservar coerência entre os objetivos da pesquisa e a estrutura argumentativa adotada, de modo que cada seção fosse desenvolvida como continuidade da anterior, estabelecendo uma progressão temática entre os fundamentos arquiteturais, os mecanismos

de integração e as estratégias de sustentação dos pipelines, essa escolha metodológica fortalece a unidade interna do texto e favorece maior consistência expositiva.

A delimitação do estudo concentrou-se no exame das arquiteturas de dados aplicadas ao contexto de *analytics*, sem avançar para análises empíricas de organizações específicas ou avaliação de ferramentas particulares, essa definição permitiu manter o foco na dimensão conceitual e estrutural do tema, valorizando a compreensão dos modelos, dos fluxos e das relações que sustentam os ambientes analíticos.

Desse modo, a metodologia adotada oferece base adequada para o alcance do objetivo proposto, uma vez que combina pesquisa bibliográfica, leitura analítica e organização temática do conteúdo, permitindo interpretar de forma articulada os elementos que compõem as arquiteturas de dados e seus fluxos operacionais, com isso, o estudo se sustenta em um percurso metodológico compatível com a natureza do objeto investigado e com a proposta analítica do artigo.

4 RESULTADOS E DISCUSSÃO

A análise das referências permitiu identificar que as arquiteturas de dados para *analytics* vêm sendo reorganizadas em torno de maior flexibilidade estrutural, maior capacidade de integração e melhor sustentação dos fluxos analíticos, Nambiar *et al*. (2022) mostram que a distinção entre *data warehouse* e *data lake* revela mudanças no modo de armazenar e disponibilizar dados, Dhaouadi *et al*. (2022) ampliam essa leitura ao evidenciar que a evolução dos processos de carga acompanha a transformação dessas arquiteturas em direção a modelos mais adaptáveis.

Nesse encadeamento, os resultados indicam que o *data warehouse* permanece associado a cenários com maior padronização e forte dependência de organização prévia dos dados, Harby e Zulkernine (2022) observam que esse modelo conserva valor em ambientes que exigem controle analítico mais rígido, Wieder e Nolte (2022) complementam que a expansão dos data *lakes* responde à necessidade de absorver diversidade de fontes e formatos em ecossistemas analíticos mais dinâmicos.

Essa diferença estrutural produz efeitos diretos sobre a integração dos dados, Machado *et al*. (2019) assinalam que a confiabilidade dos fluxos depende de mecanismos capazes de articular origem, transformação e entrega com estabilidade operacional, Sawadogo e Darmont (2021) acrescentam que, sem organização adequada da arquitetura e dos metadados, a ampliação da ingestão tende a comprometer inteligibilidade, rastreabilidade e aproveitamento analítico.

A partir dessa convergência, observa-se que a integração deixou de ser percebida como procedimento acessório e passou a compor a base funcional dos ambientes analíticos, Dhaouadi *et al*. (2022) argumentam que o reposicionamento dos processos de carga reflete a mudança de paradigma nas arquiteturas de dados, Nambiar *et al*. (2022) reforçam que a escolha do arranjo arquitetural interfere

na maneira como as organizações conciliam desempenho, estruturação e disponibilidade informacional.

No que se refere às abordagens de processamento, os resultados apontam que ETL e ELT representam respostas distintas a demandas também distintas de arquitetura e operação, Harby e Zulkernine (2022) indicam que a redistribuição das transformações para o ambiente de destino amplia a aderência do ELT a plataformas mais escaláveis, Machado *et al*. (2019) mostram que a lógica do ETL ainda oferece consistência relevante em cenários nos quais o tratamento prévio favorece maior previsibilidade do fluxo.

Em continuidade a esse ponto, percebe-se que a escolha entre ETL e ELT não se resume a preferência técnica isolada, Fikri *et al*. (2019) demonstram que ambientes com necessidade de resposta mais rápida dependem de arranjos capazes de sustentar integração e atualização com menor latência, Wieder e Nolte (2022) completam que a elasticidade das arquiteturas contemporâneas exige formas de processamento compatíveis com ingestão ampliada e diferentes ritmos de consumo analítico.

Os resultados também evidenciam que a sustentação dos pipelines ocupa lugar central na discussão sobre *analytics*, Nadal *et al*. (2022) mostram que governança operacional, visibilidade sobre os fluxos e automação das rotinas fortalecem a continuidade dos ambientes de dados, Sawadogo e Darmont (2021) acrescentam que essa sustentação depende da capacidade de descrever procedência, transformação e uso dos dados ao longo de toda a arquitetura.

Sob essa perspectiva, os metadados emergem como componente articulador entre estrutura e operação, Sawadogo e Darmont (2021) destacam que arquiteturas mais abertas exigem mecanismos refinados de descrição e catalogação, Nadal *et al*. (2022) aprofundam essa compreensão ao demonstrar que governança automatizada e documentação consistente ampliam controle, auditabilidade e segurança interpretativa nos pipelines analíticos.

Essas constatações permitem discutir que a robustez de uma arquitetura de dados não decorre apenas da tecnologia empregada, Nambiar *et al*. (2022) indicam que a adequação entre modelo arquitetural e finalidade analítica condiciona a qualidade do ambiente, Dhaouadi *et al*. (2022) completam que a modelagem dos processos de carga e transformação precisa acompanhar essa adequação para que a arquitetura opere de forma coesa e contínua.

Ao comparar os estudos, torna-se visível que as arquiteturas híbridas ganham relevância por tentarem reunir estruturação analítica e flexibilidade operacional, Harby e Zulkernine (2022) entendem o *lakehouse* como expressão dessa aproximação entre controle e elasticidade, Wieder e Nolte (2022) observam que a centralidade dos data lakes nos ecossistemas contemporâneos favorece a formação de ambientes mais integrados, com maior abertura para usos analíticos diversos.

Essa aproximação entre modelos revela que a discussão atual se desloca da oposição entre arquiteturas para a análise de suas combinações possíveis, Dhaouadi *et al*. (2022) mostram que a

evolução dos processos acompanha essa transição, Fikri *et al*. (2019) contribuem ao indicar que a adaptação arquitetural depende da capacidade de responder a requisitos concretos de desempenho, atualização e estabilidade operacional.

No plano da integração, os resultados sugerem que fluxos heterogêneos exigem coordenação permanente entre camadas e rotinas, Machado *et al*. (2019) defendem que a distribuição adequada das tarefas ao longo do pipeline favorece continuidade analítica e menor vulnerabilidade operacional, Sawadogo e Darmont (2021) completam que a ausência de mecanismos claros de organização compromete a leitura do ambiente e enfraquece o valor analítico dos dados integrados.

Em termos de discussão, verifica-se que a eficiência analítica está diretamente relacionada à coerência entre arquitetura, integração e sustentação, Nadal *et al*. (2022) apontam que governança e automação precisam acompanhar o crescimento dos ambientes de dados, Nambiar *et al*. (2022) reforçam que a arquitetura deve ser pensada em função do uso analítico pretendido, evitando dissociação entre estrutura técnica e finalidade informacional.

Os achados permitem compreender, ainda, que a manutenção dos pipelines deve ser tratada como atividade contínua e estruturante do ecossistema analítico, Wieder e Nolte (2022) observam que ambientes modernos operam sob exigências crescentes de diversidade e escala, Harby e Zulkernine (2022) acrescentam que a resposta a essas exigências passa pela combinação entre modelos arquiteturais, formas de processamento e critérios de organização dos fluxos.

Dessa forma, os resultados e a discussão convergem para a compreensão de que arquiteturas de dados para *analytics* dependem de alinhamento entre armazenamento, integração, processamento e governança para sustentar fluxos confiáveis e úteis às análises, Dhaouadi *et al*. (2022) sintetizam esse movimento ao evidenciar a transformação dos processos de carga em direção a modelos mais flexíveis, Nadal *et al*. (2022) completam essa leitura ao mostrar que a continuidade dos pipelines exige coordenação, visibilidade e controle permanente sobre todo o ciclo dos dados.

5 CONSIDERAÇÕES FINAIS

A análise desenvolvida ao longo deste estudo permitiu compreender que as arquiteturas de dados voltadas para *analytics* vêm sendo reconfiguradas em função da ampliação dos fluxos informacionais, da diversidade de fontes e da necessidade de maior estabilidade nos processos analíticos, esse movimento evidencia que a estrutura dos ambientes de dados interfere diretamente na qualidade das análises, na organização dos fluxos e na capacidade de sustentação das operações informacionais.

Ao longo da discussão, verificou-se que *data warehouse*, *data lake* e *lakehouse* representam respostas arquiteturais associadas a diferentes formas de organizar armazenamento, processamento e consumo analítico, essa constatação mostra que a escolha de uma arquitetura não deve ser orientada

por tendência tecnológica isolada, e sim pela coerência entre a configuração do ambiente e as exigências concretas de integração e uso dos dados.

Também se tornou evidente que os padrões de integração ocupam posição central na consolidação dos ecossistemas analíticos, pois é por meio deles que dados dispersos, heterogêneos e oriundos de múltiplos sistemas passam a compor fluxos mais consistentes e interpretáveis, essa compreensão reforça que a integração precisa ser tratada como elemento constitutivo da arquitetura e da continuidade operacional dos pipelines.

No exame das abordagens de processamento, observou-se que ETL e ELT respondem a necessidades distintas e se articulam de maneira diversa com os modelos arquiteturais contemporâneos, essa percepção permitiu concluir que a definição entre uma estratégia e outra depende da combinação entre capacidade computacional, frequência de atualização, volume de dados e forma de consumo analítico pretendida pelas organizações.

A discussão mostrou, ainda, que a sustentação dos pipelines analíticos exige mais do que rotinas de carga e transformação, ela depende de monitoramento, documentação, rastreabilidade e coordenação entre as diferentes camadas do ambiente de dados, por essa razão, a manutenção dos fluxos deve ser compreendida como atividade contínua, vinculada à estabilidade do ecossistema analítico e à confiabilidade das informações disponibilizadas.

Outro aspecto evidenciado pelo estudo refere-se à relevância dos metadados e da governança para a inteligibilidade dos ambientes analíticos, uma vez que a clareza sobre origem, transformação e circulação dos dados amplia a capacidade de controle e favorece maior segurança interpretativa, essa condição fortalece a compreensão de que arquitetura e governança precisam caminhar de forma articulada para sustentar análises mais consistentes.

Diante disso, o objetivo proposto foi alcançado, pois o trabalho permitiu analisar as arquiteturas de dados para *analytics* a partir de seus fundamentos estruturais, de seus padrões de integração, das abordagens de ETL e ELT e dos mecanismos de sustentação dos pipelines, essa trajetória tornou possível construir uma leitura conectada sobre as relações que estruturam os ambientes analíticos contemporâneos.

Por fim, considera-se que o estudo contribui para ampliar a compreensão teórica sobre um tema cada vez mais presente nas organizações orientadas por dados, ao oferecer uma interpretação articulada entre arquitetura, integração, processamento e continuidade operacional, o texto fornece base para pesquisas futuras que aprofundem aplicações empíricas, comparações entre contextos organizacionais e avaliações sobre a evolução dos modelos analíticos em diferentes cenários.

REFERÊNCIAS

DHAOUADI, Asma; BOUSSELMI, Khadija; GAMMOUDI, Mohamed Mohsen; MONNET, Sébastien; HAMMOUDI, Slimane. Data warehousing process modeling from classical approaches to big data and ELT: state of the art and future trends. Data, v. 7, n. 8, 2022.

FIKRI, Noussair; RIDA, Mohamed; ABGHOUR, Noureddine; MOUSSAID, Khalid; EL OMRI, Amina. An adaptive and real-time based architecture for financial data warehouse. Journal of Big Data, v. 6, 2019.

GIL, Antonio Carlos. Métodos e técnicas de pesquisa social. 6. ed. São Paulo: Atlas, 2008.

HARBY, Ahmed A.; ZULKERNINE, Farhana. From data warehouse to lakehouse: a comparative review. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA, 2022. Anais [...]. IEEE, 2022.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. Fundamentos de metodologia científica. 5. ed. São Paulo: Atlas, 2003.

MACHADO, Gustavo V.; CUNHA, Ítalo; PEREIRA, Adriano C. M.; OLIVEIRA, Leonardo B. DOD-ETL: distributed on-demand ETL for near real-time business intelligence. Journal of Internet Services and Applications, v. 10, 2019.

NADAL, Sergi; JOVANOVIC, Petar; BILALLI, Besim; ROMERO, Oscar. Operationalizing and automating data governance. Journal of Big Data, v. 9, 2022.

NAMBIAR, Athira; MUNDRA, Divyansh. An overview of data warehouse and data lake in modern enterprise data management. Big Data and Cognitive Computing, v. 6, n. 4, 2022.

SAWADOGO, Pegdwendé; DARMONT, Jérôme. On data lake architectures and metadata management. arXiv, 2021.

WIEDER, Philipp; NOLTE, Hendrik. Toward data lakes as central building blocks for data management and analysis. Frontiers in Big Data, v. 5, 2022.