Uso de *Shadow Pipelines* para Garantir a Adequação de *Pipelines* ETL à LGPD*

Raíza Albuquerque[1], Marcos Bedo[1], José Maria Monteiro[2]
Lenora Schwaitzer[3], Daniel de Oliveira[1]

[1]Universidade Federal Fluminense (UFF), Niterói, RJ, Brasil

[2]Universidade Federal do Ceará (UFC), CE, Brasil

[3]Universidade Federal do Espírito Santo (UFES), ES, Brasil

raizac@id.uff.br, monteiro@dc.ufc.br, marcosbedo@ic.uff.br
lenora.schwaitzer@ufes.br, danielcmo@ic.uff.br

Resumo. *A Lei Geral de Proteção de Dados Pessoais (LGPD) define diretrizes para a coleta, armazenamento e uso de dados pessoais no Brasil. Apesar de existirem soluções para adequação de sistemas de informação à LGPD, os pipelines de Extração, Transformação e Carga (ETL), fundamentais nas organizações, ainda carecem de soluções específicas voltadas à adequação legal. Este artigo propõe uma abordagem baseada em shadow pipelines para adaptar pipelines ETL existentes à LGPD. A abordagem permite a verificação de algumas hipóteses para tratamento de dados previstas no seu artigo 7º. A proposta foi avaliada por meio de estudo de viabilidade utilizando pipelines sintéticos. Os resultados demonstram que a abordagem contribui para a adequação à LGPD, preservando a estrutura e desempenho do pipeline original.*

Abstract. *The General Data Protection Law (LGPD) defines guidelines for the collection, storage, and use of personal data in Brazil. Although there are solutions for adapting information systems to the LGPD, Extraction, Transformation, and Loading (ETL) pipelines, which are fundamental within organizations, still lack specific legal compliance solutions. This paper proposes an approach based on shadow pipelines to adapt existing ETL pipelines to the LGPD. The approach enables the verification of legal basis for processing provided for in its article 7. The proposal was evaluated through a feasibility study using synthetic pipelines. The results demonstrate that the approach supports LGPD compliance while preserving the original pipeline's structure and performance.*

1. Introdução

Apesar de a proteção à intimidade e à vida privada já estarem previstas na Constituição Federal de 1988 [Brasil 1988], o tema ganhou maior relevância prática para organizações públicas e privadas a partir da promulgação da Lei Geral de Proteção de Dados Pessoais (LGPD)[1]. A LGPD estabelece diretrizes para o tratamento de dados pessoais no Brasil, impondo obrigações a qualquer organização que armazene ou processe informações sensíveis

*Os autores gostariam de agradecer pelo apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), do CNPq *Grant* 311898/2021-1 e da FAPERJ *Grants* E-26/204.238/2024 and E-26/204.544/2024.

[1]https://www.planalto.gov.br/ccivil_03/ _ato2015-2018/2018/lei/l13709.htm

de cidadãos brasileiros. De acordo com a lei, os titulares dos dados têm o direito de saber como suas informações são coletadas, armazenadas, utilizadas e eventualmente compartilhadas, conforme discutido em [Nascimento and Silva 2023]. Além disso, em alguns casos, os indivíduos podem consentir, restringir ou negar o uso de seus dados conforme a finalidade declarada pelo controlador de dados. Por exemplo, ao fornecer informações pessoais para participar de um sorteio, o titular dos dados deve manifestar seu consentimento para o envio de campanhas promocionais ou para outras entidades.

Apesar dos avanços proporcionados pela LGPD, ainda há uma distância entre o que a legislação estabelece e o que é efetivamente implementado pelas organizações. Um *survey* conduzido pela União Europeia (UE) [GDPR.EU 2019], que avaliou o processo de adequação de empresas ao Regulamento Geral de Proteção de Dados (GDPR, do inglês *General Data Protection Regulation*), indicou que 56% das organizações reconhecem que o uso dos dados pessoais nem sempre é claro para os titulares, o que pode comprometer a conformidade com a regulamentação. Essa constatação foi corroborada por um relatório publicado pela UE[2]. No contexto brasileiro, a situação não é distinta: trabalhos recentes [Cabral et al. 2023, Sousa et al. 2020] indicam que muitas organizações ainda se encontram em processo de adaptação à LGPD, especialmente no que se refere à conformidade de sistemas com as exigências legais. A adequação à LGPD nas organizações é complexa, uma vez que a criação e a gerência de controles sobre dados pessoais envolvem, além da compreensão da própria norma, a superação de desafios técnicos. Tais controles exigem alterações em uma pilha de *software*, composta por diversos sistemas, bibliotecas e dependências. Embora existam arcabouços que apoiam o desenvolvimento de sistemas de informação em conformidade com a LGPD [Saraiva and Soares 2023, Marques et al. 2021, Martins et al. 2020, Barros et al. 2024], observa-se uma atenção limitada aos *pipelines* de Extração, Transformação e Carga (ETL), elementos centrais no fluxo e tratamento de dados em ambientes organizacionais.

Ao contrário dos sistemas de informação, nos quais as regras de negócio são mais estáveis e geralmente são precedidas por etapas formais de levantamento de requisitos, os *pipelines* ETL são, em sua maioria, desenvolvidos de forma dinâmica e sob demanda [Yang et al. 2015, Vieira et al. 2024]. Frequentemente criados para atender a necessidades urgentes de análise por parte dos tomadores de decisão, esses *pipelines* não passam por processos sistemáticos de especificação como ocorre no desenvolvimento tradicional de sistemas. Independentemente da tecnologia utilizada para sua implementação, *e.g.*, *scripts* Python ou o Pentaho, os *pipelines* ETL exercem papel central no tratamento e distribuição de dados dentro das organizações. Por esse motivo, tornam-se elementos estratégicos para a adequação da organização à LGPD, uma vez que operam diretamente sobre os dados pessoais que circulam no ambiente corporativos. No entanto, as abordagens existentes que visam garantir a adequação dos *pipelines* com legislações de proteção de dados (LGPD ou o GDPR), ainda são incipientes [Cerqueira et al. 2023, Gruschka et al. 2018, Liu 2019], e não focam na verificação das bases legais para o tratamento de dados.

Dessa forma, este artigo foca na adaptação de *pipelines* ETL com o objetivo de adequar os mesmos aos princípios e exigências estabelecidos pela LGPD. Para isso, propomos a abordagem `Aether` desenvolvida para realizar a instrumentação semiautomática de *pipelines* ETL por meio da inserção de *shadow pipelines* [Grafberger et al. 2024]. Os *shadow pipeli-*

[2]https://fra.europa.eu/en/publication/2024/ gdpr-experiences-data-protection-authorities.

nes consistem em estruturas complementares, integradas ao código original do *pipeline* ETL, com a finalidade de verificar se o tratamento está de acordo com algumas hipóteses previstas no art. 7º da LGPC, quais sejam: o consentimento (art. 7º, I), o cumprimento de obrigação legal ou regulatória (art. 7º, II), e para a execução de contrato ou de procedimentos preliminares relacionados a contrato do qual seja parte o titular (art. 7º, V). Com isso, busca-se garantir que os dados processados estejam em conformidade com a LGPD sem comprometer a estrutura funcional do *pipeline* original. A `Aether` foi avaliada por meio de um estudo de viabilidade, no qual foram utilizados *pipelines* ETL sintéticos. Os resultados obtidos indicam que a abordagem proposta é viável e possui potencial para ser integrada a processos organizacionais com o objetivo de fortalecer a governança de dados.

2. Princípios Básicos da Lei Geral de Proteção de Dados

A proteção de dados pessoais ganhou destaque nos últimos anos, tornando-se um tema central nas discussões sobre o tratamento de informações [Zaguir 2024]. No Brasil, a coleta de dados pessoais é uma prática recorrente, como o preenchimento de cadastros em *sites*. Apesar de sermos os titulares desses dados, em geral não temos conhecimento ou controle efetivo sobre como tratados [Zaguir 2024], o que evidencia a necessidade de normativas que resguardem os direitos dos indivíduos. Nesse contexto, surgem legislações voltadas à proteção de dados, como o GDPR. O GDPR garante maior transparência e controle aos indivíduos sobre suas informações pessoais, servindo de inspiração para legislações em diversos países. Segundo [Zaguir 2024], até 2023, 162 países já haviam adotado leis semelhantes. A tendência é que, nos próximos anos, países atualmente sem regulamentações específicas devem aderir a esse movimento, evidenciando a crescente valorização da privacidade.

Seguindo essa tendência, a LGPD estabelece diretrizes para o tratamento ético e legal de dados pessoais, com inspiração no GDPR e foco na proteção dos direitos fundamentais de liberdade e privacidade [Schwaitzer 2020]. A LGPD, como disposto em seu Art. 1º, visa regular o tratamento de dados pessoais, físicos ou digitais, por pessoas naturais ou jurídicas, públicas ou privadas, de modo a assegurar a proteção de direitos fundamentais. Essa regulamentação implica que qualquer tratamento de dados deve ocorrer de forma ética, transparente e dentro dos limites legais. A diferenciação entre dados pessoais e dados pessoais sensíveis é essencial nesse processo, visto que as hipóteses para o tratamento de dados pessoais sensíveis são mais restritas.

A LGPD define *dados pessoais* como qualquer informação relacionada a uma pessoa natural identificada ou identificável, enquanto *dados pessoais sensíveis* correspondem a informações cujo uso pode ensejar discriminação, tais como origem racial ou étnica, convicções religiosas e opiniões políticas [Magrani 2019]. Para que o tratamento dessas informações seja considerado legítimo, ele deve estar respaldado em uma das bases legais estabelecidas nos artigos 7º (para dados pessoais) e 11 (para dados pessoais sensíveis) da LGPD. A Tabela 1 apresenta um resumo das hipóteses legais e seus respectivos fundamentos normativos. É importante ressaltar que este artigo foca apenas em três hipóteses: *(i)* consentimento, *(ii)* execução de contrato e *(iii)* obrigação legal (interpretada a partir da verificação da finalidade). As hipóteses $H1$, $H2$ e $H5$ foram selecionadas por serem aplicáveis a um conjunto mais amplo de *pipelines* ETL em diferentes domínios. Por exemplo, enquanto $H1$, $H2$ e $H5$ são independentes do domínio de aplicação e, portanto, precisam ser verificadas em diversos tipos de *pipelines*, a hipótese $H10$ se aplica apenas a contextos financeiros.

Tabela 1. Hipóteses de Tratamento de Dados Pessoais

	Dispositivos Legais	
Hipótese de Tratamento	**Dados Pessoais**	**Dados Pessoais Sensíveis**
H1: Mediante consentimento do titular	LGPD, art. 7º, I	LGPD, art. 11, I
H2: Para o cumprimento de obrigação legal ou regulatória	LGPD, art. 7º, II	LGPD, art. 11, II, "a"
H3: Para a execução de políticas públicas	LGPD, art. 7º, III	LGPD, art. 11, II, "b"
H4: Para a realização de estudos e pesquisas	LGPD, art. 7º, IV	LGPD, art. 11, II, "c"
H5: Para a execução ou preparação de contrato	LGPD, art. 7º, V	Não se aplica
H6: Para o exercício de direitos em processo judicial, administrativo ou arbitral	LGPD, art. 7º, VI	LGPD, art. 11, II, "d"
H7: Para a proteção da vida ou da incolumidade física do titular ou de terceiro	LGPD, art. 7º, VII	LGPD, art. 11, II, "e"
H8: Para a tutela da saúde do titular	LGPD, art. 7º, VIII	LGPD, art. 11, II, "f"
H9: Para atender interesses legítimos do controlador ou de terceiro	LGPD, art. 7º, IX	Não se aplica
H10: Para proteção do crédito	LGPD, art. 7º, X	Não se aplica
H11: Para a garantia da prevenção à fraude e à segurança do titular	Não se aplica	LGPD, art. 11, II, "g"

3. Modelo de Aplicação da `Aether`

No escopo deste artigo, um *pipeline* ETL é representado por um grafo acíclico dirigido (DAG) $P = (D \cup N, A)$, em que os vértices correspondem a dois subconjuntos distintos: N, que representa as etapas de processamento, e D, que representa os arquivos ou *datasets* manipulados. As arestas pertencentes ao conjunto A estabelecem relações de dependência entre os dados e as etapas, indicando quais *datasets* são consumidos e produzidos por cada etapa.

Cada etapa $n_i \in N$ está associada à execução de um programa ou *script*, e realiza tarefas típicas do ciclo de ETL, tais como remoção de valores ausentes, deduplicação de registros, eliminação de *outliers*, *etc*. Formalmente, define-se que uma etapa n_i realiza a transformação de um subconjunto de dados de entrada D_{input} em um novo subconjunto de saída D_{output}, conforme a função $n_i : D_{\text{input}} \mapsto D_{\text{output}} \subseteq D$. Uma vez definido o *pipeline* ETL principal, a `Aether` é capaz de injetar um *shadow pipeline* que atua paralelamente ao fluxo original, com o objetivo de assegurar a conformidade com a LGPD. Esse *shadow pipeline* pode ser igualmente modelado como um DAG, denotado por $S = (D' \cup N' \cup M, A')$. Nesse modelo, o conjunto N' corresponde às etapas responsáveis por realizar verificações relacionadas às hipóteses legais de tratamento de dados. Para o escopo do estudo, essas transformações se referem especificamente a *(i)* finalidade do uso (n_{goal}), *(ii)* base contratual (n_{contract}) e ao *(iii)* consentimento do titular (n_{consent}), *i.e.,* $N' = \{n_{\text{goal}}, n_{\text{contract}}, n_{\text{consent}}\}$.

O conjunto M representa as etapas que aplicam modificações nos dados produzidos pelo *pipeline* ETL original, com base nos resultados das verificações realizadas pelas etapas em N'. Já D' representa os conjuntos de dados gerados e transformados no contexto do *shadow pipeline*, assegurando sua adequação à LGPD. As arestas A' descrevem as relações de dependência entre os elementos de dados e as etapas de checagem e processamento, especificando quais *datasets* são consumidos e produzidos ao longo da execução do *shadow pipeline*. Formalmente, uma etapa de modificação de dados m_j realiza a adequação de um subconjunto de dados de entrada D'_{input} em um novo subconjunto de saída D'_{output}, conforme a função $m_j : D'_{\text{input}} \mapsto D'_{\text{output}} \subseteq D'$.

Uma vez que o *shadow pipeline* é injetado no *pipeline* ETL principal, a saída final do processo de transformação de dados passa a ser um conjunto $d_k \in D'_{output}$, *i.e.*, um dado resultante do processamento realizado exclusivamente após a execução das etapas de conformidade previstas no *shadow pipeline*. Dessa forma, garante-se que os dados produzidos ao término do *pipeline* estejam em conformidade com os princípios e as bases legais estabelecidas pela LGPD. A Figura 1 apresenta um exemplo de *pipeline* ETL original (blocos em

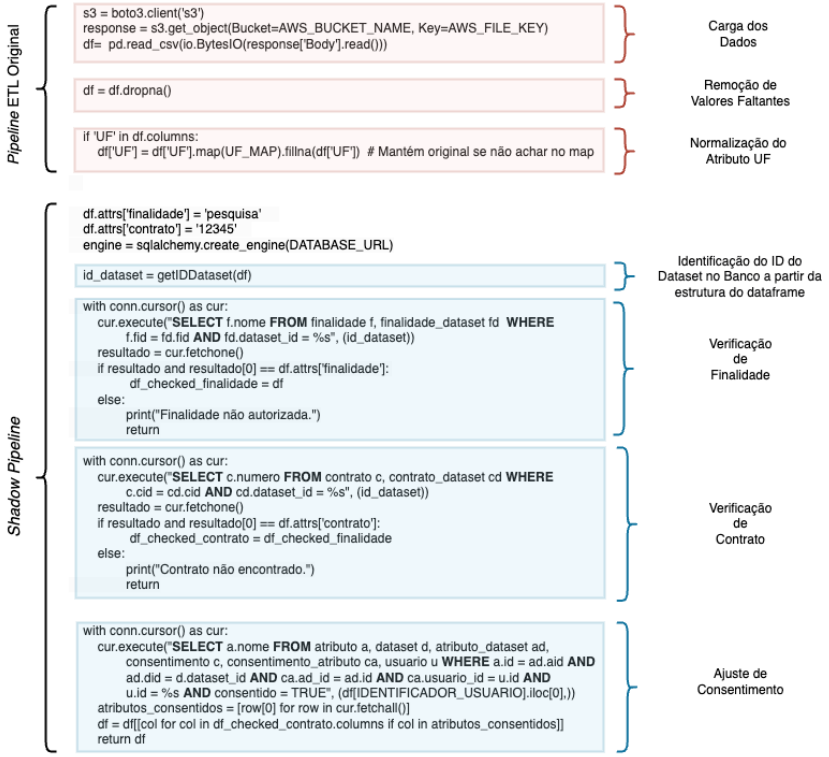

Figura 1. Exemplo de ETL (vermelho) com a injeção do *shadow pipeline* (azul).

vermelho) com a injeção de um *shadow pipeline* (blocos em azul).

4. A abordagem `Aether`

De modo a adaptar *pipelines* ETL à LGPD por meio da injeção de *shadow pipelines*, propomos a `Aether`. A arquitetura da `Aether`, ilustrada na Figura 2, é organizada em quatro camadas: *(i)* API, *(ii)* Camada de Instrumentação, *(iii)* Camada de Execução e *(iv)* Camada de Dados. Cada camada possui responsabilidades específicas no processo de adequação [Loureiro and de Oliveira 2022]. A seguir, descrevemos as funções e componentes de cada uma delas. A API da `Aether` define os pontos de entrada responsáveis pela interação entre os usuários e os demais componentes da arquitetura. Por meio dela, é possível submeter *pipelines* ETL, registrar metadados relevantes (como finalidades autorizadas, contratos vigentes e consentimentos informados), acionar a injeção do *shadow pipeline* e iniciar a execução do *pipeline* já adequado. A utilização exclusiva da API assegura que todos os *pipelines* sejam executados por meio da `Aether`, impedindo a execução direta de processos que não estão em conformidade com a LGPD.

A *Camada de Instrumentação* tem como função analisar a estrutura do *pipeline* ETL submetido, identificar seus pontos de saída de dados e injetar o *shadow pipeline* com as verificações de finalidade, contrato e consentimento. A identificação desses pontos é feita a partir da análise de estruturas de dados onde os resultados são armazenados, como arquivos ou *dataframes*. Na versão atual, essa identificação requer que o usuário informe, previamente ao processo de instrumentação, o *script* ETL original, o nome do *dataframe* que representa a saída final do *pipeline*, a finalidade de uso e o atributo do *dataframe* que é usado para

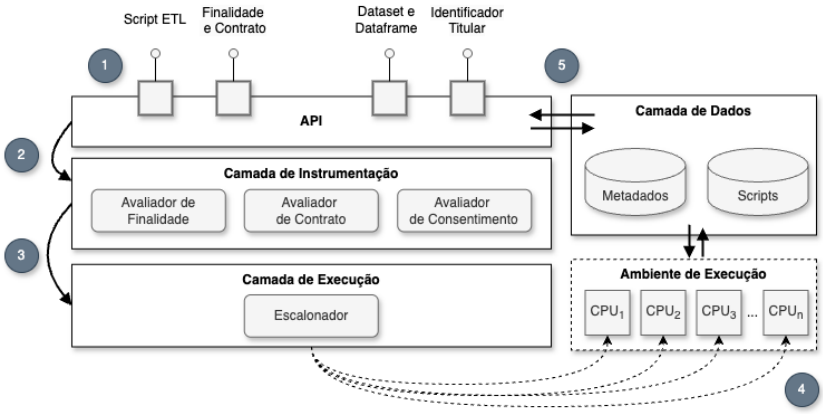

Figura 2. Arquitetura da Abordagem `Aether`

identificar o titular dos dados. Com base nessas informações, a injeta ao final do código original as instruções responsáveis por aplicar as verificações e transformações exigidas pelas políticas de conformidade com a LGPD. O procedimento seguido pela `Aether` para realizar a instrumentação é apresentada no Algoritmo 1. O *script* associado ao *pipeline* ETL é parseado e identificado o último ponto em que o *dataframe* informado pelo usuário é utilizado. Uma vez identificado esse ponto, a `Aether` insere os controles de finalidade, contrato e consentimento, similar ao *script* apresentado na Figura 1. Uma vez que o *pipeline* se encontra instrumentado, durante a execução a `Aether` verifica primeiramente a finalidade do uso dos dados. Caso a finalidade não seja a mesma da permitida, nenhum dado de saída é produzido. Em seguida, verifica-se a existência de contrato, e, caso não exista nenhum contrato vigente, nenhum dado é gerado na saída. Finalmente, são verificados os atributos que os usuários consentiram em permitir o uso. Somente esses atributos de cada tupla são disponibilizados pelo *pipeline* ETL na saída.

Algorithm 1: Instrumentação do *Script* ETL

Input: *Script* ETL P, nome do *dataframe* d, atributo identificador a
$M \leftarrow$ **Parse_script**(P);
$p \leftarrow$ **identifica_ponto_injecao**(M, d);
$M \leftarrow$ **injeta_controle_finalidade**(M, p, n_{goal});
$M \leftarrow$ **injeta_controle_contrato**$(M, n_{contract}, a)$;
$M \leftarrow$ **injeta_controle_consentimento**$(M, n_{consent}, a)$;
Salvar M sobrescrevendo S;

No exemplo apresentado na Figura 1, o usuário informa explicitamente que o *dataframe* `df` representa os dados resultantes da execução do *pipeline* ETL original. A `Aether` realiza uma análise para localizar a última ocorrência em que o `df` é utilizado no código, de modo a garantir que, a partir daquele ponto, os dados não sofrerão alterações adicionais. Em seguida, a `Aether` injeta, neste mesmo ponto do código, uma série de verificações destinadas a assegurar a conformidade com os requisitos legais, abarcando a finalidade do uso dos dados, a existência de um contrato que autorize tal utilização e a validação do consentimento dos titulares. Adicionalmente, a `Aether` atribui ao *dataframe* propriedades suplementares que especificam o número do contrato que deve ser consultado e a finalidade para a qual os dados foram autorizados a serem utilizados pelo *pipeline*. No processo de verificação do consentimento, o atributo identificador, previamente fornecido pelo usuário e que faz parte do `df`, é empregado para confirmar a existência de um registro de consentimento válido.

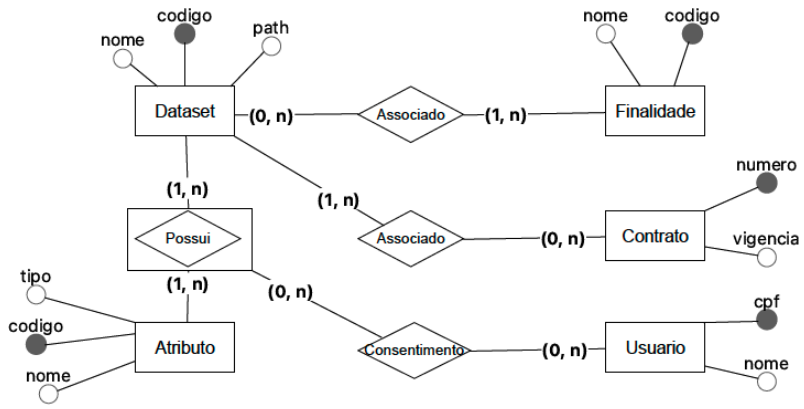

Figura 3. Diagrama conceitual simplificado da base de metadados da `Aether`

Atualmente, essa estratégia de instrumentação usada apresenta uma limitação: ela permite que dados sejam produzidos apenas para, possivelmente, descartá-los posteriormente caso não estejam em conformidade. Uma possibilidade de melhorar esse processo é injetando verificações em etapas anteriores do *pipeline*.

A *Camada de Execução* é responsável por orquestrar a execução do *pipeline* ajustado. Após o processo de injeção do *shadow pipeline*, essa camada executa todas as etapas do *pipeline*, incluindo os blocos de código inseridos pelo *shadow pipeline*, garantindo que apenas dados em conformidade com a LGPD sejam processados. Finalmente, a *Camada de Dados* contém o banco de metadados que armazena os registros atualizados sobre as finalidades autorizadas, contratos vigentes e consentimentos válidos, permitindo sua consulta pelo *shadow pipeline*. Esses metadados são utilizados pelas camadas superiores durante a instrumentação e execução para validar se cada operação está de acordo com a legislação. O banco de metadados da `Aether` segue o diagrama apresentado na Figura 3, o qual registra o nome e a estrutura dos *datasets*, seus atributos, as finalidades e contratos associados, bem como os atributos para os quais o consentimento do usuário foi registrado.

Em sua versão atual, a `Aether` encontra-se configurada para processar *pipelines* ETL implementados por meio de *scripts* Python. Entretanto, a arquitetura subjacente à ferramenta foi projetada de forma modular e extensível, permitindo que, futuramente, sejam integradas outras abordagens de implementação de *pipelines* ETL, *e.g.*, Knime ou Pentaho. A `Aether` pode ser invocada diretamente via linha de comando *python aether.py run <script_original.py> <nome_dataframe> <finalidade_desejada>*. O código-fonte da `Aether` está em fase final de preparação para disponibilização pública e, assim que concluído o processo de revisão e organização, será publicado por meio do seguinte *link* do GitHub: `https://github.com/UFFeScience/aether`.

5. Avaliação Experimental

Esta seção apresenta a avaliação experimental conduzida com o objetivo de verificar a viabilidade da `Aether` na adequação de *pipelines* ETL aos requisitos definidos pela LGPD. A avaliação é estruturada em três subseções: definição das métricas utilizadas (Subseção 5.1), descrição do ambiente e configuração dos experimentos (Subseção 5.2), e, por fim, análise e discussão dos resultados obtidos (Subseção 5.3).

5.1. Métricas

Para avaliar a capacidade da `Aether` em adaptar *pipelines* ETL às exigências da LGPD nas dimensões de finalidade, contrato e consentimento, adotou-se a abordagem GQM (*Goal, Question, Metric*) [Solingen et al. 2002]. Essa abordagem organiza o processo de avaliação em três etapas principais: *(i)* definição das metas de conformidade a serem alcançadas, *(ii)* formulação de questões que relacionam esses objetivos aos requisitos legais e às características do *pipeline* ETL, e *(iii)* definição de métricas que permitem identificar eventuais não conformidades e mensurar o grau de aderência à legislação. A Tabela 2 apresenta as questões formuladas para verificar a adequação à LGPD, bem como as métricas utilizadas na avaliação. As métricas definidas a partir do GQM foram usadas como base para a avaliação.

Tabela 2. Aplicação da GQM no contexto da `Aether`

Meta	Questão	Métrica
Finalidade	Q1. Como assegurar que a finalidade do *pipeline* está alinhada à finalidade autorizada para os dados?	Quantidade (ou percentual) de *pipelines* executados com finalidade distinta da autorizada
Consentimento	Q2. Como garantir que os dados processados tenham sido autorizados pelo titular?	Número (ou percentual) de tuplas processadas sem consentimento
Contrato	Q3. Como verificar a existência de contrato válido que permita o uso dos dados pelo *pipeline*?	Quantidade (ou percentual) de *pipelines* executados com contrato inválido ou inexistente

5.2. Configuração do Experimento e do Ambiente.

De forma a viabilizar a execução dos experimentos, foi desenvolvido um conjunto de *pipelines* ETL sintéticos, elaborados para simular cenários de processamento de dados. Esses *pipelines* foram construídos a partir de um *pipeline* de referência que: (i) lê um arquivo CSV armazenado em um bucket S3 da AWS ou localmente; (ii) carrega os dados em um *DataFrame* Python; (iii) remove registros com dados faltantes; e (iv) normaliza o atributo `Sexo`. A partir desse *pipeline* de referência, foram criadas variantes, abrangendo diferentes estratégias e etapas do processo ETL: *(i)* salvar os dados em formato Parquet em vez de CSV; *(ii)* remover registros com CPFs inválidos; *(iii)* aplicar anonimização utilizando *hash* SHA-256; *(iv)* armazenar os resultados em subdiretórios com marca temporal (*timestamp*); *(v)* carregar os dados a partir de arquivos Parquet; *(vi)* registrar detalhadamente a execução por meio da biblioteca *logging*; *(vii)* anonimizar parcialmente o atributo CPF; *(viii)* detectar e sinalizar registros com endereços ausentes; *(ix)* adicionar uma coluna com a data de processamento; *(x)* exportar os resultados também em formato JSON; *(xi)* validar CPFs com a biblioteca *pycpfcnpj*, e *(xii)* implementar uma versão orientada a objetos.

A geração das variantes utilizadas nos experimentos foi conduzida por meio de engenharia de *prompt*, tendo como suporte o modelo de linguagem GPT-4o-mini. Ao todo, foram geradas 20 variantes do *script* de referência. Tais modificações foram essenciais para avaliar a robustez da `Aether` frente a diferentes padrões de modelagem de *pipelines* ETL. O *dataset* de entrada possui os seguintes atributos: *(i)* CPF, *(ii)* nome, *(iii)* sexo, *(iv)* grupo, *(v)* vacina, *(vi)* lote, *(vii)* dose *(viii)* data de vacinação e *(ix)* local de vacinação. Esse *dataset* é disponibilizado pela Secretaria de Saúde do Recife através do portal de dados abertos [Secretaria de Saúde 2021] e contém informações detalhadas sobre as pessoas vacinadas contra a COVID-19. Nesse *dataset*, os CPFs já são parcialmente anonimizados, então para gerar variantes com o CPF completo, usamos a biblioteca *faker* do Python. As versões utilizadas do *dataset* contém entre 100.000 e um milhão de tuplas (dados de 2021 a 2024). Finalmente, o consentimento para cada usuário foi definido de forma aleatória. Os experimentos foram

conduzidos em um ambiente *notebook* com processador Apple M2 Pro de arquitetura ARM, com 16 GB de memória RAM e Sistema Operacional MacOS. Esse ambiente foi suficiente para executar todos os *pipelines* sintéticos instrumentados pela `Aether` com desempenho estável e sem restrições de recursos.

5.3. Resultados

Para cada uma das 20 variantes do *pipeline* ETL de referência que foram geradas pelo modelo de linguagem, utilizou-se a `Aether` para adequá-las à LGPD. Neste primeiro experimento, cada execução consumiu como entrada o *dataset* contendo um milhão de tuplas. Para cada uma das variantes, foi realizada uma análise a fim de verificar se a `Aether` foi capaz de avaliar adequadamente a finalidade do tratamento, a existência de base legal por meio de contratos, bem como a presença de consentimento explícito fornecido pelo titular dos dados.

A Tabela 3 apresenta um resumo dos resultados obtidos. Podemos observar que a verificação quanto à existência de um contrato vigente e à definição da finalidade foi bem-sucedida em todos os *pipelines* ETL testados. Esse resultado era esperado, tendo em vista que essas duas verificações não dependem diretamente da forma como os dados estão estruturados nos *dataframes* das variantes do *pipeline* original. Esses resultados demonstram a consistência da `Aether` no que tange à interpretação dos metadados contratuais e das finalidades de tratamento. Por outro lado, a verificação relacionada ao consentimento apresentou limitações em alguns casos. Especificamente, nos *pipelines* 9, 13 e 20, constatou-se que o CPF do titular dos dados encontrava-se parcialmente anonimizado no *dataframe*. Nessas situações, a representação do CPF seguia um padrão semelhante a "`***.584.296-**`", o que comprometeu a acurácia da correspondência entre os dados processados e a tabela de consentimento. Para contornar esse problema, a `Aether` adota uma estratégia de correspondência parcial, buscando na tabela de consentimento os CPFs que compartilhem pelo menos seis caracteres com o CPF parcialmente anonimizado no *dataframe*.

No entanto, esse mecanismo pode resultar em ambiguidades, uma vez que CPFs diferentes podem satisfazer esse critério e, assim, serem incorretamente mapeados para um mesmo registro de consentimento. Como consequência, dados pessoais podem ser processados sem a devida autorização do titular, caracterizando um vazamento de informações sensíveis. Nos casos dos *pipelines* mencionados (9, 13 e 20), essa limitação resultou na geração indevida de 34.751 tuplas pelos *pipelines*, as quais não deveriam ter sido liberadas em virtude da ausência de consentimento explícito.

Além da análise da capacidade da ferramenta `Aether` em verificar os critérios de conformidade relacionados ao consentimento, à finalidade e à existência de contrato, também foi avaliado o *overhead* computacional introduzido pelas adequações exigidas pela LGPD. Toda vez que um *pipeline* é executado com a injeção de um *shadow pipeline*, há um custo adicional decorrente do acesso ao banco de metadados, responsável pela recuperação das informações referentes a contratos, finalidades e consentimentos. Para mensurar esse impacto, executamos a `Aether` com o *shadow pipeline* integrado ao *pipeline* de referência, utilizando *datasets* com diferentes volumes: 100.000, 250.000, 500.000, 750.000 e um milhão de tuplas. A Figura 4 apresenta o *overhead* percentual médio introduzido, calculado com base em dez execuções para cada tamanho de *dataset*.

Mesmo nos cenários em que o volume de dados era reduzido,*e.g.*, no *dataset* contendo 100.000 tuplas, cujo tempo de execução do *pipeline* de referência era inferior a 1 segundo, o *overhead* introduzido permaneceu relativamente baixo, em torno de 3,7%. Tal valor pode ser

Tabela 3. Resultados da avaliação da `Aether` quanto a verificação de finalidade, contrato e consentimento para as 20 variantes do *pipeline* de referência.

Nome do *Pipeline*	Finalidade	Contrato	# tuplas sem consentimento
Pipeline 1	✓	✓	0
Pipeline 2	✓	✓	0
Pipeline 3	✓	✓	0
Pipeline 4	✓	✓	0
Pipeline 5	✓	✓	0
Pipeline 6	✓	✓	0
Pipeline 7	✓	✓	0
Pipeline 8	✓	✓	0
Pipeline 9	✓	✓	32.751
Pipeline 10	✓	✓	0
Pipeline 11	✓	✓	0
Pipeline 12	✓	✓	0
Pipeline 13	✓	✓	32.751
Pipeline 14	✓	✓	0
Pipeline 15	✓	✓	0
Pipeline 16	✓	✓	0
Pipeline 17	✓	✓	0
Pipeline 18	✓	✓	0
Pipeline 19	✓	✓	0
Pipeline 20	✓	✓	32.751

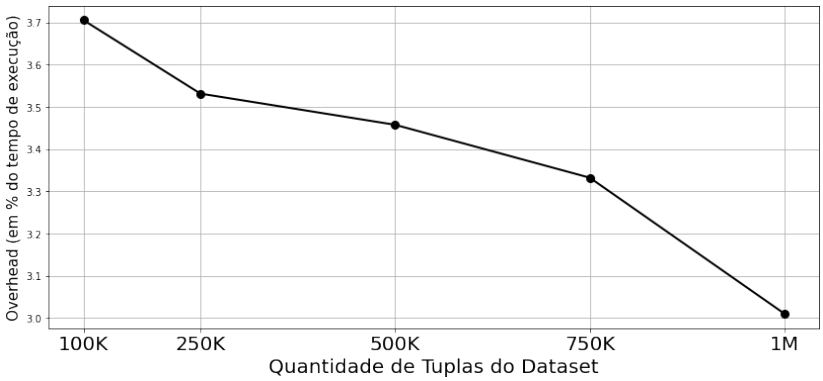

Figura 4. *Overhead* introduzido pela `Aether` no *pipeline* de referência.

considerado aceitável, dado o custo-benefício da adequação à LGPD. Observa-se, ainda, que esse *overhead* tende a decrescer à medida que o tempo de processamento do *pipeline* original aumenta. No caso da execução com um milhão de tuplas, o *overhead* observado foi de aproximadamente 3,0%, o que pode ser considerado aceitável em comparação ao tempo total de execução do *pipeline*. Além da análise do *overhead* introduzido na execução do *pipeline* de referência, realizamos também experimentos com todas as variantes, utilizando *datasets* contendo 100.000 e um milhão de tuplas, com o objetivo de avaliar o impacto da execução do *shadow pipeline* em diferentes configurações. As Figuras 5 e 6 apresentam, respectivamente, os tempos médios de execução dos *pipelines* com (barras pretas) e sem (barras cinzas) a injeção do *shadow pipeline*, para cada uma das variantes analisadas. Observa-se que, de maneira geral, o *overhead* introduzido pela `Aether` manteve-se dentro de limites aceitáveis em todos os casos, não comprometendo o desempenho dos *pipelines* originais. Entretanto, alguns comportamentos particulares merecem destaque.

Um exemplo é o *Pipeline* 10, que incorpora uma tarefa de remoção de duplicatas. No cenário com o *dataset* de um milhão de tuplas, esse *pipeline* passa a apresentar o maior tempo

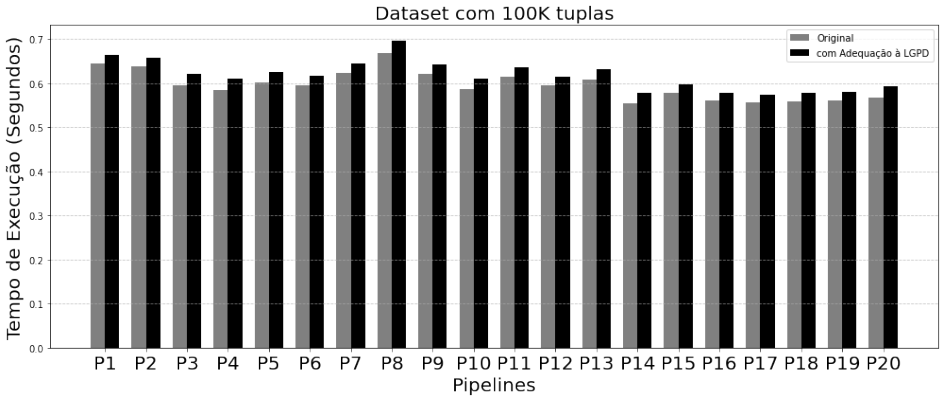

Figura 5. Tempo de execução (em segundos) para cada variante do *pipeline* de referência, com e sem *shadow pipeline*, utilizando um *dataset* com 100.000 tuplas.

de execução entre todas as variantes. Isso ocorre porque a operação de remoção de duplicatas é, por natureza, mais intensiva do ponto de vista computacional quando comparada a outras operações realizadas nos *pipelines*, como normalizações e a exclusão de tuplas com dados ausentes. Apesar disso, mesmo neste caso, o *overhead* adicional decorrente da execução do *shadow pipeline* permaneceu dentro de uma faixa considerada aceitável, reforçando a viabilidade do uso da `Aether` em cenários reais, inclusive em *pipelines* com operações mais complexas.

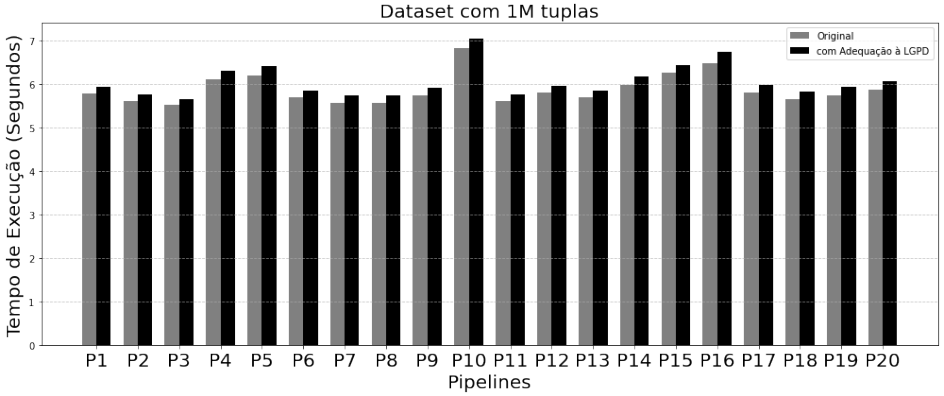

Figura 6. Tempo de execução (em segundos) para cada variante do *pipeline* de referência, com e sem *shadow pipeline* para um *dataset* de um milhão de tuplas.

6. Trabalhos Relacionados

Embora existam diversos trabalhos na literatura que propõem arcabouços e metodologias para o desenvolvimento de *pipelines* de ETL [Vassiliadis 2009], ainda é perceptível a escassez de estudos que abordem especificamente a conformidade desses *pipelines* com a LGPD (e equivalentes). Dessa forma, foi definida a seguinte questão de pesquisa: *QP1: Quais são as abordagens existentes para assegurar a conformidade de pipelines ETL com a LGPD/GDPR?* Para responder a essa questão, optamos por adotar a estratégia de *snowballing* [Jalali and Wohlin 2012] como método de mapeamento sistemático. Definimos como critérios de aceitação na revisão os artigos revisados por pares, publicados após 2016 (ano em que foi promulgado o GDPR), publicados em inglês e português, e que tratem da conformidade de *pipelines* de ETL com as regulamentações de proteção de dados. A *string*

de busca foi submetida a três bibliotecas digitais: *IEEE*, *Scopus* e *ACM Digital Library* e foram identificados 316 artigos. Destes, 99 artigos foram selecionados para revisão com base no título e no resumo. Dentre os 99 artigos selecionados na segunda fase, apenas três [Cerqueira et al. 2023, Gruschka et al. 2018, Liu 2019] abordaram diretamente a adequação de *pipelines* de ETL às leis de proteção de dados. A seguir, discutimos cada um desses trabalhos. [Cerqueira et al. 2023] propõem o LGPDCheck, uma técnica de inspeção baseada em *checklist* para facilitar a identificação da não conformidade com a LGPD em artefatos produzidos em diferentes etapas do ciclo de desenvolvimento de *software*, o que engloba o desenvolvimento de *pipelines* ETL. Apesar de não ser específica para *pipelines* ETL, as recomendações fornecidas podem ser adaptadas para tal contexto. Embora represente um avanço, a abordagem não é automatizada.

[Liu 2019] propõe um modelo para proteção de privacidade de dados em plataformas de *big data*, que englobam a execução de complexos *pipelines* ETL. No entanto, ao analisar a implementação dos *pipelines* de ETL, algumas lacunas importantes são identificadas nessa abordagem. O artigo não apresenta informações claras sobre como os dados são extraídos, transformados e carregados na plataforma, e nem tem objetivo de analisar contratos existentes, finalidades ou consentimento. Por outro lado, a abordagem proposta no artigo foca permite a adição de ruído aos dados para garantir conformidade com regulações. Finalmente, [Gruschka et al. 2018] destacam a importância da proteção de dados em projetos de pesquisa que envolvem *pipelines* ETL que processam grandes *datasets*, em conformidade com regulamentações como o GDPR. Os autores enfatizam a necessidade de consentimento explícito dos participantes e o direito de ser esquecido, porém sem detalhar a implementação prática. Os estudos de caso apresentados abordam a aplicação de métodos de proteção de dados, como anonimização e obtenção de consentimento, mas não descrevem explicitamente a implementação do *pipeline* ETL em conformidade com as leis de proteção de dados.

7. Conclusão e Trabalhos Futuros

Neste artigo, foi apresentada a abordagem `Aether`, que propõe a instrumentação semiautomática de *pipelines* ETL por meio da injeção de *shadow pipelines*, com o objetivo de assegurar o cumprimento de três requisitos centrais da LGPD: finalidade de tratamento, existência de base legal e obtenção do consentimento por parte dos titulares dos dados.

Os resultados obtidos no estudo de viabilidade executado com a `Aether` demonstram que a abordagem é promissora. A `Aether` foi capaz de verificar, com sucesso, os critérios de finalidade e contrato em todas as variantes de pipelines analisadas. A verificação do consentimento, embora bem-sucedida na maioria dos casos, apresentou limitações em cenários nos quais os dados de identificação estavam parcial ou totalmente anonimizados, comprometendo a correspondência precisa com os registros de consentimento da base de metadados. Ainda assim, a abordagem apresentou um desempenho estável e o *overhead* introduzido pela injeção dos *shadow pipelines* foi considerado baixo, mesmo em cenários com grandes volumes.

Como trabalhos futuros, pretendemos aprimorar o mecanismo de correspondência de identificadores anonimizados, incorporando técnicas de desidentificação reversa controlada e aprendizado de máquina supervisionado para aumentar a acurácia na validação de consentimento. Também pretendemos expandir o escopo da `Aether` para lidar com outros princípios da LGPD, como a minimização de dados e o direito ao esquecimento. Por fim, vislumbramos a integração da ferramenta a plataformas de orquestração de dados, como Apache Airflow, com o intuito de facilitar sua adoção em ambientes de produção.

Referências

Barros, P., Monteiro, J. M., Brayner, A., and Machado, J. (2024). Incorporando os requisitos e as restrições da lgpd ao projeto de banco de dados. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 341–353, Porto Alegre, RS, Brasil. SBC.

Brasil (1988). Constituição da república federativa do brasil promulgada em 5 de outubro de 1988: atualizada até a emenda constitucional n. 48.

Cabral, R., Vasconcelos, V., Lins, F., Santos, G., Losse, M., Medeiros, A., Sousa, E., and Felix, M. (2023). Transparência e livre acesso: Uma avaliação da disponibilidade de informações sobre a lgpd em sites de tribunais de contas no brasil. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 240–247, Porto Alegre, RS, Brasil. SBC.

Cerqueira, D., de Mello, R., and Travassos, G. (2023). Um checklist para inspeção de privacidade e proteção de dados pessoais em artefatos de software. In *Anais do XXVI Congresso Ibero-Americano em Engenharia de Software*, pages 206–213, Porto Alegre, RS, Brasil. SBC.

GDPR.EU (2019). *GDPR Small Business Survey*. GDPR.EU Library - Project co-funded by the Horizon 2020 Program and EU. GDPR.eu, São Paulo :, 38 ed edition.

Grafberger, S., Groth, P., and Schelter, S. (2024). Towards interactively improving ml data preparation code via "shadow pipelines". In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, DEEM '24, page 7–11, New York, NY, USA. Association for Computing Machinery.

Gruschka, N., Mavroeidis, V., Vishi, K., and Jensen, M. (2018). Privacy issues and data protection in big data: a case study analysis under gdpr. In *IEEE International Conference on Big Data*, pages 5027–5033. IEEE.

Jalali, S. and Wohlin, C. (2012). Systematic literature studies: database searches vs. backward snowballing. In Runeson, P., Höst, M., Mendes, E., Andrews, A. A., and Harrison, R., editors, *ACM-IEEE International Symposium on Empirical Software Engineeringand Measurement*, pages 29–38. ACM.

Liu, H. (2019). Research on feasibility path of technology supervision and technology protection in big data environment. In *International Conference on Intelligent Transportation, Big Data & Smart City*, pages 293–296. IEEE.

Loureiro, J. and de Oliveira, D. (2022). Orbiter: um arcabouço para implantação automática de aplicações big data em arquiteturas serverless. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 379–384, Porto Alegre, RS, Brasil. SBC.

Magrani, E. (2019). *Entre dados e robôs: ética e privacidade na era da hiperconectividade*, volume 5. Arquipélago Editorial.

Marques, S., Lisboa, A., Érico Amaral, and Lampert, V. (2021). Pdagro: Uma proposta de protocolo para compliance à lgpd. In *Anais do XIII Congresso Brasileiro de Agroinformática*, pages 378–381, Porto Alegre, RS, Brasil. SBC.

Martins, A. D., Barros, P., Monteiro, J., and Machado, J. (2020). Lgpd: A formal concept analysis and its evaluation. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 259–264, Porto Alegre, RS, Brasil. SBC.

Nascimento, B. L. C. d. and Silva, E. M. d. (2023). Lei geral de proteção de dados (lgpd) e repositórios institucionais: reflexões e adequações. *Em Questão*, 29:127314.

Saraiva, J. and Soares, S. (2023). Adoption of the lgpd inventory in the user stories and bdd scenarios creation. In *Anais do XXXVII Simpósio Brasileiro de Engenharia de Software*, page 416–421, Porto Alegre, RS, Brasil. SBC.

Schwaitzer, L. (2020). Lgpd e acervos históricos: impactos e perspectivas. *Archeion Online, João Pessoa*, 8(2):36–51.

Secretaria de Saúde (2021). Relação de pessoas vacinadas - Covid 19 - Datasets - Portal de Dados Abertos da Cidade do Recife. Atualização semanal. Acesso em: 21 de Junho de 2024, 19:02 (UTC-03:00).

Solingen, R., Basili, V., Caldiera, G., and Rombach, D. (2002). *Goal Question Metric (GQM) Approach*.

Sousa, T., Coutinho, M., Coutinho, L., and Albuquerque, R. (2020). Lgpd: Levantamento de técnicas criptográficas e de anonimização para proteção de bases de dados. In *Anais do XX Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 55–68, Porto Alegre, RS, Brasil. SBC.

Vassiliadis, P. (2009). A survey of extract-transform-load technology. *Int. J. Data Warehous. Min.*, 5(3):1–27.

Vieira, M., de Oliveira, T., Cicco, L., de Oliveira, D., and Bedo, M. V. N. (2024). From tracking lineage to enhancing data quality and auditing: Adding provenance support to data warehouses with provetl. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 26th International Conference on Enterprise Information Systems, ICEIS 2024, Angers, France, April 28-30, 2024, Volume 1*, pages 313–320. SCITEPRESS.

Yang, Y., Meneghetti, N., Fehling, R., Liu, Z. H., and Kennedy, O. (2015). Lenses: an on-demand approach to etl. *Proc. VLDB Endow.*, 8(12):1578–1589.

Zaguir, N. A. (2024). *Desafios e habilitadores para a conformidade com a GDPR e LGPD: modelo de Governança da Informação sobre dados pessoais.* PhD thesis, Universidade de São Paulo.