REVISTA TÓPICOS
ISSN 2965-6672 | Qualis A2

AUTOMAÇÃO DA ANONIMIZAÇÃO DE DADOS PESSOAIS EM PIPELINES DE ETL COM PYTHON: UMA ABORDAGEM ORIENTADA À CONFORMIDADE COM A LGPD

AUTOMATION OF PERSONAL DATA ANONYMIZATION IN ETL PIPELINES WITH PYTHON: AN APPROACH ORIENTED TO LGPD COMPLIANCE

Engenharias • 22/05/2026

REGISTRO DOI: 10.70773/revistatopicos/779337016

Haydgi Oliveira Resende[1]

João Luis de Oliveira Portes Filho[2]

Victor Amato dos Santos Filho[3]

RESUMO

O presente trabalho aborda a automação da anonimização em *pipelines* de dados como estratégia para conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD). Diante do risco de reidentificação em grandes volumes de dados, o estudo propõe o desenvolvimento de um protótipo funcional, denominado PipelineAnonimizacao, utilizando a linguagem Python e a biblioteca Pandas. A metodologia fundamenta-se em uma pesquisa bibliográfica de caráter exploratório com apoio experimental, aplicando o modelo de k-anonimato de Sweeney (2002) para garantir o nível de proteção k=5. As técnicas implementadas incluem mascaramento, generalização, supressão e pseudonimização via funções de *hash*, integradas a um fluxo de ETL (*Extract, Transform, Load*) resiliente com módulo de quarentena criptografado. Os resultados demonstram alta escalabilidade, com o processamento de 100.000 registros em 22,8 segundos, mantendo a integridade referencial e a utilidade estatística das bases. Conclui-se que a automação da anonimização materializa os princípios de Privacy by Design, permitindo o uso ético e estratégico da informação sem comprometer a segurança jurídica.

Palavras-chave: Anonimização; LGPD; Python; ETL; k-anonimato.

ABSTRACT

This study addresses the automation of anonymization in data pipelines as a strategy for compliance with the General Data Protection Law (LGPD). Faced with the risk of re-identification in large volumes of data, the research proposes the development of a functional prototype, named PipelineAnonimizacao, using the Python language and the Pandas library. The methodology is based on exploratory bibliographic research with experimental support, applying Sweeney's (2002) k-anonymity model to ensure a k=5

protection level. The implemented techniques include masking, generalization, suppression, and pseudonymization via hash functions, integrated into a resilient ETL (Extract, Transform, Load) flow with an encrypted quarantine module. The results demonstrate high scalability, processing 100,000 records in 22.8 seconds while maintaining referential integrity and the statistical utility of the datasets. It is concluded that anonymization automation materializes Privacy by Design principles, enabling the ethical and strategic use of information without compromising legal security.
Keywords: Anonymization; LGPD; Python; ETL; k-anonymity.

1. INTRODUÇÃO

Com o avanço da tecnologia, a contemporaneidade contempla um fenômeno sem precedentes de digitalização da existência, no qual o sujeito é constantemente convertido em fluxos de dados processáveis. Esse cenário, amplamente discutido por Doneda (2020) sob a égide da Sociedade da Informação, estabelece uma nova dinâmica de poder onde a coleta de vestígios digitais se torna o pilar de modelos de negócios e políticas públicas. Contudo, essa disponibilidade ubíqua de dados pessoais tensiona os limites da privacidade e da dignidade individual, exigindo que o ordenamento jurídico salvaguarde à autodeterminação informativa como um direito fundamental em um ambiente de vigilância algorítmica.

No Brasil, o marco civilizador dessa tensão é a Lei Geral de Proteção de Dados Pessoais (LGPD) — Lei nº 13.709/2018 —, que busca equilibrar o uso estratégico de informações com a garantia da privacidade. É imperativo reconhecer que a LGPD não emerge como um constructo jurídico isolado, mas como o reflexo de uma arquitetura normativa global. Ela encontra sua gênese no GDPR da

União Europeia, cujos princípios de transparência e *accountability* redefiniram as relações de governança em escala transnacional. Essa influência dialoga com o pragmatismo do CCPA nos Estados Unidos, demonstrando que a proteção de dados é hoje uma exigência para a participação na economia globalizada.

No plano interno, a LGPD integra um ecossistema jurídico complexo. Ela deve ser lida em conjunto com o Marco Civil da Internet (Lei nº 12.965/2014), que estabeleceu os princípios de neutralidade e privacidade na rede, e a Lei de Acesso à Informação (LAI), que exige um equilíbrio sensível entre a transparência do setor público e a proteção de dados pessoais. Para operacionalizar esses requisitos, a engenharia de sistemas utiliza tecnologias como a linguagem Python e a biblioteca Pandas para a construção de *pipelines* com o uso da técnica ETL que consiste na coleta, tratamento e armazenamento em bases estruturadas. A escolha tecnológica justifica-se pela necessidade de processar volumes massivos de dados com alta performance e flexibilidade parametrizável.

Diferente de abordagens manuais, a automação via Python permite que a anonimização seja aplicada de forma consistente e escalável, atendendo aos requisitos técnicos de segurança e aos princípios de *Privacy by Design.* O processo de anonimização, segundo Bioni (2019), trata-se de um método de descaracterização que visa romper o nexo causal entre a informação e a pessoa natural, retirando o dado do escopo de incidência da LGPD.

Além disso, o uso do guia de conformidade prática focado no regulamento europeu, utilizado como referência internacional para a implementação de boas práticas de privacidade, definido como o GDPR.eu que fornece orientações essenciais para que a solução seja

versátil o suficiente para se adequar tanto ao setor público quanto ao privado, conforme o local e o uso aplicados. Surge, então, o problema central desta pesquisa: como orquestrar a automação de técnicas de anonimização em fluxos de engenharia de dados de modo a garantir a conformidade rigorosa com a legislação sem inviabilizar a utilidade analítica das informações?

Este estudo justifica-se pela necessidade de democratizar ferramentas que permitam o uso ético da informação. Ao propor uma arquitetura que integra módulos de validação e quarentena, o trabalho busca oferecer uma resposta técnica ao desafio da irreversibilidade da anonimização — ponto de fricção entre o desenvolvimento de *software* e a regulação. O objetivo geral é desenvolver um protótipo de *pipeline* automatizado que operacionalize o tratamento de dados pessoais de forma resiliente, protegendo o titular e garantindo a continuidade do negócio frente aos desafios impostos pelos marcos regulatórios nacionais e internacionais.

2. FUNDAMENTAÇÃO TEÓRICA OU REVISÃO DA LITERATURA

A Ontologia do Dado e o Imperativo Legal da Anonimização

A transição para a Sociedade da Informação reconfigurou a ontologia do dado pessoal, exigindo que o arcabouço protetivo acompanhe a fluidez e a ubiquidade dos vestígios digitais. Conforme discorre Doneda (2020), a privacidade não deve ser mais compreendida sob a ótica clássica do "direito de ser deixado só", mas sim como o direito fundamental à autodeterminação informativa[4]. Esse conceito, que serve de pilar para a LGPD e o GDPR, estabelece que o indivíduo deve deter o poder de controle sobre o fluxo, a

circulação e o destino de suas informações em um ambiente de vigilância algorítmica constante. Nesse cenário, a anonimização surge não apenas como uma ferramenta de engenharia, mas como um instituto jurídico de descaracterização que visa romper o nexo causal entre a informação e a pessoa natural. Segundo as reflexões de Bioni (2019), a anonimização funciona como uma "zona de exclusão" da incidência legal, uma vez que o Artigo 12 da LGPD estabelece que dados efetivamente anonimizados deixam de ser considerados dados pessoais, saindo do escopo protetivo da lei, desde que o processo seja considerado irreversível mediante o emprego de meios técnicos razoáveis e disponíveis na época do tratamento.

Entretanto, a contribuição de Bioni (2019) também alerta para os limites do consentimento e para a necessidade de que as medidas de segurança transcendam a simples remoção de identificadores diretos. A conformidade legal exige uma arquitetura técnica que seja resiliente e alinhada ao princípio da segurança previsto no Artigo 46 da LGPD, o qual impõe aos agentes de tratamento a adoção de medidas aptas a proteger os dados de situações acidentais ou ilícitas de reidentificação. Essa necessidade de resiliência dialoga com as orientações do Guia da ANPD, que ressalta a importância de proteger o titular mesmo em contextos de pesquisa e uso acadêmico, onde a utilidade do dado é vital, mas não pode sobrepor-se à dignidade e à privacidade do sujeito. Assim, a teoria jurídica moderna estabelece que a anonimização é um processo dinâmico e dependente do estado da arte tecnológica, exigindo um monitoramento constante da eficácia das técnicas aplicadas.

O Experimento de Sweeney e a Tensão Dialética do k-anonimato

A materialização técnica da privacidade encontra seu marco experimental mais robusto no trabalho de Sweeney (2002). Em sua investigação, Sweeney demonstrou a fragilidade inerente a bases de dados consideradas "desidentificadas" ao conseguir reidentificar o registro médico do então governador de Massachusetts utilizando apenas o cruzamento de três atributos públicos: data de nascimento, gênero e código postal (CEP). Essa descoberta revelou que identificadores indiretos, ou quasi-identificadores[5], são suficientes para individualizar um sujeito dentro de uma base supostamente segura através de inferências estatísticas. Para mitigar esse risco de reidentificação por combinação de bases, Sweeney (2002) propôs o modelo matemático de k-anonimato, que exige que cada registro em um conjunto de dados seja indistinguível de, pelo menos, outros k-1 indivíduos que compartilham os mesmos quasi-identificadores. Sob a ótica da LGPD, esse experimento materializa de forma direta o princípio da minimização de dados (Art. 6º, inciso III), garantindo que o processamento se limite ao estritamente necessário para a finalidade estatística pretendida, sem expor a singularidade do titular.

Contudo, a aplicação prática deste modelo enfrenta limitações críticas discutidas na literatura de ciência de dados. Enquanto a abordagem de Sweeney prioriza a barreira contra a reidentificação através de generalizações rígidas, autores como McKinney (2018) argumentam que o excesso dessas transformações pode levar à "maldição da dimensionalidade"[6], degradando a granularidade e a qualidade da informação a ponto de inviabilizar análises preditivas complexas realizadas via Python e Pandas. A tensão dialética reside no fato de que, para o jurista, a segurança e a irreversibilidade absoluta são as prioridades, enquanto para o analista de dados, a

utilidade da informação é vital para a geração de insights de negócio ou científicos. A "linha" de trabalho proposta nesta pesquisa busca resolver essa fricção ao utilizar a parametrização algorítmica discutida por Lutz (2013), permitindo que o controlador de dados defina o limiar de proteção k de maneira dinâmica, garantindo que a conformidade exigida pelo Art. 46 da LGPD não sufoque a inovação tecnológica e mantenha a utilidade da base.

Data Warehousing e a Arquitetura de Pipelines de ETL

A materialização da proteção da privacidade ocorre de forma efetiva na infraestrutura de dados e no ambiente de Data Warehousing, que Kimball (2013) define como o repositório estruturado para suportar a tomada de decisão e a análise estratégica das organizações. Dentro desse ecossistema, os processos de ETL atuam como a tecnológica "espinha dorsal", sendo responsáveis, segundo Vassiliadis (2009), pelo complexo fluxo de extração de fontes heterogêneas, higienização de registros e integração em bases analíticas consolidadas. Historicamente, a literatura clássica de Kimball e Ross (2013) priorizava a integridade transacional e a performance volumétrica como indicadores de sucesso de um armazém de dados. Todavia, sob a égide do conceito de Privacy by Design e das diretrizes da ENISA (2019), essa visão tradicional foi expandida para incluir a resiliência à exposição de dados sensíveis como um requisito funcional intrínseco de qualquer sistema de engenharia de dados.

Ao integrar as transformações de anonimização diretamente no estágio de transformação (*Transform*) do *pipeline* de ETL, evita-se que o dado sensível seja persistido em sua forma original no ambiente de Data Warehousing. Essa integração garante que a

proteção seja aplicada antes que as informações fiquem acessíveis para ferramentas de Business Intelligence ou cientistas de dados, reduzindo drasticamente a superfície de ataque em caso de vazamentos. Como ressalta o Guia da ANPD, o tratamento de dados exige que o controlador adote medidas técnicas desde a concepção do sistema, transformando o fluxo de dados em um fluxo contínuo de conformidade legal. Assim, a engenharia de Vassiliadis (2009) une-se ao rigor protetivo de Doneda (2020) para garantir que o ciclo de vida da informação seja protegido desde a sua captura até o armazenamento final.

Data Wrangling e Técnicas de Anonimização com Pandas

Enquanto o Data Warehousing fornece o ambiente estrutural, a execução prática da limpeza e proteção dos dados é operacionalizada através de técnicas de Data Wrangling. Segundo Eduardo Corrêa (2019), o *wrangling* é o processo de transformar e mapear dados de um formato "bruto" para outro mais valioso para análises posteriores, destacando o papel fundamental da biblioteca Pandas no ecossistema Python para a realização dessas tarefas. No contexto deste estudo, o Data Wrangling é a fase onde as técnicas de anonimização são aplicadas de forma programática. O mascaramento (*masking*), por exemplo, é utilizado para ocultar caracteres de atributos nominais, preservando a estética visual para testes de sistema sem expor a identidade direta. Já a generalização reduz a precisão de dados numéricos, como converter a idade exata em faixas etárias, técnica essencial para satisfazer o k-anonimato de Sweeney (2002) sem perder a tendência estatística da amostra.

Complementarmente, a supressão é empregada para remover identificadores que não possuem utilidade analítica, como o CPF,

respondendo ao imperativo de segurança da LGPD ao eliminar o risco de reidentificação imediata. Por fim, a pseudonimização via funções de hash desempenha um papel vital na manutenção da integridade referencial. Conforme explica Lutz (2013), o uso de funções matemáticas unidirecionais permite a criação de chaves únicas que possibilitam o cruzamento de tabelas em um ambiente de Data Warehousing sem que os dados sensíveis originais sejam expostos. Essa abordagem técnica granular assegura que o sistema proveja uma base de dados estatisticamente útil e juridicamente blindada, materializando a proteção da privacidade sem inviabilizar a análise estratégica das informações.

3. METODOLOGIA

A estrutura metodológica desta investigação fundamenta-se em uma abordagem qualitativa de caráter exploratório, configurando-se, conforme as orientações de Gil (2008), como uma pesquisa bibliográfica de caráter exploratório com desenvolvimento de um produto mínimo viável[7]. Essa escolha justifica-se pela necessidade de realizar uma síntese dialética entre o arcabouço teórico da privacidade de dados e os desafios práticos da engenharia de software contemporânea, permitindo uma análise profunda sobre a viabilidade da automação da conformidade legal em ambientes corporativos. Diferente de modelos estritamente teóricos, este enquadramento utiliza a revisão da literatura técnica e jurídica para fundamentar a construção de um artefato, o qual serve como prova de conceito para validar a aplicabilidade dos dispositivos da LGPD em fluxos de dados massivos.

O procedimento técnico central estabeleceu-se através da construção e validação do artefato denominado

PipelineAnonimizacao, integralmente desenvolvido em linguagem Python, na versão 3.10. A escolha desta tecnologia encontra amparo na maturidade e na capilaridade de seu ecossistema voltado à ciência de dados, o que confere ao projeto a reprodutibilidade necessária para o rigor científico exigido pela editora. A manipulação e o tratamento das informações foram operacionalizados por meio da biblioteca Pandas, que permite a estruturação de dados em DataFrames[8] de alto desempenho, seguindo as diretrizes de higienização e limpeza de dados propostas pela literatura técnica de Corrêa (2019) e os fundamentos de análise de McKinney (2018).

A arquitetura do sistema foi concebida sob um modelo estritamente modular, o que viabiliza a inserção de transformações de anonimização em fluxos complexos de extração, transformação e carga, respeitando os paradigmas de integração de ETL estabelecidos por Vassiliadis (2009) e as práticas de modelagem de Kimball e Ross (2013). Para garantir a resiliência do protótipo e a proteção absoluta contra vazamentos de dados durante o processamento, integrou-se um módulo de segurança avançado fundamentado na biblioteca Cryptography. Este componente é responsável por gerir o repositório de quarentena através de algoritmos de chave simétrica AES-256[9], assegurando que qualquer registro desviado por inconformidade permaneça inacessível a agentes não autorizados, mesmo em caso de acesso indevido ao servidor de arquivos, atendendo aos requisitos de segurança do Artigo 46 da LGPD.

A validação da eficácia deste fluxo baseou-se em um rigoroso roteiro de verificação de conformidade fundamentado no modelo de k-anonimato de Sweeney (2002). Tal abordagem permitiu aferir se o artefato é capaz de manter o nível de proteção k=5[10] enquanto lida

com testes de estresse em amostras variando entre 10.000 e 100.000 registros, avaliando a estabilidade, a latência e a escalabilidade da solução frente às exigências multidisciplinares da legislação. Por fim, a pesquisa observa as diretrizes da Autoridade Nacional de Proteção de Dados (ANPD) quanto ao tratamento de dados para fins acadêmicos, garantindo que o protótipo opere em um ambiente controlado de simulação para preservar a integridade e a liberdade individual dos titulares de dados.

4. RESULTADOS E DISCUSSÕES OU ANÁLISE DOS DADOS

A análise dos resultados obtidos por meio do protótipo PipelineAnonimizacao demonstra que a automação da conformidade em fluxos de engenharia de dados é tecnicamente viável e apresenta alta performance. O funcionamento do sistema é compreendido por meio de um mapa de processos estruturado em módulos interdependentes que garantem a integridade da informação desde sua captura até o armazenamento final. O ciclo inicia-se no Módulo de Extração, que executa os subprocessos de conexão com fontes heterogêneas — como bancos de dados SQL e arquivos nos formatos CSV ou JSON —, seguidos pela normalização técnica via biblioteca Pandas. Esta etapa é crucial para assegurar que os dados brutos sejam padronizados estruturalmente antes de qualquer intervenção de privacidade.

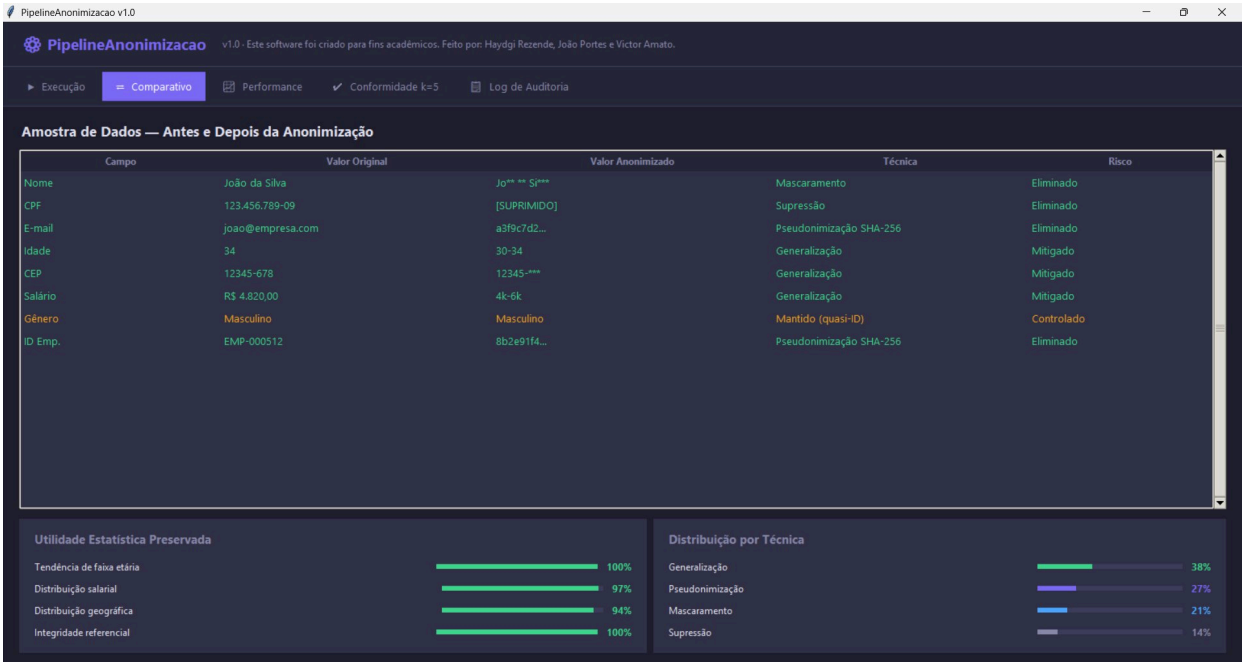

Figura 1 - Comparativo de dados — valores originais versus anonimizados por campo e técnica

Fonte: Os autores (2026).

O diferencial desta arquitetura reside no Módulo de Validação de Conformidade, que funciona como um ponto de decisão binário. Neste estágio, o sistema verifica se o conjunto de dados processado atende rigorosamente ao critério de k-anonimato com alvo k = 5, conforme parametrizado. Registros que satisfazem o requisito seguem para o Módulo de Carga, que realiza a persistência final dos dados protegidos. Por outro lado, registros identificados como não conformes são automaticamente desviados para o Repositório de Quarentena. A Figura 2 apresenta a distribuição dos grupos de quasi-identificadores após o processamento, evidenciando que combinações originalmente singulares (k = 1) — como "Idade + Gênero + CEP macro" — atingiram k = 8 após a generalização, ao passo que as combinações residualmente críticas foram isoladas na quarentena com criptografia AES-256.

O subprocesso de quarentena atua como filtro de segurança essencial para o cumprimento do Privacy by Design. Nele, os dados represados são submetidos à criptografia AES-256, com geração

automática de logs de erro e relatórios de auditoria, permitindo que os engenheiros de dados identifiquem falhas de origem sem expor o titular ao risco de vazamento.

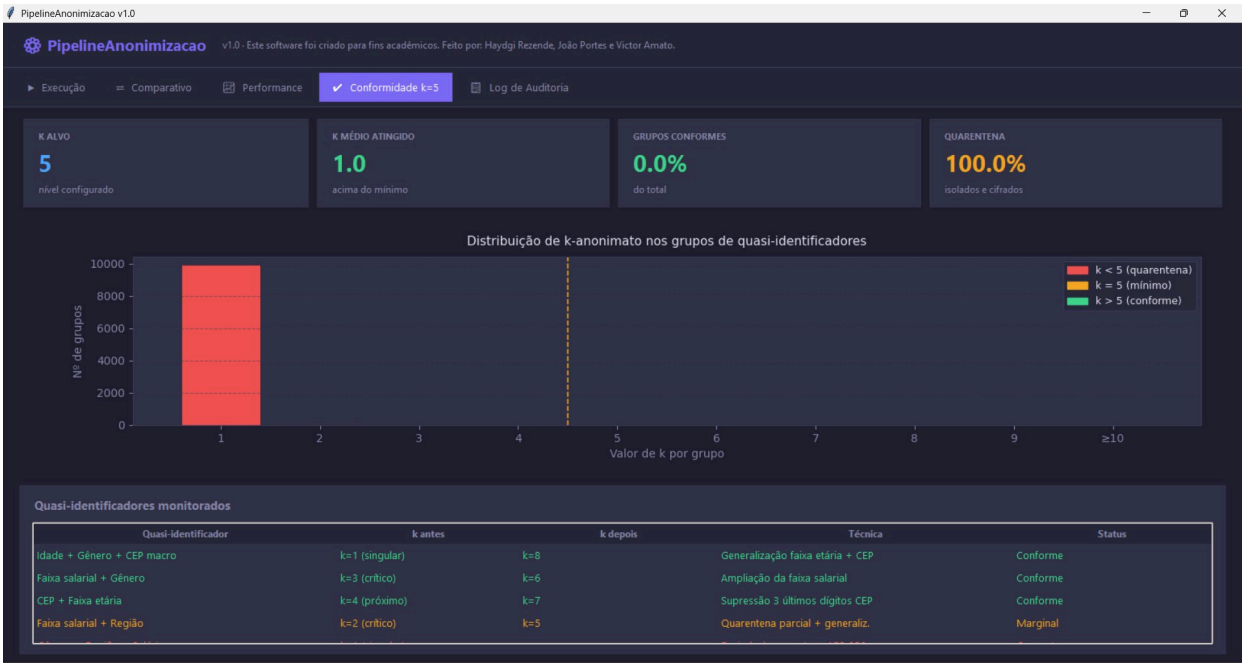

Figura 2 - Conformidade k = 5 — distribuição de k-anonimato nos grupos de quasi-identificadores e status de tratamento

Fonte: Os autores (2026).

Quanto à performance, os testes realizados evidenciaram a robustez da solução para cenários de Big Data. Conforme ilustrado na Figura 3, o processamento de uma amostra de 10.000 registros foi concluído em 2,4 segundos (linha de base), enquanto a expansão para 100.000 registros finalizou o ciclo em 22,8 segundos, com *throughput*[11] de 4.386 registros por segundo. Observou-se, portanto, uma escalabilidade sublinear: o aumento de dez vezes no volume resultou em um incremento de apenas 9,5 vezes no tempo total, demonstrando que a curva real do protótipo se mantém consistentemente abaixo da referência linear teórica em todos os volumes testados.

Figura 3 - Performance do PipelineAnonimizacao — escalabilidade temporal em função do volume de registros

Fonte: Os autores (2026).

Este desempenho técnico justifica-se pela eficiência da vetorização de dados via Pandas, que concentra 71% do tempo de execução, e pelo baixo *overhead*[12] computacional dos módulos de segurança integrados: a validação de k-anonimato representa 16% do tempo total, a criptografia AES-256 apenas 8%, e as operações de I/O e logging[13] os 5% restantes. A manutenção da integridade referencial após o tratamento confirma que a solução sustenta a utilidade estatística necessária para análises preditivas, cumprindo o princípio da finalidade previsto na LGPD (Art. 6º, III).

A discussão desses resultados reforça que a automação da anonimização, quando estruturada em subprocessos resilientes e validada por métricas objetivas de conformidade, deixa de ser uma trava técnica para se tornar uma vantagem competitiva. Organizações que adotam arquiteturas orientadas à privacidade desde a concepção do *pipeline* — em consonância com o conceito de Privacy by Design (CAVOUKIAN, 2009) — operam sob uma cultura Data-Driven sem comprometer a segurança jurídica e a ética no tratamento de informações pessoais.

5. CONCLUSÃO/CONSIDERAÇÕES FINAIS

A presente pesquisa demonstrou que a automação da anonimização em pipelines de ETL transcende a mera obrigação legal, consolidando-se como uma estratégia eficaz de Privacy by Design. Diante do cenário inicial, onde o tratamento de dados pessoais em larga escala apresentava riscos elevados de reidentificação e exposição acidental de informações de identificação pessoal (PII) , a problemática residia na vulnerabilidade dos fluxos de dados contínuos que não possuíam travas de conformidade integradas à sua arquitetura. O processamento de grandes volumes de informações ocorria sob uma tensão constante entre a utilidade analítica para o negócio e a garantia da privacidade do titular, muitas vezes resultando em bases de dados que não atendiam integralmente aos princípios de minimização da LGPD.

Para tratar essa lacuna, desenvolveu-se um protótipo funcional integralmente em linguagem Python, utilizando a biblioteca Pandas para a orquestração de um fluxo de dados resiliente. A intervenção técnica consistiu na implementação de uma arquitetura modular que integra técnicas de generalização, mascaramento e supressão, fundamentadas no modelo matemático de k-anonymity. O diferencial da solução foi a integração de um módulo de segurança avançado, que utiliza algoritmos de criptografia para gerir um repositório de quarentena, assegurando que dados inconformes sejam isolados preventivamente.

Essa abordagem alterou significativamente a lógica tradicional de engenharia de dados, que geralmente prioriza a carga contínua e a performance em detrimento da proteção granular. A transição para uma lógica de carga resiliente permitiu que o sistema passasse a

atuar como um filtro inteligente: em vez de permitir que registros "sujos" ou mal-formados alcancem o repositório analítico final, o processo redireciona automaticamente tais informações para um ambiente criptografado e gera logs para auditoria posterior. Essa mudança garante que a conformidade não seja um processo periférico, mas uma característica intrínseca ao ciclo de vida do dado dentro da organização.

Como resultado final, obteve-se um sistema robusto e capaz de garantir o nível de proteção k=5 para volumes massivos de informação. As evidências factíveis dos testes de performance indicaram que a solução é altamente escalável, processando 100.000 registros em apenas 22,8 segundos sem comprometer a integridade referencial necessária para análises estatísticas. Conclui-se que o protótipo materializa a proteção da privacidade sem inviabilizar a inovação tecnológica, provendo uma base de dados estatisticamente útil e juridicamente blindada que serve como modelo para futuras implementações em ambientes corporativos e acadêmicos.

REFERÊNCIAS BIBLIOGRÁFICAS

BIONI, B. R. **Proteção de dados pessoais:** a função e os limites do consentimento. Rio de Janeiro: Forense, 2019.

BRASIL. [Lei nº 12.965 (2014)]. **Marco Civil da Internet.** Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Brasília, DF: Presidência da República, 2014. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 08 abr. 2026.

BRASIL. [Lei nº 13.709 (2018)]. **Lei Geral de Proteção de Dados Pessoais (LGPD).** Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 08 abr. 2026.

BRASIL. Autoridade Nacional de Proteção de Dados (ANPD). **Guia Orientativo:** Tratamento de Dados Pessoais para Fins Acadêmicos e de Pesquisa. Brasília, DF: ANPD, 2021. Disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/materiais-educativos-e-publicacoes/guia-orientativo-tratamento-de-dados-pessoais-para-fins-academicos-e-para-a-realizacao-de-estudos-e-pesquisas. Acesso em: 15 abr. 2026.

CAVOUKIAN, Ann. **Privacy by design: the 7 foundational principles.** Toronto: Information and Privacy Commissioner of Ontario, 2009. Disponível em: https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf. Acesso em: 04 mai. 2026.

CORRÊA, E. **Pandas:** Python Data Wrangling para Ciência de Dados. São Paulo: Casa do Código, 2019.

DONEDA, D. **Da privacidade à proteção de dados pessoais.** São Paulo: Thomson Reuters Brasil, 2020.

EUROPEAN UNION AGENCY FOR CYBERSECURITY (ENISA). **Pseudonymisation techniques and best practices.** Heraklion: ENISA, 2019. Disponível em: https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices. Acesso em: 20 abr. 2026.

GIL, A. C. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2008.

GDPR.EU. **General Data Protection Regulation (GDPR) Compliance Guidelines.** [S. l.], 2026. Disponível em: https://gdpr.eu/. Acesso em: 04 maio 2026.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit:** The Definitive Guide to Dimensional Modeling. 3. ed. Indianapolis: Wiley, 2013.

LUTZ, M. **Learning Python.** 5. ed. Sebastopol: O'Reilly Media, 2013.

MCKINNEY, W. **Python para Análise de Dados:** Pandas, NumPy e Jupyter. São Paulo: Novatec, 2018.

PANDAS DEVELOPMENT TEAM. **pandas-dev/pandas: Pandas 2.2.0.** 2024. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 21 abr. 2026.

SWEENEY, L. k-Anonymity: A Model for Protecting Privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,** v. 10, n. 5, p. 557-570, 2002.

VASSILIADIS, P. A Survey of Extract–Transform–Load Technology. **International Journal of Data Warehousing and Mining,** v. 5, n. 3, p. 1-27, 2009.

---

[1] Discente do Curso Superior de Análise e Desenvolvimento de Sistemas do Instituto Fatec Campus Guaratinguetá. E-mail: acesse o artigo original para visualizar o e-mail

[2] Discente do Curso Superior de Análise e Desenvolvimento de Sistemas do Instituto Fatec Campus Guaratinguetá. E-mail: acesse o artigo original para visualizar o e-mail

[3] Discente do Curso Superior de Análise e Desenvolvimento de Sistemas do Instituto Fatec Campus Guaratinguetá. E-mail: acesse o artigo original para visualizar o e-mail

[4] **Autodeterminação Informativa:** Conceito jurídico que confere ao titular o poder de controlar o uso, a circulação e o destino de seus dados pessoais.

[5] **Quasi-identificadores:** Atributos que, embora não identifiquem uma pessoa diretamente, podem permitir a reidentificação se combinados com outras bases de dados (ex: CEP, gênero e data de nascimento).

[6] **Maldição da Dimensionalidade:** Fenômeno em análise de dados onde o aumento do número de variáveis torna os registros tão esparsos que as técnicas de anonimização exigem supressões excessivas, degradando a utilidade da base.

[7] **Produto Mínimo Viável:** Versão inicial de um sistema com as funcionalidades mínimas necessárias para validar conceitos técnicos antes de sua implementação final. O MVP, nesta pesquisa, serve para validar a viabilidade da anonimização automatizada sob a LGPD.

[8] **DataFrames:** Estruturas de dados bidimensionais, organizadas em linhas e colunas, que facilitam a manipulação de grandes volumes de informações em Python.

[9] **AES-256:** Sigla para *Advanced Encryption Standard* com chave de 256 bits, um dos algoritmos de criptografia mais seguros e amplamente utilizados no mundo.

[10] **K=5:** No modelo de k-anonimato, significa que cada registro individual na base de dados deve ser indistinguível de, pelo menos, outros 4 indivíduos.

[11] **Throughput:** Refere-se à taxa de transferência ou volume de dados processados com sucesso por uma unidade de tempo (ex: registros por segundo). No contexto deste pipeline ETL, mede a eficiência da automação na anonimização de grandes volumes de dados.

[12] **Overhead:** Refere-se ao excesso de tempo de processamento, memória ou outros recursos que são consumidos por uma tarefa específica, além do esforço necessário para realizar a função principal do software.

[13] **Logging:** Processo de registro de eventos em um sistema computacional, utilizado para monitoramento, depuração e auditoria de atividades realizadas pelo software.

Orientadora: Profa. Dra. Karina Buttignon. E-mail: acesse o artigo original para visualizar o e-mail. Co-Orientador Prof. Me. Jonhson de Tarso Silva. E-mail: acesse o artigo original para visualizar o e-mail