Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Tecnologia da Informação
Mestrado Profissional em Tecnologia da Informação

Proposta de uma arquitetura para (pseudo)anonimização multinível de dados em saúde

Pedro Henrique Rodrigues Emerick

Natal-RN
Fevereiro de 2023

Pedro Henrique Rodrigues Emerick

Proposta de uma arquitetura para (pseudo)anonimização multinível de dados em saúde

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Tecnologia da Informação da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Mestre em Tecnologia da Informação.

Orientador

Dr. Roger Kreutz Immich

Co-orientador

Dr. Silvio Costa Sampaio

Universidade Federal do Rio Grande do Norte – UFRN
Instituto Metrópole Digital – IMD
Programa de Pós-Graduação em Tecnologia da Informação – PPGTI

Natal-RN

Fevereiro de 2023

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Emerick, Pedro Henrique Rodrigues.
Proposta de uma arquitetura para (pseudo)anonimização multinível de dados em saúde / Pedro Henrique Rodrigues Emerick. - 2023.
114 f.: il.

Dissertação (mestrado) - Universidade Federal do Rio Grande do Norte, Instituto Metrópole Digital, Programa de Pós-Graduação em Tecnologia da Informação, Natal, RN, 2023.
Orientador: Prof. Dr. Roger Kreutz Immich.
Coorientador: Dr. Silvio Costa Sampaio.

1. Privacidade de dados - Dissertação. 2. Dados pessoais - Dissertação. 3. Anonimização - Dissertação. 4. Pseudonimização - Dissertação. 5. LGPD - Dissertação. I. Immich, Roger Kreutz. II. Sampaio, Silvio Costa. III. Título.

RN/UF/BCZM CDU 004.056

Elaborado por Ana Cristina Cavalcanti Tinoco - CRB-15/262

Dedico este trabalho à Deus e aos meus familiares, amigos e professores que de alguma forma contribuíram com meu progresso.

Agradecimentos

Agradeço a Deus, por Seu inexplicável amor, ensinamentos e bençãos derramadas em minha vida.

Agradeço a minha noiva, Izabela, por seu apoio, incentivo, compreensão e amor, que não me deixaram desanimar e desistir.

Agradeço aos meus pais, Dercy e Fernanda, por todo incentivo e educação, me ensinando que o conhecimento é infinito.

Agradeço aos meus irmãos, Guilherme, Raphael, Samuel e Rebeca, por toda alegria e paz que vocês me trazem.

Agradeço aos meus familiares e amigos, que me incentivaram e demonstraram seu apoio e torcida de alguma forma. Em especial Elzony, Edite, Francielle, Gabriel, Lucas e Rodrigo.

Agradeço ao meu amigo e co-orientador, Silvio, por seus conselhos, incentivos e colaborações, e por sempre acreditar no meu potencial quando nem eu mesmo acreditava.

Agradeço ao meu orientador, Roger, por sua orientação, incentivos e colaboração com o desenvolvimento do trabalho.

Agradeço ao PPGTI, pela oportunidade e por todo aprendizado e crescimento adquirido.

Agradeço ao meu psicólogo, Gustavo, por seu trabalho e profissionalismo que foram de extrema importância durante todo o desenvolvimento do trabalho.

Por fim, agradeço a mim, por aproveitar as boas oportunidades da vida e por não ter desistido, mesmo quando parecia ser a melhor opção.

"Tudo neste mundo tem o seu tempo; cada coisa tem a sua ocasião."

Eclesiastes 3,1

Proposta de uma arquitetura para (pseudo)anonimização multinível de dados em saúde

Autor: Pedro Henrique Rodrigues Emerick
Orientador: Dr. Roger Kreutz Immich
Co-orientador: Dr. Silvio Costa Sampaio

Resumo

Nas últimas décadas, a evolução tecnológica trouxe inúmeros avanços, mas também permitiu a coleta, o processamento e o armazenamento intensivos de dados pessoais. São muitas as evidências, principalmente de revelações sobre as operações e fuga de dados de grandes empresas que têm os dados como seu maior ativo, a exemplo do Facebook, Google, Amazon e Ubber. Diante desta constatação, nota-se um preocupação crescente com a utilização destes dados, evidenciada pela profusão de legislações mundo afora, que visam proteger a privacidade dos indivíduos. As diversas legislações apontam para a necessidade de implementação de processos e técnicas que garantam a privacidade dos dados, dentre as quais está a (pseudo)anonimização dos dados. É neste contexto e buscando contribuir para a proteção da privacidade, que, neste trabalho, é proposta uma arquitetura para a (pseudo)anonimização multinível de dados em saúde. Multinível, pois os dados são pseudonimizados em dois níveis diferentes, um local e um global, garantindo assim que dados de múltiplos provedores de dados possam ser relacionados, ainda que (pseudo)anonimizados. O foco na área da saúde é, por um lado, uma aplicação desafiadora, dada a sensibilidade dos dados. A arquitetura proposta neste trabalho foi implementada como uma prova de conceito e avaliada a partir de um conjunto de testes. Os resultados dos testes sugerem que a arquitetura possibilita uma correta anonimização na fonte, uma ligação segura dos dados (pseudo)anonimizados em múltiplas fontes e ainda permite a reidentificação para casos que envolvam a segurança dos indivíduos envolvidos.

Palavras-chave: privacidade de dados, dados pessoais, anonimização, pseudonimização, LGPD.

Proposal of an architecture for multilevel (pseudo)anonymization of healthcare data

Author: Pedro Henrique Rodrigues Emerick
Supervisor: Dr. Roger Kreutz Immich
Co-supervisor: Dr. Silvio Costa Sampaio

Abstract

In recent decades, technological evolution has brought numerous advances allowing intensive collection, processing, and storage of personal data. There is much evidence, mainly revelations, about the operations and data breaches of large companies with data as their most significant asset, such as Facebook, Google, Amazon, and Uber. Due to this finding, there is a growing concern about using these data, evidenced by the profusion of laws worldwide that aim to protect individuals' privacy. The various legislations point to the need to implement processes and techniques that guarantee data privacy, among which is the (pseudo)anonymization of data. It is in this context and seeking to contribute to the protection of privacy that, in this work, an architecture is proposed for the multilevel (pseudo)anonymization of health data. Multilevel, as data is pseudonymized at two levels, one local and one global, thus ensuring that data from multiple providers can be related yet (pseudo)anonymized. The focus on the health area is, on the one hand, a challenging application, given the sensitivity of the data. The architecture proposed in this work was implemented as a proof of concept and evaluated from a set of tests. Test results suggest that the architecture enables correct anonymization at the source, secure linking of (pseudo)anonymized data across multiple sources, and even allows reidentification for cases involving the security of the individuals involved.

Keywords: data privacy, personal data, anonymization, pseudonymization, LGPD.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

RGPD – Regulamento Geral de Proteção de Dados

GDPR – General Data Protection Regulation

LGPD – Lei Geral de Proteção de Dados

ANPD – Autoridade Nacional de Proteção de Dados

HIV – Human Immunodeficiency Virus

HIPPA – Health Insurance Portability and Accountability Act

UFRN – Universidade Federal do Rio Grande do Norte

IMD – Instituto Metrópole Digital

IMT – Instituto de Medicina Tropical

CCHLA – Centro de Ciências Humanas, Letras e Artes

CCS – Centro de Ciências da Saúde

CB – Centro de Biociências

EAJ – Escola Agrícola de Jundiaí

FACISA – Faculdade de Ciências da Saúde do Trairi

EMCM – Escola Multicampi de Ciências Médicas

PEP – Policy Enforcement Point

S-RES – Sistema de Registro Eletrônico de Saúde

PoC – Prova de Conceito

CPF – Cadastro de Pessoa Física

CCPA – California Consumer Privacy Act

PIPEDA – Personal Information Protection and Electronic Documents Act

LGPDPPSO – Ley General De Protección De Datos Personales En Posesión De Sujetos

Obligado

LFPDPPP – Ley Federal De Protección De Datos Personales En Posesión De Los Particulares

CEP – Código de Endereçamento Postal

SDV – Synthetic Data Vault

EMD – Earth Mover's Distance

CSV – Comma-separated values

ARGUS – AntiRe-identification General Utility System

SDC – Statistical Disclosure Control

SGBD – Sistema de Gerenciamento de Banco de Dados

API – Application Programming Interface

REST – Representational State Transfer

TIAMAT – Tool for Interactive Analysis of Microdata Anonymization Techniques

NLM – National Library of Medicine

PII – Personally Identifiable Information

SECRETA – System for Evaluating and Comparing RElational and Transaction Anonymization algorithms

CAT – Cornell Anonymization Toolkit

GUI – Graphical User Interface

HB – Honest Broker

TTP – Trusted Third Party

OpenAIRE – Open Access Infrastructure for Research in Europe

AaaS – Anonymization as a Service

OsloMet – Oslo Metropolitan University

NAV – Norwegian Labour and Welfare Administration

ARXaaS – ARX as a Service

DP – Data Provider

DL – Data Lake

LB – Local Broker

UI – User Interface

Listagens

Sumário

1 Introdução

Embora seja comum a utilização de expressões como “Informação é dinheiro”, “A economia dos dados”, “A era da informação”, para evidenciar o valor e a importância dos dados, ainda o fazemos com algum distanciamento, quase que impersonalizando os dados. Contrariamente, em (Gates e Matthews 2014), num artigo intitulado “Dado é a nova moeda” (tradução livre do inglês “*Data is the new currency*”), os autores traçam um perfil evolutivo da utilização de dados pessoais, inicialmente relembrando os primórdios da computação, onde empresas de serviços públicos compilavam dados de clientes e, posteriormente, vendiam acesso a outras organizações para iniciativas de marketing e vendas. A coleta de dados acelerou em escala exponencial, resultando em novas áreas de negócios e hoje os agregadores ou corretores de dados variam desde o óbvio, como a Google, a Amazon ou o Facebook, até o mais oculto, como a sua companhia de seguro de saúde ou seu banco. Um aspecto preocupante é a constante reivindicação dos direitos de propriedade e exploração sobre os dados de outras pessoas.

No campo legal, a preocupação com o abuso na coleta e utilização de dados já apresenta um histórico de anos, com origens nas primeiras leis europeias nas décadas de 1970 e 1980, como a iniciativa da Alemanha, no estado de Hesse em 1970, que preocupada com o avanço da computação e da indústria nos países mais desenvolvidos, teria impulsionado o estado alemão a criar normas para regular a privacidade no país, sendo então considerada a primeira legislação voltada à proteção de dados (Stepanova e Jechel 2021). Ao longo dos anos seguintes, várias iniciativas semelhantes foram implantadas em outros países europeus isoladamente, culminando no Regulamento Geral de Proteção de Dados (RGPD) (do inglês, *General Data Protection Regulation (GDPR)*), em vigor desde 2018. O RGPD representa um grande avanço na proteção de dados pessoais e se impõe como um regulamento comum aos signatários da União Europeia, que recepcionaram este regulamento em suas legislações nacionais. No Brasil, embora a preocupação com a privacidade do indivíduo esteja expressa na Constituição Federal do Brasil de 1988 e seja também manifestada no Código de Defesa do Consumidor de 1993, ao disciplinar a relação entre

empresas e consumidores, apenas em 2013, com o Marco Civil da Internet, que temas como privacidade, confidencialidade e segurança das informações e dados pessoais prestados ou coletados, inclusive por meio eletrônico, passaram a ser diretamente abordados numa visão legislativa. Mas, somente em 2018, foi aprovada a Lei Geral de Proteção de Dados (LGPD), com início de sua vigência a iniciar em agosto de 2020. A mesma medida provisória que definiu a vigência da LGPD também previa a criação da Autoridade Nacional de Proteção de Dados (ANPD) como órgão de fiscalização. A LGPD foi fortemente inspirada na experiência da RGPD.

Toda essa movimentação legislativa é reflexo de uma evidente mudança dos usuários, ou melhor, das pessoas, que passaram a ter uma maior consciência de que a maioria destes dados dizem respeito a elas e suas vidas, passando então a exigir uma maior proteção, transparência e controle sobre a coleta e utilização destes dados. Esta discussão se intensificou na última década, com a disseminação no uso de tecnologias de coleta, transmissão e processamento de dados que observamos uma profusão de atos legislativos voltados à proteção de dados pessoais.

Como resposta a esta demanda da sociedade por maior segurança das informações e privacidade (Rodrigues et al. 2019, Ferrao et al. 2019), é evidente uma necessidade crescente por novas soluções para esta finalidade. Existe uma miríade de propostas, que abrangem desde novos sistemas que facilitem a configuração segura de equipamentos de rede (Cesário et al. 2022, Fiorenza et al. 2021) e novos protocolos para realização de criptografia dos dados (Oliveira et al. 2021), passando também por soluções de autenticação de usuários (Kreutz et al. 2020, Fernandes et al. 2020, Fernandes et al. 2019). Apesar de soluções válidas, estas não serão abordadas neste trabalho devido a restrições do escopo. Além das soluções mencionadas anteriormente, e dentro do escopo do trabalho, é possível notar um esforço crescente no desenvolvimento de técnicas e mecanismos habilitadores de processos de desidentificação dos dados (Liu, Feng e Zhu 2022, Marques e Bernardino 2020, Brito e Machado 2017).

Um processo de desidentificação envolve a remoção ou alteração de identificadores pessoais, seguido pela aplicação de técnicas ou controles adicionais necessários para remover, obscurecer, agregar, alterar e/ou proteger dados de alguma forma para que não se trata mais de um indivíduo identificável. O processo de desidentificação pode resultar em dados anonimizados ou pseudonimizados.

No setor da saúde, a preocupação com a privacidade é um fator primordial, pois armazenam dados confidenciais e pessoais dos pacientes. Os registros médicos podem

conter os detalhes mais íntimos de uma pessoa, ou seja, informações que não podem ou não devem ser compartilhadas com outro grupo além dos médicos. Manter essas informações seguras coopera para uma relação de confiança entre pacientes e instituições de saúde, que é crucial para prática da saúde bem-sucedida. A confiança do paciente faz com que sejam fornecidos dados com qualidade, onde ambas partes são beneficiadas.

A divulgação ou o vazamento destes dados sensíveis podem trazer danos sociais e até econômicos aos pacientes, por exemplo, a revelação de que um indivíduo testou positivo para *Human immunodeficiency virus (HIV)* pode causar isolamento e discriminação social. Além de que, pessoas podem deixar de procurar ajuda médica com medo da exposição que o compartilhamento indesejado de seus registros médicos pode causar, tendo potencial para causar complicações na prevenção, tratamento e pesquisa de doenças. Segundo o *HIPPA (Health Insurance Portability and Accountability Act) Journal* (Journal 2022), entre 2009 e 2021, mais de 4.400 violações de dados com 500 ou mais registros foram expostas ao *Health and Human Services Office for Civil Rights*. Com essas violações, houve perda, roubo, exposição ou divulgação inadmissível de mais de 314 milhões de registros de saúde, o que equivale a 94,63% da população dos Estados Unidos da América em 2021. A anonimização de dados de saúde é um tema atual, onde tem sido realizado um grande esforço pelas comunidades médicas e de ciência da computação para tornar o compartilhamento de dados entre hospitais e centros de pesquisa mais seguro e confiável (Gentili, Hajian e Castillo 2017). Um dos grandes desafios na anonimização destes dados é proporcionar uma relação saudável entre a utilidade dos dados e a privacidade dos indivíduos.

Na área da saúde, os dados pessoais de indivíduo podem ser utilizados em diversos momentos, seja em hospitais, consultórios ou laboratórios de exames, possibilitando a coleta de diversos tipos de dados. O uso dos dados pessoais neste contexto é caracterizado como o uso primário das informações, isto significa que foram utilizados dentro de sua finalidade original, a prestação do serviço de saúde. Porém, os dados deste indivíduo podem ser utilizados para além da assistência a sua saúde, sendo realizado o uso secundário de suas informações, isto é, a utilização dos dados para finalidades distintas da original, como pesquisa científica, desenvolvimento de novos tratamentos/produtos/serviços, treinamento de inteligência artificial, entre outros. O uso secundário dos dados traz grandes benefícios para todo sistema de saúde (Júnior et al. 2022). A preocupação com a divulgação ou uso de suas informações para fins secundários é latente em muitas pessoas. Como resultado disto, em diversas situações, os indivíduos apontam essa preocupação com a privacidade e confidencialidade como razão para que suas informações de saúde não se-

jam utilizadas para fins de pesquisa (Gentili, Hajian e Castillo 2017). Assim, qualquer sistema de saúde que naturalmente manipule dados de seus pacientes deve implementar mecanismos de preservação da privacidade a fim de permitir extrair os benefícios do uso secundário das informações, mas sem prejuízo aos seus titulares.

Um exemplo prático de utilização primária de dados pessoais no contexto da saúde é o SigSaúde. Este projeto é uma iniciativa da UFRN que tem como objetivo geral o desenvolvimento de uma plataforma computacional integrada a ser utilizada pelas unidades da UFRN que prestam atendimento às pessoas na área da saúde, mas que também possa eventualmente ser integrado a outros sistemas de saúde do governo (SigSaúde 2018). Atualmente, esta plataforma é utilizada na otimização da gestão de processos e na melhoria da integração entre os serviços-escola de saúde da universidade. O seu desenvolvimento também prevê que os dados coletados possam ser disponibilizados, de forma segura e estruturada, para a produção de pesquisas e estudos acadêmicos que podem trazer grandes benefícios à comunidade, como com o uso para controle epidemiológico, respeitando sempre a natureza legal e confidencial desses dados (Filho et al. 2020).

1.1 Problematização e motivação

Esta dissertação de mestrado buscou aplicar o conhecimento adquirido na área de privacidade e proteção de dados, visando endereçar os problemas de compartilhamento seguro e uso secundário de dados de saúde. Muitos exemplos de plataformas de dados de saúde possuem implementados uma série de soluções e mecanismos de segurança, como por exemplo acessos autenticados e autorizados, trilhas de auditoria, entre outros sistemas que permitem o compartilhamento dos dados dos pacientes seguindo normas de segurança. Porém, em sua maioria, o objetivo deste compartilhamento é para utilização primária dos dados, ou seja, dentro da sua finalidade original que é a utilização estritamente na área médica. Existe ainda uma grande deficiência nestas plataformas para a disponibilização de dados visando a utilização secundária, como por exemplo, em pesquisas científicas objetivando novos tratamentos e serviços, bem como em treinamento de pessoas ou sistemas de detecção automática de doenças.

Sendo assim, pode-se dizer que uma solução que permita preservar a privacidade de dados que venham a ser exportados é um requisito fundamental para que estas plataformas ampliem seus modos de utilização, que é justamente o compartilhamento de dados de saúde para dar suporte a pesquisas e estudos clínicos. Neste sentido, este trabalho buscou

preencher essa lacuna ao fornecer uma arquitetura que pode ser conectada à plataformas de saúde para a disponibilização de conjuntos de dados, seguindo as especificações de privacidade definidas. Adicionalmente, a solução desenvolvida foi desenhada para funcionar com diferentes provedores de dados (ou fontes de dados) e, também viabiliza a ligação destes dados em um repositório unificado, por exemplo um *data lake*, permitindo assim a preservação da jornada do paciente nos diferentes provedores, quando for o caso. É importante ressaltar que, embora o problema de compartilhamento e uso secundário de dados pessoais, com preservação da privacidade dos indivíduos a quem os dados dizem respeito, seja um problema que pode ser mapeado para diferentes contextos, nesta dissertação será mais direcionado a plataformas de saúde. Desta forma, será possível colocar em prática os conceitos estudados ao longo desta dissertação, pois o tema Privacidade de Dados é de simples entendimento, mas de difícil implementação. Aqui, os desafios técnicos acabam por também servirem de motivação para o trabalho proposto.

Finalmente, podemos dizer que a proposta apresentada nesta dissertação não deve ter um efeito específico e direto nos vazamentos de dados, porém, esperamos sim, que através dela, consigamos obter ao menos uma dupla de efeitos indiretos. Primeiro, através da oferta da (pseudo)anonimização, ou seja, pseudonimização e anonimização, dos dados espera-se que seja coibido o compartilhamento de dados puros, evitando assim que existam diversas cópias destas informações, e consequentemente, limitando as possibilidades de vazamentos. Segundo, através da possibilidade de compartilhamento legal das informações através de dados (pseudo)anonimizados, espera-se uma redução no interesse de dados obtidos ilegalmente.

1.2 Objetivos

A seguir, são apresentados os objetivos Geral e Específicos para este trabalho.

1.2.1 Objetivo Geral

O objetivo geral deste trabalho de dissertação de mestrado foi desenhar uma arquitetura para o compartilhamento de dados de saúde para uso secundário com a preservação da privacidade dos indivíduos inicialmente relacionados aos dados. Além de garantir a privacidade dos dados, reduzindo a possibilidade de reidentificação dos dados em conformidade com a Lei Geral de Proteção de Dados (LGPD), a solução construída teve ainda como princípio ser de fácil integração/utilização por terceiros.

1.2.2 Objetivos Específicos

Para atingir o objetivo geral, foram definidos como objetivos específicos:

- Conceber um estudo do estado-da-arte em técnicas de (pseudo)anonimização e modelos de privacidade;
- Elaborar um estudo das ferramentas de anonimização disponíveis;
- Idealizar uma arquitetura de (pseudo)anonimização para dados de saúde;
- Implementar uma Prova de Conceito (PoC) para exercitar as principais funcionalidades da arquitetura proposta;
- Realizar testes e validações da PoC em um ambiente real e controlado, visando demonstrar a viabilidade da obtenção dos resultados propostos.

1.3 Contribuições

De forma alinhada com os objetivos gerais e específicos descritos anteriormente, o esforço de pesquisa da dissertação pode ser resumido em três contribuições principais:

- Estudo do estado-da-arte em técnicas de (pseudo)anonimização e modelos de privacidade. O estudo inclui ainda um resumo das ferramentas de desidentificação mais citadas na literatura e que deve servir de referência para novos trabalhos;
- Desenho de uma arquitetura de (pseudo)anonimização que permite: (i) (pseudo)anonimizar dados em saúde, de múltiplas fontes; (ii) ligar os dados (pseudo)anonimizados a partir das diferentes fontes. Estas duas características da arquitetura proposta mostraram-se suficientes para viabilizar o compartilhamento e uso secundário dos dados (pseudo)anonimizados de forma centralizada (por exemplo, num Data Lake);
- Implementação de uma prova de conceito (PoC) que permite a reprodução e continuidade das principais funcionalidades apresentadas pela arquitetura. Além da implementação, este trabalho documenta o processo de implementação dos principais componentes, e das diversas interfaces e realiza testes e validações na PoC.

1.4 Estrutura do trabalho

O restante deste documento está organizado da seguinte forma:

Capítulo 2 - Apresenta o Referencial Teórico desta dissertação, abordando conceitos relevantes de privacidade de dados, algumas das iniciativas legislativas sobre o tema, técnicas básicas de preservação da privacidade, os modelos de privacidade mais relevantes, além de outras tecnologias relacionadas.

Capítulo 3 - Apresenta uma revisão de trabalhos relacionados com a solução proposta nesta dissertação. Assim, são listadas as ferramentas de (pseudo)anonimização mais relevantes. Além disso, estão também relacionados um conjunto de projetos no tema que serviram de inspiração para a solução aqui apresentada.

Capítulo 4 - Apresenta a arquitetura proposta, com detalhamento de seus componentes, responsabilidades e fluxos. É também apresentado o processo de implementação da PoC realizada, seguido de uma descrição detalhada do estado da implementação.

Capítulo 5 - Apresenta os testes realizados sobre a plataforma implementada a fim de verificar o seu correto funcionamento e o atingimento dos objetivos inicialmente declarados. São apresentados quatro diferentes testes: o primeiro voltado para a verificação do correto funcionamento do mecanismo de (pseudo)anonimização; o segundo buscando validar a geração dos pseudônimos global e local de acordo com os requisitos de segurança da arquitetura; o terceiro voltado para a avaliação do correto funcionamento da arquitetura no processo de (pseudo)anonimização dos dados fim a fim num cenário de múltiplos Data Providers; e o quarto e último explorando o cenário de reidentificação de um indivíduo a partir de dados consolidados no Data Lake.

Capítulo 6 - Apresenta as considerações finais do trabalho realizado, incluindo uma revisão crítica dos resultados obtidos, limitações e dificuldades encontradas. São ainda visionadas algumas possibilidades de trabalhos futuros.

2 Referencial teórico

Este capítulo reúne conceitos relevantes de privacidade de dados, iniciativas legislativas sobre o tema de algumas regiões do mundo, técnicas primárias de preservação da privacidade, os modelos de privacidade mais pertinentes, além de outras tecnologias relacionadas ao tema.

2.1 Dados Anonimizados e Pseudonimizados

Num simples processo de desidentificação, a privacidade do indivíduo é supostamente garantida através da remoção dos identificadores explícitos do conjunto de dados. Porém outros dados podem ser usados com informações adicionais para reidentificar um indivíduo (Brito e Machado 2017). A anonimização é o processo de remoção de identificadores diretos (Neumann et al. 2019) e alteração irreversível de identificadores indiretos, que podem levar à identificação de um indivíduo (OLIVEIRA, MADEIRA e MONTEIRO 2020). Desta forma, é possível atingir um nível de privacidade dos indivíduos (Liu, Feng e Zhu 2022), permitindo o compartilhamento e o uso secundário dos dados associados, sem que isto ofereça riscos à privacidade dos donos dos dados (Marques e Bernardino 2020).

No processo de anonimização, os dados são classificados como identificadores diretos ou atributos explícitos, quase-identificadores ou semi-identificadores, sensíveis (Simi, Nayaki e Elayidom 2017) e não-sensíveis ou insensíveis. O atributo explícito é aquele que sozinho permite identificar o indivíduo (Simi, Nayaki e Elayidom 2017), como por exemplo o número do Cadastro de Pessoa Física (CPF). Os quase-identificadores são aqueles que sozinhos geralmente não permitem identificar um indivíduos, mas podem ser cruzados com informações complementares (por exemplo, externas) para identificar o indivíduo, como, por exemplo, sexo, estado civil e data de nascimento. Os dados sensíveis são aqueles que a sua exposição pode causar algum constrangimento ao indivíduo, por exemplo, uma doença (Simi, Nayaki e Elayidom 2017). Já os não-sensíveis são aqueles que

podem ser publicados, pois não permitem identificar e nem causam constrangimento ao usuário.

Na maioria das vezes, a anonimização é aplicada sobre os atributos explícitos e quase-identificadores, fazendo com que seja difícil a identificação do indivíduo, e deixando os atributos sensíveis sem transformações para preservar a sua utilidade (Domingo-Ferrer, Martínez e Sánchez 2022). A utilidade e qualidade dos dados anonimizados é um ponto importante, pois a anonimização tende a trazer perda de dados por supressão (Shin e Kim 2021), criando a dicotomia em que quanto maior a privacidade, menor é a utilidade e a qualidade dos dados.

Diferente da anonimização, onde os dados identificadores diretos são removidos, na pseudonimização, há a substituição por pseudônimos (OLIVEIRA, MADEIRA e MONTEIRO 2020). Um pseudônimo é um dado gerado artificialmente, seja ele uma palavra ou um código (Oliveira 2020). Esta técnica pode ser encarada como uma anonimização reversível (Dubagunta, van Son e Magimai.-Doss 2022), pois os pseudônimos gerados contém uma relação entre eles e o dado real, permitindo assim a reidentificação dos dados pseudonimizados para o dado real (Aamot et al. 2013). Por esta razão, legislações, como o RGPD, exigem que este mapeamento esteja em um ambiente seguro, controlado e separado dos dados para dificultar a reidentificação em caso de vazamento de dados.

Embora o conceito de pseudonimização pareça simples, é necessário ter muito cuidado na sua implementação. Na literatura, é possível encontrar algumas sugestões de metodologias que podem ser adotadas. Uma boa fonte é o trabalho de (Aamot et al. 2013), que apresenta os métodos: (i) de Noumeir, em que a ligação entre pseudônimo e o dado real é direto por um terceiro confiável; (ii) de Pommerening, em que o pseudônimo (pseudônimo 1) é também pseudonimizado, gerando assim outro pseudônimo (pseudônimo 2), e então para chegar ao dado real é necessário que um terceiro confiável faça a ligação entre o pseudônimo 1 e o pseudônimo 2, e outro terceiro confiável faça a ligação entre o pseudônimo 1 e o dado real; e (iii) o método de De Moor, que é bem parecido com o de Pommerening, porém, a ligação entre o dado real e o pseudônimo 1 é feita por um Provedor de Dados. A Figura 1 demonstra os três métodos citados.

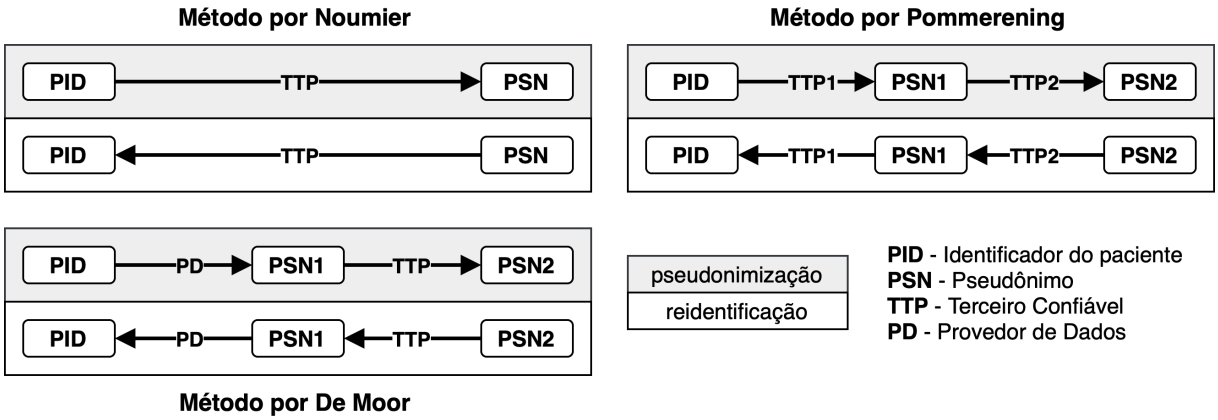


Figura 1: Métodos de pseudonimização. Adaptado de (Aamot et al. 2013).

2.2 Técnicas de Anonimização

As técnicas de anonimização de dados ajudam a remover ou transformar o conteúdo da informação dos dados. Essas técnicas podem envolver transformações estatísticas, baseadas em algoritmos ou personalizadas e devem garantir que o valor transformado e o valor original pertençam ao mesmo domínio. A seguir, são listadas as técnicas de anonimização de dados mais recorrentes na literatura sobre o tema.

2.2.1 Generalização

A generalização é uma técnica aplicada aos dados quase-identificadores, substituindo-os por valores menos específicos, mas semanticamente consistentes (Balusamy e Muthusundari 2014, Jr, Machado e Monteiro 2014, Mimoto., Basu. e Kiyomoto. 2016, Brito e Machado 2017, Hore et al. 2021). Os atributos quase-identificadores sujeitos à generalização podem ser contínuos ou categóricos. Um atributo contínuo é numérico e pode assumir um número infinito de valores reais diferentes, enquanto que um atributo categórico assume um valor em um conjunto limitado. A técnica de generalização requer a definição de uma hierarquia para cada atributo quase-identificador. Cada hierarquia contém pelo menos dois níveis. A raiz é o valor mais geral. Representa o nível mais alto. As folhas correspondem aos valores dos dados originais e constituem o nível mais baixo (Fredj, Lammari e Comyn-Wattiau 2015). Por exemplo, considerando o atributo quase-identificador categórico CIDADE, é possível definir uma hierarquia onde o primeiro nível (após a raiz, ou seja, o valor original) da

generalização seja alterar a cidade pelo estado, no nível seguinte para país, e por fim o seu respectivo continente. Para um quase-identificador contínuo IDADE, é possível definir para o primeiro nível um conjunto de intervalos a cada 5 anos, seguido de um intervalo maior de 10 anos e assim por diante.

Embora de conceito simples, dependendo do domínio dos quase-identificadores, o uso desta técnica pode se mostrar difícil e complexo, principalmente para a definição das hierarquias, visto que o domínio do atributo pode não ser bem definido, como por exemplo a raça de um indivíduo. Além disso, (Brito e Machado 2017) ressalta que a utilização inadequada da generalização, pode fazer com que os dados se tornem inúteis para determinadas análises e, portanto, é necessário buscar um conjunto mínimo necessário de alterações para que mantenha uma utilidade adequada dos dados ao mesmo tempo que atenda aos requisitos de privacidade. Na prática, uma generalização ótima de dados quase certamente acaba sendo um problema computacionalmente desafiador (NP-difícil) e, por esta razão, a maioria dos algoritmos propostos na literatura utiliza alguma heurística gulosa (do inglês, *greedy*) para busca que leva a soluções localmente ótimas (Hore et al. 2021).

Por fim, (Brito e Machado 2017, Fung et al. 2010) apontam que a generalização pode ser aplicada aos quase-identificadores de diferentes formas: i) alterando todo o domínio com sua respectiva hierarquia, todas em um mesmo nível; ii) aplicada em apenas um domínio, então todos os registros que estão em um nível de hierarquia devem ser generalizados, por exemplo generalizar uma cidade de MG para o estado, então todas as cidade de MG devem ser generalizadas; iii) aplicada em apenas um atributo específico, então a generalização da cidade X, implicará na generalização de todos os registros da cidade X; e por fim, iv) também ser aplicada a apenas um quase-identificador de um registro qualquer.

Cidade	Idade	Saldo
Natal	30	300.00
Belo Horizonte	43	450.00
Parnamirim	32	300.00
Sete Lagoas	40	500.00

(a) Dados originais.

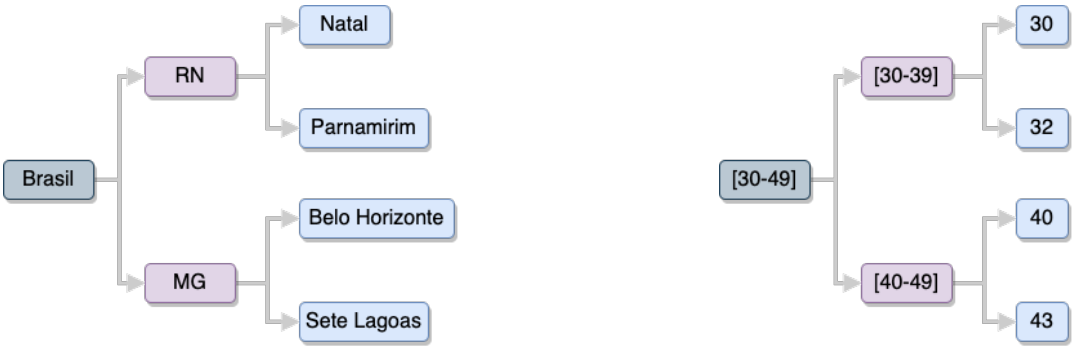
Cidade	Idade	Saldo
RN	[30-39]	300.00
MG	[40-49]	450.00
RN	[30-39]	300.00
MG	[40-49]	500.00

(b) Dados generalizados.

Figura 2: Generalização dos dados baseada nos atributos CIDADE e IDADE.

Exemplo de Generalização: Para o conjunto de dados ilustrado na Figura 2a, se aplicarmos aos quase-identificadores CIDADE e IDADE as suas respectivas hierarquias

ilustradas na Figura 3, é possível obter, como resultado, o conjunto de dados generalizado ilustrado na Figura 2b.



(a) Hierarquia para o atributo CIDADE. (b) Hierarquia para o atributo IDADE.

Figura 3: Hierarquias de generalização para os atributos CIDADE e IDADE.

2.2.2 Supressão

A supressão é uma técnica que, assim como a generalização, é aplicada aos quase-identificadores, removendo o atributo sempre que o mesmo não for passível de anonimização, não seja relevante ou não seja necessário para a análise (OLIVEIRA, MADEIRA e MONTEIRO 2020, Marques e Bernardino 2020, Jr, Machado e Monteiro 2014, Brito e Machado 2017). Esta técnica também pode ser encarada como um caso extremo da generalização, em que registros são generalizados para além do nó raiz na hierarquia de generalização, gerando uma supergeneralização que impõe a remoção do dado (Jr, Machado e Monteiro 2014, Hore et al. 2021, Brito e Machado 2017). Como apontado por (Lee et al. 2017), os *outliers* - dados que se diferenciam drasticamente do restante - são os principais alvos desta supergeneralização, já que a baixa ocorrência destes valores não permitem a criação de sua própria classe de equivalência, sendo necessário aplicar a supressão destes valores.

A supressão pode ser aplicada globalmente ou localmente. Na supressão global todos os valores de um determinado atributo são removidos, garantindo assim que nenhuma informação será disponibilizada sobre o atributo. Já na supressão local apenas alguns registros de um valor do atributo são removidos, porém sendo necessário garantir que os demais valores daquele atributo não possam ser descobertos (Brito e Machado 2017, Liu, Feng e Zhu 2022). Existem três tipos de supressão: a supressão de registro, de valor e de célula. A supressão de registro implica na remoção de um registro inteiro do conjunto de dados, logo nenhum atributo deste registro é mantido. Na supressão de valor todas as instâncias de um determinado valor são removidas do conjunto de dados, por exemplo,

os valores do atributo idade que sejam menores ou iguais a 15 em um conjunto de dados podem ser removidos, enquanto os demais valores não sofrem alteração. E, por fim, na supressão de célula, ou supressão celular, apenas algumas instâncias de um determinado valor de atributo são removidas, representando uma supressão local, por exemplo, pode-se remover apenas um terço dos valores do atributo idade que sejam iguais a 15, assim ainda seria encontrado com o valor 15 para para o atributo idade, mas não na proporção original (Fung et al. 2010, Brito e Machado 2017).

(Marques e Bernardino 2020) ressalta que a principal vantagem desta técnica é tornar impossível a recuperação das informações que foram suprimidas, uma vez que foram removidas permanentemente. Porém, (Lee et al. 2017) alerta que a supressão prejudica a veracidade dos dados, pois conforme mencionado anteriormente a supressão realiza uma remoção permanente de dados que podem vir a ser úteis no em uma análise futura. Portanto esta operação deve ser realizada em quantidades insignificantes de registros para preservar a qualidade dos dados.

Exemplo de Supressão: Para o conjunto de dados ilustrado na Figura 4a, é possível suprimir globalmente o quase-identificador TELEFONE (caso não tenha relevância para a análise), removendo essa informação, como ilustrado na Figura 4b. Outra forma de supressão poderia ser aplicada sobre todo o registro de ID com valor igual a 2, resultando no conjunto de dados ilustrado na Figura 4c.

ID	Nome	Telefone
1	Batman	99824-2489
2	Charada	94752-9284
3	Coringa	97258-2849
4	Robin	99263-5526

(a) Dados originais.

ID	Nome
1	Batman
2	Charada
3	Coringa
4	Robin

(b) Dados após a supressão global do atributo TELEFONE.

ID	Nome	Telefone
1	Batman	99824-2489
3	Coringa	97258-2849
4	Robin	99263-5526

(c) Dados após a supressão de registros com ID com valor 2.

Figura 4: Exemplos de supressão de dados.

2.2.3 Mascaramento

O mascaramento é uma técnica que altera os caracteres do valor de um atributo no conjunto de dados por um símbolo constante, como por exemplo, “*” ou “X”, sendo tipicamente aplicada de maneira parcial (OLIVEIRA, MADEIRA e MONTEIRO 2020,

Marques e Bernardino 2020, Nelson 2015, Singapore 2018). Esta técnica é utilizada quando ocultar parte do valor de um atributo é o suficiente para fornecer o anonimato do indivíduo, onde dependendo do atributo, pode ser definido a substituição de um número fixo de caracteres, por exemplo para número de cartão de crédito, ou de um número variável de caracteres, por exemplo para o endereço de e-mail (Singapore 2018). De acordo com o atributo, apenas parte de seu valor pode já trazer utilidade para a análise dos dados, por exemplo, o Código de Endereçamento Postal (CEP), que é composto por oito caracteres e cada um deles representa uma informação, poderia ser aplicado o mascaramento nos cinco últimos caracteres, e ainda sim seria possível obter a região, sub-região e setor do CEP em questão (Marques e Bernardino 2020).

(Singapore 2018) adverte que a aplicação da técnica deve levar em consideração se o comprimento do valor de um atributo traz informações relevantes sobre os dados originais. (Nelson 2015) lembra ainda que a técnica, se aplicada de forma simples, como exemplo substituindo apenas o primeiro caractere do valor, tem seu uso limitado, pois o valor original do dado pode ser facilmente reconstruído.

Exemplo de Mascaramento: Para o conjunto de dados ilustrados na Figura 5a, é possível aplicar o mascaramento do atributo CEP de modo a manter apenas a informação sobre a sua região, sub-região, setor, subsetor e divisor de subsetor. O conjunto de dados resultante é ilustrado na Figura 5b.

ID	Nome	CEP
1	Batman	11400-873
2	Charada	26500-723
3	Coringa	35702-763
4	Robin	11400-872

(a) Dados originais.

ID	Nome	CEP
1	Batman	11400-***
2	Charada	26500-***
3	Coringa	35702-***
4	Robin	11400-***

(b) Dados mascarados.

Figura 5: Exemplo de Mascaramento do atributo CEP.

2.2.4 Embaralhamento

O embaralhamento é uma técnica que não altera o valor dos atributos, mas sim, os alterna de forma aleatória com outro registro, onde o valor do atributo A na tupla C_I é substituído pelo valor do atributo A na tupla C_K, onde $I \neq K$, fazendo com que todos os valores do conjunto de dados seja real, mas atribuídos a indivíduos "errados" (Marques e Bernardino 2020, Nelson 2015, Brito e Machado 2017, PINHO 2020).

(Raghunathan 2013) considera o embaralhamento uma técnica simples de ser implementada e cita uma variação da técnica, o embaralhamento em grupo. Embaralhamento em grupo é utilizado quando as informações agrupadas precisam ser anonimizadas em conjunto, onde um grupo de atributos é embaralhado. Embaralhamento tem maior utilidade quando a análise a ser realizada será executada sob apenas um atributo e não é necessário relacioná-los com os demais atributos do registro (Marques e Bernardino 2020). Outro exemplo do uso do embaralhamento, é a criação de conjunto de dados para testes de software (Nelson 2015).

Apesar de ser considerada uma técnica simples, devem ser tomados alguns cuidados. O embaralhamento nem sempre fornece a anonimização dos dados, sendo possível a reorganização dos dados a sua forma original, por isto, deve ser utilizado em conjunto com outras técnicas (Marques e Bernardino 2020, PINHO 2020). Além disto, o embaralhamento é mais eficaz quando aplicado a um grande conjunto de dados, pois em um conjunto de dados com poucos registros é fácil rastrear os dados confidenciais originais. (Singapore 2018) ressalta que deve ser realizada uma avaliação sobre os atributos para que se decida quais deverão sofrer embaralhamento, pois, em determinadas situações, pode-se chegar a conclusão, por exemplo, que apenas os identificadores precisam ser alternados.

Exemplo de Embaralhamento: Para o conjunto de dados ilustrado na Figura 6a, é possível aplicar o embaralhamento em grupo nestas informações, resultando no conjunto de dados ilustrado na Figura 6b.

Nome	Telefone	Idade
Batman	99824-2489	32
Charada	94752-9284	18
Coringa	97258-2849	23

(a) Dados originais.

Nome	Telefone	Idade
Batman	97258-2849	23
Charada	99824-2489	32
Coringa	94752-9284	18

(b) Dados embaralhados.

Figura 6: Exemplo de Embaralhamento de dados.

2.2.5 Adição de ruído

A adição de ruído é uma técnica geralmente aplicada a atributos numéricos ou de datas, onde o valor original de um atributo v é substituído por $v+r$, sendo r o ruído adicionado (Brito e Machado 2017, Fung et al. 2010, Virupaksha e Dondeti 2021). A adição de ruído traz grande utilidade quando aplicada a conjunto de dados que irão sofrer análises estatísticas simples, pois após sua aplicação podem ser mantidas informações, como

média e correlação (Brito e Machado 2017, Fung et al. 2010). Porém, deve-se ter cautela com o nível de ruído que será aplicado sobre o conjunto de dados para que a análise dos dados e a privacidade dos indivíduos sofra pouco impacto (Marques e Bernardino 2020). Além disso, o nível de perturbação dos dados deve ser proporcional ao intervalo de valores dos atributos, ainda que essa definição seja uma tarefa complexa (PINHO 2020).

Exemplo de Ruído: Para o conjunto de dados ilustrado na Figura 7a, é possível adicionar ruído a fim de proteger o atributo sensível SALDO. Neste exemplo, o ruído é adicionado a partir do incremento de 13% ao valor do atributo SALDO, resultando no conjunto de dados ilustrado na Figura 7b.

Nome	Saldo
Batman	302.50
Charada	124.98
Coringa	200.00

(a) Dados originais.

Nome	Saldo
Batman	341.83
Charada	141.23
Coringa	226.00

(b) Dados após a adição de ruído.

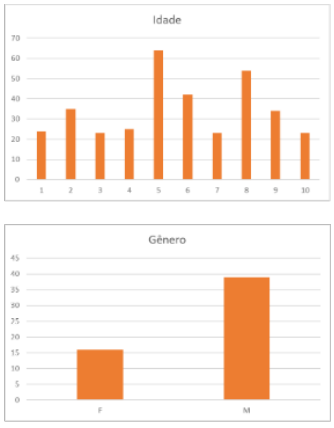
Figura 7: Exemplo de adição de ruído.

2.2.6 Dados Sintéticos

Distinto das técnicas apresentadas anteriormente, os dados sintéticos são dados gerados de forma artificial ao invés de serem criados por eventos reais (James et al. 2021). Dados sintéticos são gerados através de um modelo estatístico construído a partir do conjunto de dados original e que busca preservar e espelhar as suas propriedades estatísticas (James et al. 2021, Fung et al. 2010, Brito e Machado 2017). A grande vantagem dessa técnica é que as características estatísticas dos dados gerados representam, na medida do possível, uma boa aproximação dos dados originais, porém não dizem respeito a pessoas reais e, portanto, não estão sujeitos à mesma proteção legal e imposição de medidas de privacidade. Porém, é necessário ter cuidado com os dados gerados, pois podem conter valores sem sentido ou irreais para determinados atributos (Brito e Machado 2017).

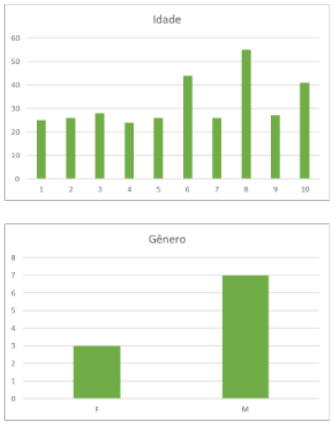
Exemplo de geração de dados sintéticos: Utilizando o pacote *Synthetic Data Vault (SDV)* (SDV 2022, Patki, Wedge e Veeramachaneni 2016) para gerar dados sintéticos com base no conjunto de dados ilustrado na Figura 8a, é possível obter em uma de suas execuções, como resultado, o conjunto de dados ilustrado na Figura 8b.

ID	Gênero	Idade
1	F	24
2	F	35
3	M	23
4	F	25
5	M	64
6	M	42
7	M	23
8	M	54
9	F	34
10	M	23



(a) Dados originais.

ID	Gênero	Idade
1	M	25
2	M	26
3	F	28
4	F	24
5	F	26
6	M	44
7	M	26
8	M	55
9	M	27
10	M	41



(b) Dados sintéticos gerados através do pacote *SDV* (SDV 2022).

Figura 8: Exemplo de geração de dados sintéticos.

2.3 Modelos de Privacidade

Nas últimas duas décadas, vários modelos de anonimato de dados foram propostos na literatura. Esta seção reúne alguns dos mais referenciados na literatura sobre o tema.

2.3.1 *k*-anonimato

O modelo de privacidade k-anonimato é uma técnica que garante que cada registro seja semelhante a, pelo menos, $k-1$ outros registros do conjunto de dados, onde k é um valor inteiro positivo (Sweeney 2002, Lee et al. 2017). Este processo evita a violação de privacidade por quase-identificadores. Com isso, cada registro tem probabilidade menor ou igual a $\frac{1}{k}$ de ser ligado a um indivíduo (Murthy et al. 2019, Li e Lai 2017, Medková 2020, Liu, Feng e Zhu 2022, Jr, Machado e Monteiro 2014). Assim, o valor de k define o nível de privacidade do conjunto de dados e, deste modo, quanto maior o valor de k, menor a probabilidade de reidentificação de um indivíduo. Porém, o valor de k também afeta diretamente a perda de informação, visto que registro que não possam ser agrupados em grupos de equivalência serão suprimidos, reduzindo a utilidade dos dados (Brito e Machado 2017, Marques e Bernardino 2020).

O k-anonimato é o mais conhecido modelo de privacidade, sendo utilizado como forma de proteção ao ataque de ligação ao registro (Jr, Machado e Monteiro 2014, Singapore 2018, Brito e Machado 2017). Porém, apesar de ser eficaz contra o ataque de ligação ao registro - ataque na qual o adversário tem como objetivo identificar que um

registro refere-se a um indivíduo em específico - o k-anonimato não é eficaz com o ataque de ligação ao atributo - ataque em que o adversário pode inferir atributos sensíveis a um indivíduo mesmo sem o reidentificar, apenas baseado na classe de equivalência que a vítima pertence - e contra o ataque de ligação à tabela - ataque na qual o adversário tem como objetivo perceber a presença ou não do indivíduo no conjunto de dados (Fung et al. 2010, PINHO 2020, Jr, Machado e Monteiro 2014).

Exemplo de *k*-anonimato: Para o conjunto de dados ilustrado na Figura 9a, se aplicarmos aos quase-identificadores CIDADE e IDADE o k-anonimato, com $k = 2$, é possível obter, como resultado, o conjuntos de dados ilustrado na Figura 9a. Veja que foi aplicado a técnica de generalização sobre os dados para que se tornasse possível atingir o 2-anonimato.

ID	Cidade	Idade	Saldo
1	Natal	30	300.00
2	Belo Horizonte	43	450.00
3	Parnamirim	30	345.00
4	Sete Lagoas	40	500.00

(a) Dados originais. Não atendem à generalização 2-anonimato.

Cidade	Idade	Saldo
RN	[30-39]	300.00
MG	[40-49]	450.00
RN	[30-39]	345.00
MG	[40-49]	500.00

(b) Dados após a generalização 2-anonimato.

Figura 9: Exemplo de generalização k-anonimato.

2.3.2 *l*-diversidade

O modelo l-diversidade introduzido como um complemento ao k-anonimato (Machanavajjhala et al. 2006), ou seja, devendo ser aplicado a um conjunto de dados já k-anônimo. Esta técnica busca fornecer proteção contra o ataque de ligação ao atributo ao qual o k-anonimato é sensível (Fung et al. 2010, Brito e Machado 2017). Em seu funcionamento geral, cada classe de equivalência gerada pelo k-anonimato, devem conter pelo menos l - sendo l um número inteiro positivo - valores distintos para os atributos sensíveis, assim, mesmo que seja inferido a classe de equivalência de um indivíduo, não será possível deduzir qual o seu atributo sensível com uma probabilidade maior que $\frac{1}{l}$ (PINHO 2020, Fung et al. 2010, Virupaksha e Dondeti 2021, Jr, Machado e Monteiro 2014, Brito e Machado 2017).

Entre as limitações no uso do l-diversidade, estão: (i) o impacto desta técnica na utilidade dos dados verifica-se quando aplicada em um cenário onde existe muitos valo-

res repetidos para os atributos sensíveis de uma classe de equivalência e pouco/nenhum valores distintos, sendo necessário realizar muita alteração ou supressão de valor, diminuindo significativamente a qualidade dos dados; (ii) esta técnica é suscetível ao ataque de assimetria (Brito e Machado 2017, Jr, Machado e Monteiro 2014, Fung et al. 2010), permitindo ao atacante inferir o valor de um atributo sensível a um indivíduo com uma chance maior do que a proporcionada pela distribuição global; (iii) esta técnica está exposta também ao ataque de similaridade (Brito e Machado 2017, Jr, Machado e Monteiro 2014) - este tipo de ataque ocorre quando mesmo os atributos sensíveis sendo distintos, eles são semanticamente semelhantes; e (iv) o l-diversidade é incapaz de lidar com semântica dos novos valores que irão permutar com os originais, ou seja, a técnica assume que os atributos sensíveis são categorizados, desconsiderando outros tipo de valores, como por exemplo o numérico, onde a descoberta de valores aproximados pode ser o suficiente (Brito e Machado 2017, Jr, Machado e Monteiro 2014).

Exemplo de *l*-diversidade: Para o conjunto de dados que atende ao 2-anonimato, ilustrado na Figura 10a, se aplicarmos ao atributo sensível SALDO o l-diversidade, com $l = 2$, é possível obter, como resultado, o conjunto de dados ilustrado na Figura 10b. Vale ressaltar que apenas os registros de ID 1 e 3 não atendiam à 2-diversidade no conjunto dados originais.

ID	Cidade	Idade	Saldo
1	RN	[30-39]	300.00
2	MG	[40-49]	450.00
3	RN	[30-39]	300.00
4	MG	[40-49]	500.00

(a) Dados originais. Atende à generalização 2-anonimato, mas não atendem à generalização 2-diversidade.

Cidade	Idade	Saldo
RN	[30-39]	300.00
MG	[40-49]	450.00
RN	[30-39]	230.00
MG	[40-49]	500.00

(b) Dados após a generalização 2-diversidade.

Figura 10: Exemplo de generalização 2-anonimato com 2-diversidade.

2.3.3 *t*-proximidade

De forma similar ao l-diversidade, o t-proximidade foi proposto para ser um modelo de privacidade complementar ao k-anonimato e ao l-diversidade (Li, Li e Venkatasubramanian 2006). Desta forma, precisar ser aplicado a um conjunto de dados já k-anônimo e l-diverso. A técnica objetiva superar a limitação do l-diversidade em relação ao ataque de assime-

tria, pois ela infere que informações sobre atributos sensíveis podem ser alcançadas a partir da frequência de ocorrência destes atributos na tabela (Brito e Machado 2017, Jr, Machado e Monteiro 2014, Virupaksha e Dondeti 2021). Uma classe de equivalência tem t-proximidade quando a distância da distribuição do atributo sensível desta classe para a distribuição desse mesmo atributo em todo o conjunto de dados não for maior que o limite t definido. Assim, um conjunto de dados pode ser dita que possui t-proximidade quando todas as classes de equivalência têm t-proximidade (Li, Li e Venkatasubramanian 2006, Domingo-Ferrer e Soria-Comas 2015, Rajendran, Jayabalan e Rana 2017). A distância entre as distribuições da classe de equivalência e global é medida usando a métrica chamada *Earth Mover's Distance (EMD)* e seu resultado é um valor real no intervalo $[0, 1]$, onde quanto maior o valor, menor é a proteção (Jr, Machado e Monteiro 2014, Brito e Machado 2017, Fung et al. 2010, Li, Li e Venkatasubramanian 2006, Soria-Comas et al. 2015).

Entre os benefícios do uso do t-proximidade, estão (Rajendran, Jayabalan e Rana 2017): (i) impedir a divulgação de atributos que preservam a privacidade dos dados, (ii) fornecer proteção contra ataques que o k-anonimato e l-diversidade não oferecem e (iii) identificar a proximidade semântica dos atributos com a EMD. Porém, existem as seguintes limitações: (i) independente do nível de privacidade de cada atributo sensível, será sempre usada a mesma especificação, trazendo uma falta de flexibilidade, (ii) o uso da EMD não é apropriada para ataques de ligação ao atributo quando os mesmos são numéricos e (iii) ter uma base de dados com t-proximidade pode diminuir a utilidade dos dados para que seja garantido a mesma distribuição nas classes de equivalência (Brito e Machado 2017).

2.3.4 Privacidade Diferencial

Diferente dos modelos de privacidade anteriores, a privacidade diferencial não é utilizada para fornecer um conjunto de dados anonimizados, mas sim responder a consultas aos dados (Singapore 2018). O objetivo desta técnica é tornar os dados dos indivíduos anônimos acrescentando um ruído as respostas das consultas realizadas na base de dados, fornecendo informações estatísticas sobre o conjunto de dados sem afetar a privacidade dos indivíduos (Singapore 2018, Senavirathne e Torra 2019, Sei et al. 2019, Brito e Machado 2017, Domingo-Ferrer e Soria-Comas 2015, Li e Lai 2017). A utilização da privacidade diferencial pode se tornar complexa. Primeiro, pela dificuldade em calcular a sensibilidade da consulta aos dados, podendo gerar demasiado ruído, afetando de forma

significativa a utilidade dos dados. Mesmo com a aplicação de técnicas para minimizar essa perda, é difícil encontrar um equilíbrio no ruído adicionado e garantir a privacidade. Segundo, os valores retornados nas consultas podem não ser pertinentes em áreas específicas devido a incerteza de como os dados são gerados e é por isso que técnicas de anonimização como k-anonimato acabam por serem mais utilizadas na prática (Brito e Machado 2017).

3 Trabalhos relacionados

A desidentificação de dados pessoais, seja através da anonimização ou pseudonimização, é um processo complexo que requer habilidades técnicas e que envolvem a compreensão dos requisitos de privacidade, a compreensão da natureza dos dados, a compreensão dos requisitos legais associados, além de outros aspectos. Embora existam várias ferramentas de desidentificação disponíveis para auxiliar nesta tarefa, a seleção entre estas diversas ferramentas disponíveis é, por si, um desafio devido à necessidade de coletar, uniformizar e comparar informações de cada uma. Assim, a fim de auxiliar neste desafio, este capítulo reúne algumas das ferramentas mais referenciadas na literatura sobre o tema. Adicionalmente, é listado um conjunto de projetos de pesquisa que apresentam processos de desidentificação arquiteturalmente interessantes e que forneceram boas ideias para a solução aqui proposta.

3.1 Ferramentas de Desidentificação de Dados

Esta seção reúne um conjunto de ferramentas de anonimização ou pseudonimização de dados com o objetivo de entender seu potencial, avaliar seus prós e contras, e extrair ideias e os principais requisitos para o desenvolvimento da arquitetura de desidentificação de dados que foi proposta na dissertação de mestrado. Estas ferramentas serão dividas em três grandes grupos, as de código aberto, as sem custo financeiro e as proprietárias.

3.1.1 Ferramentas de código aberto

O ARX (Prasser et al. 2020) é considerado uma das principais ferramentas de anonimização (Tomás, Rasteiro e Bernardino 2022), pois suporta uma ampla variedade de (i) modelos de privacidade e riscos, como k-anonimato, l-diversidade, t-proximidade, δ-presença e privacidade diferencial; (ii) métodos de transformação de dados, como esquemas de transformação local e global, generalização de valor, amostragem aleatória, exclusão de

registro, atributo e células, microagregação, *top and bottom coding* e categorização; análise de risco de métodos de reidentificação e utilidade dos dados resultantes, fornecendo modelos de uso geral, orientados a células, registros e atributos (Prasser et al. 2020). Por este motivo, na literatura (Vovk, Piho e Ross 2021) é utilizado em diversos estudos, quer para a anonimização de conjuntos de dados (Silva, Basso e Moraes 2017, Jakob et al. 2020, Gentili, Hajian e Castillo 2017, Boeck et al. 2021, Prasser et al. 2016), quer para a análise dos riscos de reidentificação (Jyothi e Rao 2017).

O Amnesia (Amnesia 2022) é outra ferramenta popular de animização (Tomás, Rasteiro e Bernardino 2022), seguindo as diretrizes do RGPD e suportando alguns modelos de privacidade, como k-anonimato e k^m-anonimato (Kulkarni e Bedekar 2022, Crutzen, Peters e Mondschein 2019).

O μ-ARGUS (ARGUS 2022) é um software que visa auxiliar a produção de microdados seguros. O nome ARGUS é um acrônimo para *AntiRe-Identification General Utility System*. Inicialmente projetado como uma ferramenta privada, os últimos lançamentos foram transferidos para código aberto. Para tornar os microdados seguros, o modelo de privacidade k-anonimato é usado na maioria das etapas, mas também é possível aplicar transformações adicionais, como: supressão local, agrupamento de categorias, adição de ruído e dados sintéticos (Stenersen 2020).

Assim como o μ-ARGUS, o sdcMicro (Templ, Kowarik e Meindl 2015) é um pacote R que permite a anonimização de microdados. SDC é uma abreviação de *Statistical Disclosure Control*. O sdcMicro foi desenvolvido para auxiliar pesquisas na geração de microdados para uso público. No sdcMicro, são utilizados dois modelos de privacidade, k-anonimato e l-diversidade, bem como métodos para transformação de dados, como randomização, *top and bottom coding*, supressão e *recoding* (Zuo et al. 2021).

O Anonimatron (Anonimatron 2022) realiza a anonimização de dados por meio de pseudônimos e permite gerar nomes romanos falsos, endereços de e-mail e identificadores únicos universais, e afirma ser compatível com o RGPD (Zuo et al. 2021).

O CHORUS (Johnson et al. 2020) é um *framework* que fornece uma biblioteca Scala que permite a implementação de métodos diferenciais de privacidade em um modelo cooperativo.

O g9 Anonymizer (esito 2022) é uma ferramenta que vem como *plugin* do Eclipse, que fornece lógica de anonimização programável. Para obter a desidentificação de dados, o *plugin* oferece suporte a transformações de dados, como mascaramento, embaralhamento,

síntese de geração de dados e supressão.

O *The University of Texas at Dallas Anonymization Toolbox* (Dallas 2022) é um software desenvolvido no *UT Dallas Data Security and Privacy Lab* e implementa seis métodos de anonimização: Datafly (Sweeney 1997), Mondrian Multidimensional k-Anonymity (LeFevre, DeWitt e Ramakrishnan 2006), Incognito (LeFevre, DeWitt e Ramakrishnan 2005), Incognito com l-diversidade (Han, Yu e Yu 2008), Incognito com t-proximidade (Li, Li e Venkatasubramanian 2007) e Anatomy (Xiao e Tao 2006).

O *Cornell Anonymization Toolkit* (CAT) (Xiao, Wang e Gehrke 2009) é outra ferramenta gratuita que permite anonimização de dados com uma interface intuitiva. A ferramenta suporta o algoritmo Incognito e os modelos de privacidade l-diversidade e t-proximidade (Maier 2013).

3.1.2 Ferramentas gratuitas

Algumas outras ferramentas também são gratuitas para uso, mas seu código-fonte não está disponível. A *Tool for Interactive Analysis of Microdata Anonymization Techniques* (TIAMAT) (Dai et al. 2009) suporta diferentes algoritmos de anonimização, como Mondrian (LeFevre, DeWitt e Ramakrishnan 2006) e k-Member (Byun et al. 2007), bem como vários modelos para análise e otimização do utilitário de dados de saída, bem como modelos de privacidade k-anonimato, l-diversidade e t-proximidade.

O *System for Evaluating and Comparing RElational and Transaction Anonymization algorithms* (SECRETA) (Poulis et al. 2014) tem como foco a análise da eficácia e eficiência de algoritmos de anonimização para dados tabulares e com valor definido. SECRETA suporta nove algoritmos, quatro para lidar com conjuntos de dados com atributos relacionais (Incognito (LeFevre, DeWitt e Ramakrishnan 2005), Cluster (Poulis et al. 2013), Top-down (Fung, Wang e Yu 2005) e Full subtree bottom-up) e cinco para lidar com conjuntos de dados com atributos da transação (COAT (Loukides, Gkoulalas-Divanis e Malin 2011), PCTA (Gkoulalas-Divanis e Loukides 2012), Apriori, LRA e VPA (Terrovitis, Mamoulis e Kalnis 2011)).

O NLM-Scrubber (Medicine 2022) foi desenvolvido pela *National Library of Medicine* (NLM) para desidentificar textos clínicos. O objetivo do NLM-Scrubber é gerar informações de saúde adequadas com o *Health Insurance Portability and Accountability Act* (HIPPA). Diferentemente de outras ferramentas, o NLM-Scrubber realiza a desidentifi-

cação de textos, substituindo termos que representam informações como idade, endereço, dados e *Personally Identifiable Information* (PII), por uma tag que identifica apenas o tipo de informação no texto. Por exemplo, no texto: "Dr. Pedro visitou o paciente de 98 anos...", o texto desidentificado conteria "Dr. [PERSONALNAME] visitou o paciente [AGE90+]..."

3.1.3 Ferramentas proprietárias

Do lado proprietário, a necessidade de ferramentas de anonimização de dados que protejam a atividade privada de indivíduos e corporações em conformidade com o RGPD criou um mercado lucrativo que explica o surgimento de uma miríade de ferramentas. Aircloak Insights (Aircloak 2022) é uma ferramenta privada que atua como um proxy entre os analistas de dados e o conjunto de dados. O Aircloak Insights consiste em dois componentes: Insights Air e Insights Cloak. O Insights Cloak é responsável por realizar a análise e anonimização dos dados sensíveis, conectando-se aos bancos de dados que contêm os dados sensíveis, sem a necessidade de alterações. A anonimização é baseada em uma combinação de técnicas usadas ao longo do tempo, como k-anonimato, supressão, ruído de privacidade diferencial e *top and bottom coding*.

Como as limitações dos métodos tradicionais de desidentificação estão se tornando mais evidentes, ferramentas modernas são desenvolvidas para produzir resultados efetivos com dados estruturados e não estruturados em uma vasta gama de campos e setores. As novas ferramentas misturam métodos tradicionais de desidentificação com novos, como dados sintéticos, processamento de linguagem natural e inteligência artificial. Por exemplo, o CloverDX (CloverDX 2022) está focado na desidentificação de conjuntos de dados em nível de produção para desenvolvimento, visualização, teste, análise ou prototipagem. A ferramenta permite um conjunto de transformações de dados com base em uma combinação de mascaramento e geração de dados sintéticos.

Da mesma forma, o foco do BizDataX (BizDataX 2022) permite o anonimato dos dados de produção para desenvolvimento e teste e oferece uma caixa de ferramentas de mascaramento de dados para ocultar identidades e dados confidenciais, alcançando a conformidade com o RGPD. Criada para atender às necessidades de anonimização de uma grande empresa farmacêutica, a solução de anonimização de dados da Gramener (Gramener 2022) usa NLP para redigir informações privadas de pacientes de documentos de ensaios clínicos, de acordo com HIPPA e RGPD.

Outra ferramenta útil para redigir documentos sensíveis é o *Docbyte's Real-time Au-*

tomated Anonymization (Docbyte 2022). A ferramenta usa inteligência artificial e aprendizado de máquina para anonimizar os dados. Ela pode escurecer ou desfocar imagens e redigir texto considerado sensível usando algoritmos focados em imagens e reconhecimento de objetos.

3.1.4 Comparativo

A fim de fornecer uma visão geral das ferramentas discutidas neste capítulo, apresentamos um comparativo na Tabela 1, considerando as seguintes características para cada uma:

- ***Nome da Ferramenta***: Nome comercial ou de projeto da ferramenta.
- ***Data da última versão***: Data da última versão lançada. Diz sobre a continuidade da ferramenta, mostrando a busca constante dos seus autores por melhorias, correções e/ou novas funcionalidades.
- ***Open source***: Possui código aberto. O uso do *open source* permite a modificação, estudo e distribuição da ferramenta de forma gratuita.
- ***GUI***: Fornece uma Interface Gráfica do Usuário (do inglês, *Graphical User Interface (GUI)*). Com uma GUI o uso do software se torna mais simples, atingindo um número maior de usuários.
- ***Gratuita***: Uso e disponibilização da ferramenta de forma gratuita.
- ***Multiplataforma***: A ferramenta pode ser executada em mais de uma plataforma, como por exemplo Windows, Linux e macOS. Uma ferramenta multiplataforma atinge mais usuários, traz flexibilidade e mobilidade no seu uso.
- ***Linguagem de programação***: Linguagem de programação utilizada na implementação do software. Este atributo ajuda pessoas interessadas em desenvolver uma ferramenta de anonimização na tomada de decisão de qual linguagem utilizar. Se uma linguagem é mais utilizada dentre as ferramentas pode ser um sinal de um suporte maior a esta finalidade.
- ***Anonimização***: Utilização de técnicas de anonimização ou modelos de privacidade para tornar os dados anônimos, como generalização, K-anonimato, L-diversidade e

supressão. Nesta característica a pseudonimização e o modelo de privacidade diferencial não foram consideradas, visto que estas técnicas não transformam os dados em busca do anonimato de fato.

- ***Análise de risco***: Possibilidade de analisar o risco de reidentificação dos dados após o processo de anonimização.
- ***Disponível***: Disponibilidade da ferramenta para utilização, seja ela de forma gratuita ou comercial. Existem ferramentas que foram desenvolvidas apenas para pesquisas e cenários específicos, não estando disponível para o uso do público.

Tabela 1: Principais características das ferramentas de desidentificação apresentadas neste estudo.

Nome da Ferramenta	Data da última versão	Open source	GUI	Gratuita	Multiplataforma	Linguagem de programação	Anonimização	Análise de risco	Disponível
Aircloak Insights	✣	☆	★	☆	★	Ruby[a]	★	☆	★
Amnesia	04/2022	★	★	★	★	Java	★	☆	★
Anonimatron	07/2021	★	☆	★	★	Java	☆	☆	★
ARX	01/2021	★	★	★	★	Java	★	★	★
BizDataX	✣	☆	★	☆	☆	✣	★	☆	★
CHORUS	06/2021[b]	★	☆	★	★	Scala	☆[c]	☆	★
CloverDX	06/2022	☆	★	☆	★	✣	★	☆	★
Cornell Anonymisation Toolkit	2009[d]	★	★	★	★	C++	★	★	★
Docbyte's Real-time Automated Anonymization	✣	☆	✣	☆	✣	✣	★	☆	☆
g9 Anonymizer	12/2020	☆	★[e]	☆	★	Java	★	☆	★
Gramener's Data Anonymization Solution	✣	☆	✣	☆	✣	✣	★	☆	☆
μ-ARGUS	01/2022	★	★	★	★[f]	Java	★	★	★
NLM-Scrubber	04/2019	☆	★	★	★	Perl	☆	☆	★
sdcMicro	08/2022	★	★	★	★	R	★	★	★
SECRETA	2014[d]	☆	★	★	★	C++	★	☆	★
TIAMAT	2009[d]	☆	★	★	★	Java	★	☆	★
UTD Anonymization Toolbox	03/2012	★	★	★	★	Java	★	☆	★

★ Apresenta a propriedade.
✣ Não foi encontrado ou não se aplica.
[b] Última atualização do código no seu repositório Git.
[d] Ano de publicação do artigo que expõe a ferramenta.
[f] Versões mais recentes disponibilizadas apenas para plataforma *Windows*.
☆ Não apresenta a propriedade.
[a] De acordo com (Zuo et al. 2021).
[c] Oferece a técnica de privacidade diferencial.
[e] Através da IDE Eclipse.

3.2 Projetos que apresentam soluções interessantes de desidentificação de dados

Esta seção reúne um conjunto de projetos de pesquisa que apresentam processos de desidentificação arquiteturalmente interessantes e que forneceram boas ideias para a arquitetura proposta.

3.2.1 SOP HARMONY

A HARMONY *Alliance* é uma iniciativa que apoia projetos que utilizam tecnologias de *Big Data* para melhorar o tratamento dos diversos tipos de câncer sanguíneo (*blood cancer*) (HARMONY 2022). Para dar suporte a estes projetos e permitir a coleta e análise de dados em uma escala tão ampla e internacional, a HARMONY *Alliance* desenvolveu a HARMONY *BigData Platform* que implementa um conceito de anonimização "de fato" (Butler et al. 2018), que corresponde a um procedimento de pseudonimização em duas etapas que transforma os dados antes que sejam armazenados na plataforma HARMONY, sendo a primeira junto ao Provedor de Dados e a segunda numa entidade honesta ou terceira parte confiável, para só então serem ingeridas pela plataforma HARMONY. A anonimização em duas etapas utilizada na arquitetura proposta foi diretamente influenciada pelo projeto SOP HARMONY.

3.2.2 OpenAIRE

O *Open Access Infrastructure for Research in Europe (OpenAIRE)* (OpenAIRE 2022) é uma organização sem fins lucrativos que apoia a pesquisa europeia. O catálogo de serviços oferecidos pela organização é grande, contendo mais de 10 serviços. Dentre estes serviços está o Amnesia, uma ferramenta para anonimização de dados disponibilizada como um uma aplicação desktop e também como uma API REST, que possibilita a sua utilização como um auxílio à criação de novas aplicações de desidentificação. A API fornece um conjunto completo de endpoints para o processo de anonimização, abrangendo etapas de configuração e análise dos dados, como a definição de hierarquias e geração de estatísticas dos atributos quasi-identificadores. O projeto OpenAIRE e, particularmente a solução Amnesia, serviram de inspiração para o funcionamento baseado em APIs que orientou o desenvolvimento da arquitetura proposta.

3.2.3 Anonymization as a Service - OsloMet

A *Anonymization as a Service (AaaS)*, do português, Anonimização como Serviço, é um projeto de tese de bacharelado concluído na *Oslo Metropolitan University (OsloMet)* em colaboração com a *Norwegian Labour and Welfare Administration* (NAV) que oferece a anonimização como serviço (Groseth et al. 2019). A solução criada utiliza a biblioteca ARX[1] para a funcionalidade de anonimização. O projeto é *open source* e todos os seus produtos são distribuídos sob a *MIT License*. O ecossistema da Anonimização como Serviço é composto por três produtos, o ARX as a Service (ARXaaS), peça central do microsserviço, e o PyARXaaS e WebARXaaS, clientes que fornecem interfaces de usuário, python e web, para o ARXaaS. O ARXaaS é um serviço web implementado em Java utilizando o *framework Java Spring Boot*. A biblioteca do ARX é utilizada para implementar as funcionalidades nas áreas de análise de risco e anonimização do conjunto de dados e geração de hierarquias de generalização que são aplicadas na anonimização. O serviço disponibiliza uma API RESTful, com suporte ao *stand-alone HTTPS*, para que os aplicativos clientes usufruam de suas funcionalidades. O ARXaaS serviu de inspiração para a implementação da plataforma em múltiplos aspectos, sendo os mais relevantes as tecnologias empregadas (Java e Spring Boot), uso de APIs e uso da biblioteca ARX.

[1] A biblioteca *open source* ARX Java fornece funcionalidades de anonimização de dados para outros sistemas de software e está disponível para download em: https://arx.deidentifier.org/downloads/

4 Arquitetura Proposta

Neste capítulo, é apresentada e discutida a arquitetura proposta. São discutidas as principais decisões que moldaram o desenho da arquitetura. A seguir, são detalhados os principais componentes, responsabilidades e fluxos. Por fim, é discutido o estado atual da implementação da arquitetura.

4.1 Decisões de Projeto

Nesta seção são apresentadas e discutidas as principais decisões que influenciaram o desenho da arquitetura, incluindo a implementação dos componentes e fluxos principais, com foco nas estratégias implementadas visando alcançar a segurança e privacidade dos dados.

4.1.1 (Pseudo)Anonimização na fonte

Uma das primeiras decisões que influenciaram o desenho da arquitetura, foi a opção por anonimizar os dados na fonte, ou seja, ainda no provedor dos dados. O objetivo foi garantir que somente dados (pseudo)anonimizados sejam enviados para fora do provedor, em conformidade com o que é definido na LGPD, para o caso do Brasil, em que a legislação brasileira recomenda a anonimização dos dados sempre que possível na realização de estudos por orgãos de pesquisa (inciso IV do artigo 7) e estudos em saúde (artigo 13), assim como em muitas outras regulamentações de privacidade de dados mundo afora. Esta decisão é manifestada no componente Local Broker da arquitetura proposta.

4.1.2 Dois níveis de (pseudo)anonimização

A decisão de aplicar dois níveis de (pseudo)anonimização foi inicialmente inspirada no modelo apresentado em (Moor, Claerhout e Meyer 2003), que utiliza uma primeira

(pseudo)anonimização na fonte ou pré-pseudonimização, seguida da geração de um novo pseudônimo no nível do TTP (*Trusted Third Party*). No desenho da arquitetura aqui proposta, o papel do TTP é desempenhado pelo Honest Broker de forma similar. No entanto, para a (pseudo)anonimização na fonte, diferente de (Moor, Claerhout e Meyer 2003) que utiliza a mesma função de geração do pseudônimo nas diferentes fontes de dados e depois criptografa-os para envio ao TTP, aqui é utilizada uma função *hash* com um *salt* randômico (criado na inicialização do Local Broker) que, em seguida, é encriptado para o Honest Broker. O uso da função *hash* tem o objetivo de evitar que o Honest Broker seja capaz de inferir o valor original do identificador, enquanto que o uso do *salt* permite que um individuo tenha pseudônimos diferentes em diferentes provedores. Essa camada adicional de segurança, visa garantir que, mesmo quando do comprometimento dos dados enviados ao Honest Broker, um provedor não será capaz de identificar um indivíduo a partir dos dados de outro provedor. Esta estratégia de dupla anonimização é também utilizada no projeto SOP HARMONY (Butler et al. 2018), descrito na seção 3.2, que também faz uso de uma primeira (pseudo)anonimização no lado do provedor e uma outra no lado do TTP. Esta decisão é manifestada no componente intitulado como Honest Broker na arquitetura proposta.

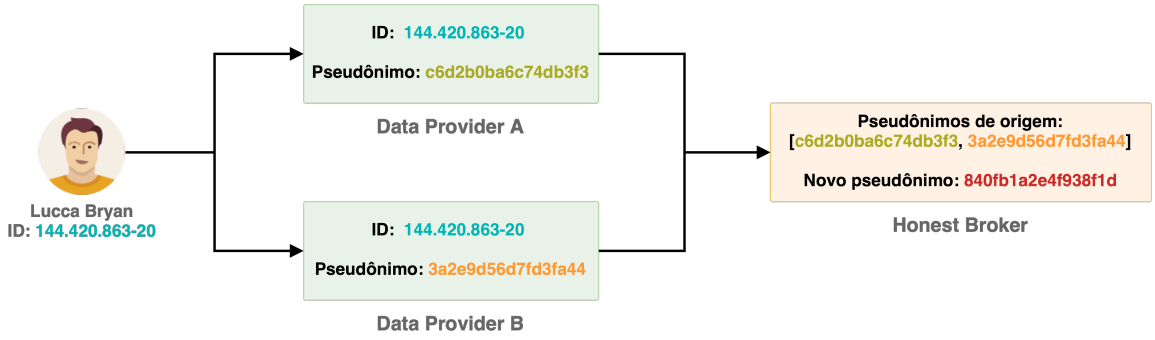


Figura 11: Exemplo do esquema de anonimização em dois níveis na arquitetura proposta.

Um exemplo deste processo é ilustrado na Figura 11, onde dados referentes ao mesmo indivíduo "Lucca Bryan" estão presentes nos Data Providers A e B. Embora os dados sejam do mesmo indivíduo, no primeiro nível de (pseudo)anonimização são gerados diferentes pseudônimos em cada um dos provedores. Apenas no Honest Broker, ou seja, no segundo nível de (pseudo)anonimização, ocorre a ligação dos dados (pseudo)anonimizados, a partir da geração de um pseudônimo único para a arquitetura proposta.

4.1.3 Definição dos conjuntos de dados no Data Lake

A fim de uniformizar o processo de (pseudo)anonimização a partir de múltiplas fontes, optou-se por manter a definição dos parâmetros a serem utilizados de forma centralizada no Data Lake. Assim, para cada conjunto de dados a ser alimentado no Data Lake, é criado um *template* com as definições estruturais do conjunto de dados e dos parâmetros a serem usados no processo de (pseudo)anonimização, como discutido na Seção 4.2.3. Essa estratégia também tem impacto na redução dos esforços de harmonização, conversão e categorização dos dados recebidos.

4.2 Arquitetura proposta

A Figura 12 apresenta uma visão geral da arquitetura proposta.

De uma forma geral, a arquitetura é composta por três elementos principais: o Data Provider (DP), o Honest Broker (HB) e o Data Lake (DL). O Data Provider é o responsável pela coleta e uso primário dos dados, e é uma fonte de dados a serem compartilhados na arquitetura, como por exemplo o SigSaúde. É ainda neste elemento que o processo de (pseudo)anonimização tem início, juntamente com a análise de risco dos dados a serem publicados. A arquitetura suporta múltiplos provedores simultaneamente. O Honest Broker é a entidade considerada honesta por todos as entidades envolvidas (Data Providers e responsável pelo Data Lake) e a sua responsabilidade é realizar uma nova etapa de pseudonimização que garante a ligação entre os dados (pseudo)anonimizados de diferentes fontes. O Data Lake é o componente responsável pelo armazenamento dos dados para uso secundário e sua disponibilização para pesquisadores, cientistas e demais público alvo.

As seções seguintes apresentam em detalhe cada um dos elementos da arquitetura, incluindo os fluxos de informação na comunicação entre estes.

4.2.1 Data Provider

O Data Provider (DP) é o responsável por coletar, armazenar e fazer uso primário dos dados pessoais no contexto desta dissertação, sendo como o próprio nome diz, o provedor de dados. Na prática, o Data Provider é um hospital, clínica, clínica escola ou qualquer outra instituição prestadora de serviços de saúde, como por exemplo o SigSaúde. Seu papel na arquitetura proposta é fornecer dados clínicos, incluindo dados pessoais dos pacientes, para alimentar o Data Lake com o objetivo de viabilizar pesquisas ou até mesmo monetizar

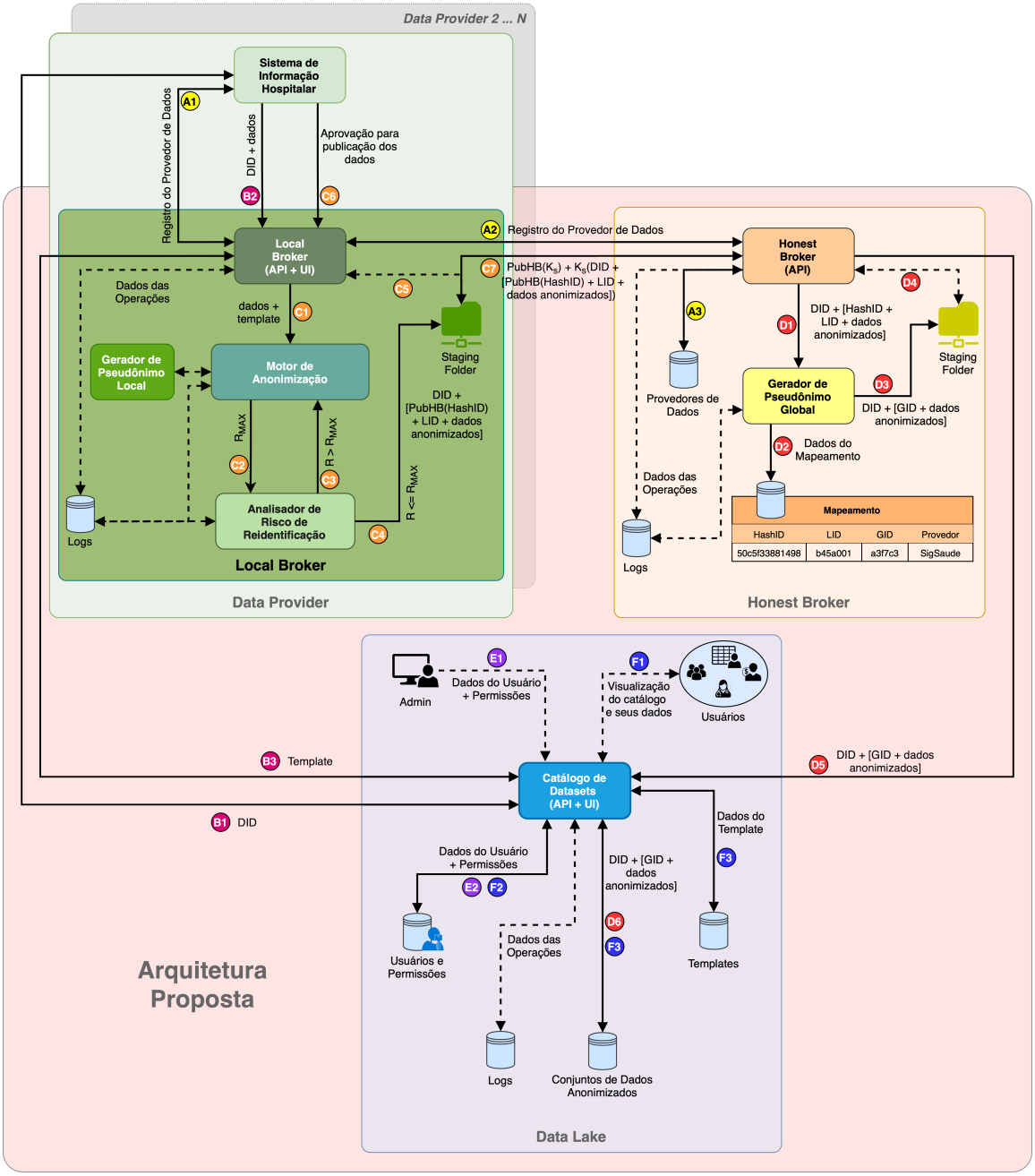


Figura 12: Visão geral da arquitetura proposta. Fluxos de dados: A1..AN: Registro de um Data Provider; B1..BN: Preparação dos dados no provedor; C1..CN: (Pseudo)anonimização dos dados na fonte; D1..DN: Ligação e publicação dos dados anonimizados; E1..EN: Gerenciamento de usuários e permissões; F1..FN: Acesso ao catálogo de conjuntos de dados.

a informação - cabe aqui salientar que o objetivo do uso secundário da informação não é tratado nesta dissertação.

Toda a interação do Data Provider com a arquitetura é realizada através do componente Local Broker (LB). Assim, cada provedor tem uma instância LB sendo executada

em seu ambiente. O LB fornece uma API e uma interface do usuário para sua utilização, a fim de ocultar todo o processo de (pseudo)anonimização dos dados e envio ao Data Lake. O LB é acessível apenas por seu Data Provider, e se comunica apenas com o Honest Broker, não sendo necessária sua exposição a qualquer outro elemento.

É no LB que ocorre a (pseudo)anonimização na fonte, ou seja, no Data Provider. Com isso, medidas adicionais de controle e segurança podem ser colocadas em ação para auxiliar na privacidade do processo, conforme exigido na legislação. Este é, na verdade, o primeiro nível de anonimização em dois níveis proposta na arquitetura, sendo então gerado um pseudônimo único para cada indivíduo (paciente), chamado de pseudônimo local (LID). Os dados do mapeamento entre o LID e o(s) atributo(s) identificador(es) são armazenados no provedor, ficando longe dos dados (pseudo)anonimizados - como manda a lei - que serão enviados ao Data Lake. Este mapeamento é usado em situações específicas onde é necessário reidentificar um indivíduo a partir de dados (pseudo)anonimizados, por exemplo, num estudo clínico, a descoberta e desenvolvimento de uma cura para uma determinada doença pode exigir a reidentificação do grupo de indivíduos que participaram do ensaio para que estes possam ser devidamente tratados.

O envio de dados sensíveis entre LB e o HB é sempre criptografado com o uso de chaves simétricas e assimétricas. Visando trazer mais segurança, a comunicação entre os componentes, após o registro do provedor, se faz necessário o uso da autenticação. A fim de auxiliar a auditoria, informações acerca de todos os fluxos executados no Data Provider, são devidamente armazenadas em log.

4.2.2 Honest Broker

O Honest Broker (HB) é o componente da arquitetura que deve ser executado em uma entidade considerada honesta - ou seja, que irá processar os dados de forma correta, idônea e segura - por todas as entidades envolvidas. Na prática, um HB deve ser executado em instituições que passam confiança e credibilidade à sociedade, como universidades, centros de pesquisa, entidades governamentais, etc. No HB são realizados dois processos gerais: o registro do Data Provider e a ligação dos dados (pseudo)anonimizados, através da geração de um novo pseudônimo, o pseudônimo global (GID). A lógica utilizada para a geração do novo pseudônimo deve garantir que registros provenientes de um mesmo titular devem conter o mesmo pseudônimo global, independente de sua origem. Para isto, o LB envia o valor de hash criptografado com a chave pública do HB, referente ao atributo identificador do indivíduo, visto que, o pseudônimo local é diferente em cada provedor. A geração do

GID representa o segundo nível de anonimização da arquitetura proposta. Importante ressaltar que o HB se comunicará apenas com o Local Broker dos provedores e o Data Lake, não permitindo a solicitação de envio de dados de fontes não registradas.

Assim como no LB, o mapeamento entre os pseudônimos local e global, são armazenados no HB, ficando longe dos dados (pseudo)anonimizados - como determina a legislação vigente - que serão enviados ao Data Lake. Assume-se ainda que o HB implementará um conjunto complementar de mecanismos técnicos e organizacionais de segurança. Este mapeamento é também usado em situações específicas onde é necessário reidentificar um indivíduo a partir de dados (pseudo)anonimizados. Diferente da comunicação entre LB e HB, o HB irá comunicar com o Data Lake somente para o envio dos dados, que já estarão anonimizados e prontos para publicação e, portanto, esta comunicação não precisa ser criptografada. A auditoria é também auxiliada com o armazenamento em log das informações acerca de todos os fluxos executados no HB.

4.2.3 Data Lake

O Data Lake (DL) é o repositório dos dados (pseudo)anonimizados e, na prática, pode ser implementado por diferentes entidades públicas (por exemplo, universidades, centros de pesquisa e órgãos governamentais ou privadas (por exemplo, empresas especializadas na comercialização de dados ou informações). É através deste elemento que é realizado o acesso aos dados (pseudo)anonimizados, disponibilizados pelos Data Providers. Os dados são organizados em conjuntos de dados, definidos a partir de *templates*. A cada conjunto de dados é atribuído um identificador único chamado *Dataset ID* (DID). O acesso ao Data Lake poderá ser realizado através de API ou interface de usuário, sendo possível consultar o catálogo de conjunto de dados disponíveis, adicionar um *template*, além de gerenciar os usuários e suas permissões.

Um *template* para um conjunto de dados é composto pelos seguintes campos:

- Nome para o conjunto de dados.
- Descrição da finalidade conjunto de dados.
- Organização responsável pelo conjunto de dados.
- Informações de contato.
- Atributos do conjunto de dados, incluindo obrigatoriamente o nome, o tipo (identificador, quase-identificador, sensível ou insensível) e a sua hierarquia se existir.

- Dados sobre os modelos de privacidade a serem aplicados e as suas parametrizações, por exemplo o valor de k para o uso do k-anonimato.
- Valor máximo de supressão de dados admitido.
- Valor do risco máximo de reidentificação aceito.

O acesso ao DL deverá ser autenticado e auditável. Seguindo a mesma estratégia aplicada a outros elementos da arquitetura, informações acerca de todos os fluxos executados no Data Lake serão devidamente armazenadas em log.

4.2.4 Fluxos

Após a apresentação dos componentes gerais da arquitetura, nesta seção, são apresentados os principais fluxos entre eles. Em cada fluxo, a fim de facilitar o entendimento, é apresentada uma ilustração contendo apenas os elementos que participam do fluxo, ou seja, está imagem é um recorte da figura da arquitetura geral.

A. Registro de Data Provider

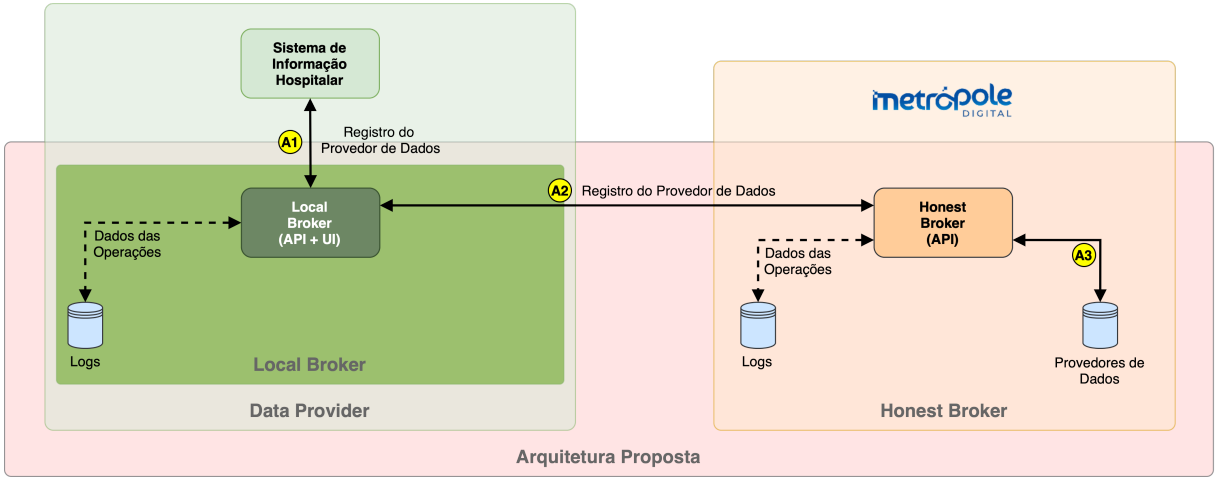


Figura 13: Fluxo de registro de Data Providers.

O Data Provider que desejar enviar dados ao Data Lake, deve, antes, efetuar o seu registro junto ao Honest Broker, através de seu Local Broker, como ilustrado na Figura 13. Para isso, o provedor deve enviar os dados necessários ao registro para o seu Local Broker (A1). O Local Broker, ao receber a solicitação, envia uma mensagem de intenção de registro ao Honest Broker [A2.1]. Em resposta, o Honest Broker envia a sua chave pública

(PubHB) [A2.2]. Em seguida, o Local Broker envia os dados para registro simetricamente criptografados com uma chave de sessão K_s, que segue criptografada com a PubHB, para o Honest Broker [A2.3]. A cada comunicação deve ser gerada e utilizada uma nova chave de sessão, sendo assim, cada chave K_s será utilizada uma única vez.

Ao receber as informações necessárias para o registro, o Honest Broker descriptografa, valida e armazena os dados recebidos em uma base de dados exclusiva para o registro de Data Providers (A3). Essa base de dados é utilizada para identificação e validação da origem dos dados recebidos. Para finalizar o fluxo, o Honest Broker devolve uma mensagem a indicar que o Data Provider está devidamente registrado e apto a enviar dados para o Data Lake, ou uma mensagem de erro em caso de falha no processo de registro [A2.4]. Todo o processo de troca de mensagens e validações são devidamente auditados em um serviço de log. O registro do Data Provider deve ser o primeiro fluxo a ser executado pelo provedor ou ficará impedido de executar qualquer outra operação na arquitetura. Além disto, é importante ressaltar que deve existir um processo de solicitação, seja ele burocrático ou não, para participar da plataforma antes de o provedor ser capaz de instanciar um Local Broker e solicitar o registro junto a ele.

B. Preparação dos Dados no Provedor

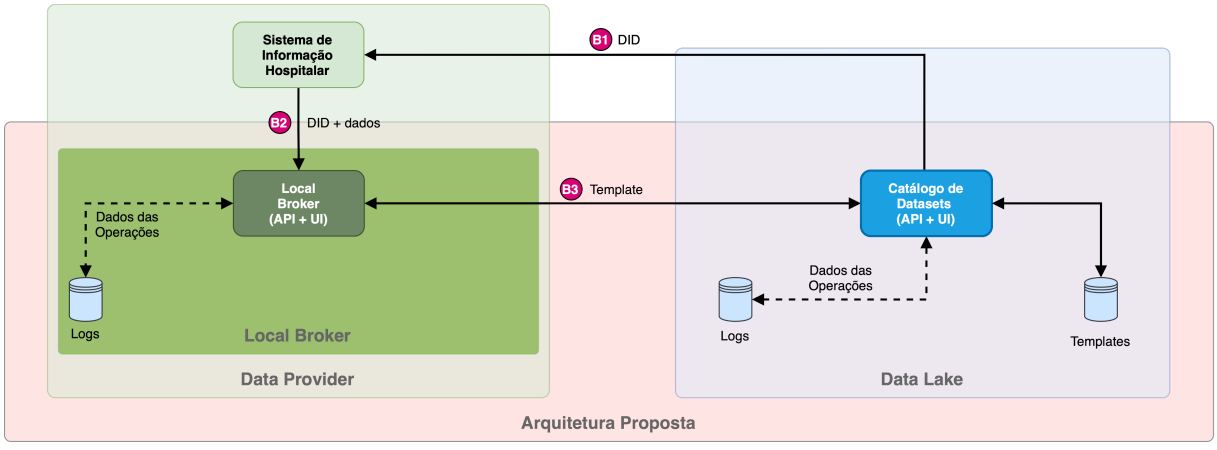


Figura 14: Fluxo de envio de dados.

Como já descrito na Seção 4.2.3, o Data Lake reúne um catálogo de conjuntos de dados, onde cada conjunto de dados é identificado pelo seu *Dataset ID* ou DID. Assim, para iniciar a preparação dos dados a serem (pseudo)anonimizados, o Data Provider deve consultar o catálogo - por exemplo, utilizando a Interface do Usuário (UI) (do inglês *User Interface*) do Data Lake - a fim de obter o DID correspondente ao conjunto de dados para o qual deseja enviar os dados (B1), como ilustrado na Figura 14.

De posse do DID, o provedor deve organizar os dados que serão enviados de acordo com as definições (colunas, tipos, etc) no Data Lake. Em seguida, deve enviar os dados e o DID para o seu Local Broker (B2). A fim de preparar os dados para o processo de (pseudo)anonimização, o Local Broker então deve obter o *template* correspondente ao DID recebido (B3). Cabe lembrar que o *template* traz as informações necessárias ao processo de (pseudo)anonimização, como os modelos de privacidade a serem aplicados e as suas parametrizações, hierarquias a serem aplicadas, etc. De posse do *template* correto, o Local Broker valida os dados repassados pelo provedor.

C. (Pseudo)Anonimização dos Dados

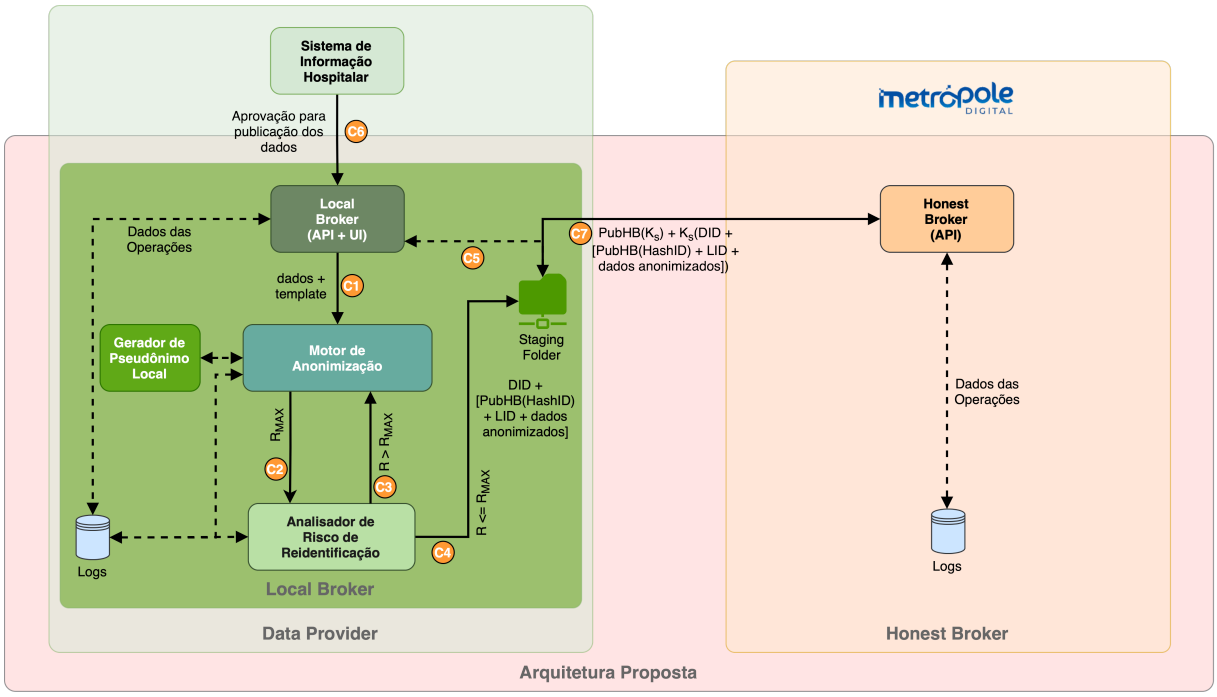


Figura 15: Fluxo de anonimização dos dados.

O Local Broker dá início ao processo de (pseudo)anonimização com o envio dos dados e do *template* a ser aplicado para o Motor de Anonimização (C1), como ilustrado na Figura 15.

O primeiro passo do processo de (pseudo)anonimização é a geração dos pseudônimos locais (LIDs). Em seguida, é executado um *loop* de anonimização, onde os dados são transformados pelo Motor de Anonimização, segundo as parametrizações descritas no *template*, e submetidos ao Analisador de Risco de Reidentificação (C2) que irá verificar se o risco de reidentificação para os dados transformados é igual ou inferior ao R_{MAX} (definido no *template*). Caso o risco de reidentificação calculado ainda seja superior ao R_{MAX}, o Motor

de Anonimização é acionado para otimizar os parâmetros a fim de reduzir ainda mais o risco (C3). São realizadas sucessivas otimizações, até que o risco de reidentificação dos dados transformados seja igual ou inferior ao R_{MAX} e, só então, são enviados para o *Staging Folder* o DID e os dados transformados, indicando que estão prontos para envio (C4). Cada registro dos dados transformados inclui: o *hash* do atributo identificador, criptografado com a chave pública do Honest Broker; o pseudônimo local (LID); e o conjunto de dados anonimizados. Periodicamente, o Local Broker verifica se existem dados a serem enviados, informando ao Data Provider (C5). Assincronamente, o provedor visualiza os dados anonimizados pendentes de envio, podendo aprovar a publicação dos mesmos (C6). Caso não aprove o envio, os dados serão eliminados e não serão publicados. Ao receber a aprovação para publicação dos dados, o Local Broker realiza a criptografia simétrica dos dados (DID, hash dos identificadores e os dados anonimizados) com uma chave de sessão K_s, que é também enviada, porém criptografada com a chave pública do HB (C7).

D. Ligação e Publicação dos Dados Anonimizados

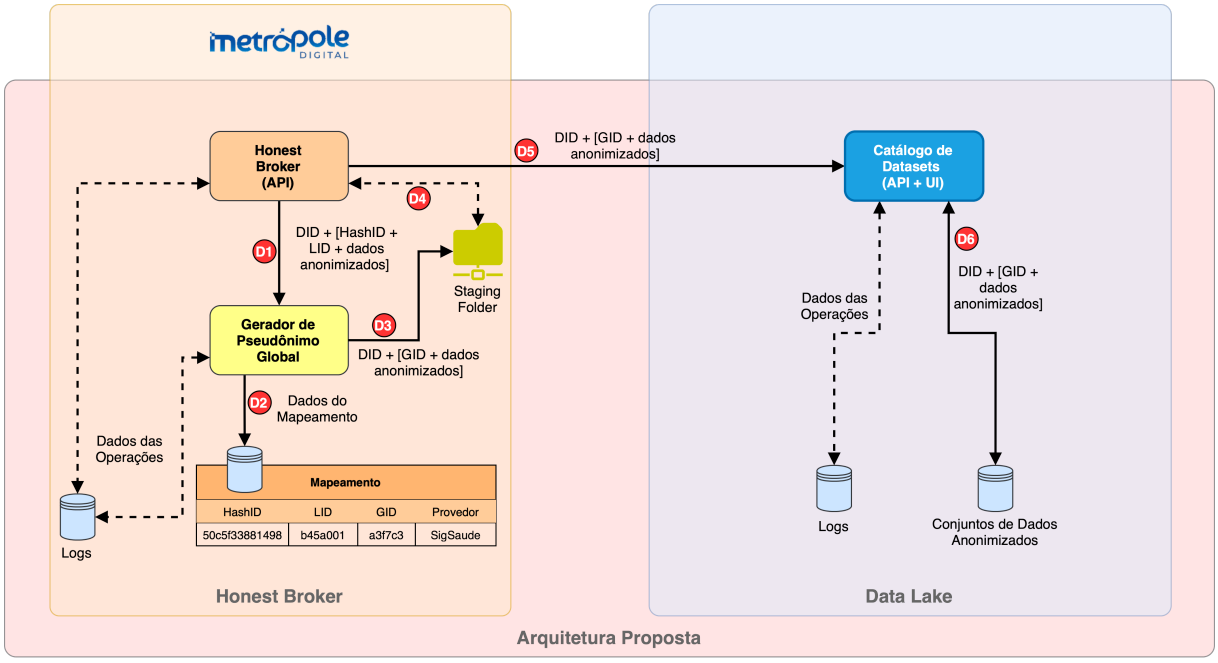


Figura 16: Fluxo de ligação e publicação dos dados anonimizados.

Como descrito na Seção 4.2.2, cabe ao Honest Broker ligar os dados anonimizados de diferentes provedores, através do mapeamento de pseudônimos locais (LID) para um pseudônimo global (GID). Este processo tem início com o processamento dos dados (pseudo)anonimizados recebidos do Data Provider (D1), como ilustrado na Figura 16.

Este processamento implica na validação e geração do GID para cada registro no conjunto de dados (pseudo)anonimizado. Para a geração do GID, em cada registro, o *hash* do identificador do registro é descriptografado e então é verificada a sua existência no banco de dados que contêm o mapeamendo de LID para GID (D2). Caso o *hash* não exista, é gerado um novo GID para o registro em questão, sendo então inserido na base de dados, juntamente com o *hash*, o LID e a identificação do provedor. Caso já exista, é realizada uma verificação adicional sobre a lista de registros retornada em busca do LID a ser mapeado. Caso não seja encontrado, uma nova entrada é criada com o GID encontrado, o *hash*, o LID e a identificação do provedor. Caso já exista, nada é feito. Após o processamento do conjunto de dados e geração dos pseudônimos globais, estes dados são disponibilizados em uma *Staging Folder* para envio ao Data Lake (D3). Periodicamente, o Honest Broker verifica a existência de dados pendentes de envio no *Staging Folder* (D4), enviando-os para o Data Lake (D5). Ao receber os dados, o Data Lake os armazena em uma base de dados dedicada (D6). Todo o processo é devidamente registrado em logs.

E. Gerenciamento de Usuários

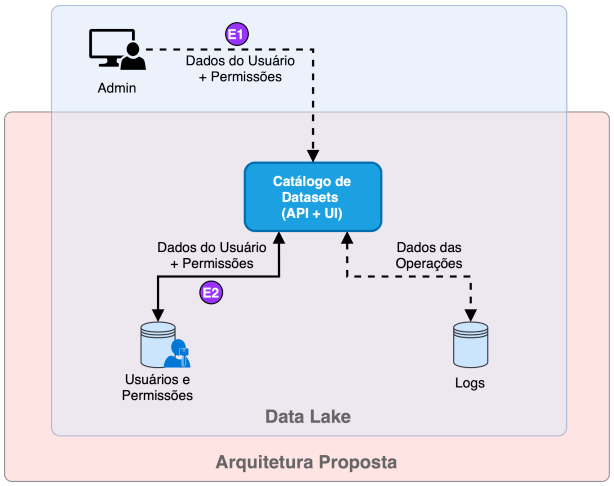


Figura 17: Fluxo de gerenciamento de usuários.

Embora os dados armazenados no Data Lake sejam considerados (pseudo)anonimizados, ainda assim, na arquitetura proposta, prevê-se que o acesso ao Data Lake seja restrito a usuários devidamente autenticados e autorizados. Assim, de forma simplificada, assume-se um usuário administrador que será responsável pelo cadastro e definição de permissões dos diferentes utilizadores do Data Lake (E1), como ilustrado na Figura 17. Estes dados e permissões serão armazenadas numa base de dados (E2) que será utilizada para garantir a segurança no acesso e execução das diferentes

operações do Data Lake. Inicialmente, são previstos três tipos de usuário:

Visualizador. Usuário interessado em visualizar o catálogo de conjunto de dados do Data Lake, com acesso limitado apenas aos dados descritivos contidos nos *templates*.

Provedor. Usuário que representa um Data Provider, com acesso total aos dados dos *templates*.

Injetor. Usuário responsável por criar *templates* no catálogo.

F. Acesso ao Data Lake

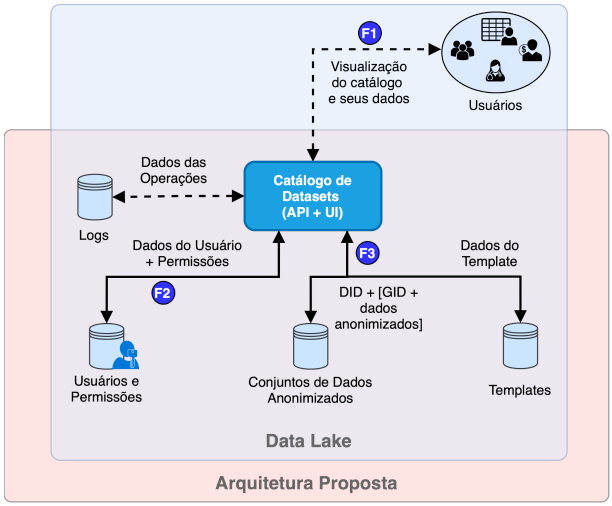


Figura 18: Fluxo de acesso ao Data Lake.

Como descrito anteriormente, o acesso autenticado e autorizado de usuários ao Data Lake pressupõe que os pedidos de acesso (visualização, inserção, alteração) (F1) sejam devidamente autenticados e autorizados (F2), a fim de limitar o acesso aos dados (F3) dos *templates* referentes aos diferentes conjuntos de dados armazenados no Data Lake. Este fluxo é ilustrado de forma simplificada na Figura 18.

4.3 Especificação e desenvolvimento da Prova de Conceito (PoC)

Um dos maiores desafios para a conclusão deste trabalho foi a implementação e validação de todos os componentes e a comunicação entre eles atempadamente. Embora as

principais funcionalidades propostas foram implementadas, em alguns casos, foram realizadas simplificações para permitir exercitar todo o conjunto de operações inicialmente definido. Assim, nesta seção, é apresentado e discutido o protótipo desenvolvido para servir como prova de conceito para a implementação dos componentes e fluxos previstos.

4.3.1 Tecnologias relevantes

A implementação dos componentes da arquitetura foi preparada para ser executada usando *containers* Docker. Assim, cada elemento tem a sua imagem e é instanciada como um *container* Docker. Além das vantagens gerais de utilização de *containers* - como agilidade, portabilidade, escalabilidade, compatibilidade entre plataformas, desempenho, segurança e outras (AlSalamah, Cámara e Kelly 2018, Bellavista e Zanni 2017) - a virtualização foi fundamental para os testes e validação, permitindo a instanciação de múltiplos Data Providers. Além disto, a execução dos componentes em *container* permite que a arquitetura seja executada em diferentes plataformas.

Os componentes foram desenvolvidos em linguagem de programação Java em sua versão 19, com o framework Spring Boot versão 2.7.4.

4.3.2 Geração de logs

Embora seja um dos controles importantes para este tipo de solução, o mecanismo de log, de forma estruturada e persistente, não foi implementado. Como a arquitetura é modular, pode ser acoplado um serviço de log ou *audit trail* para auxiliar a auditoria das operações realizadas. É importante lembrar que os logs devem ser gerados de forma padronizada, facilitando a sua integração com ferramentas como Fluentd[1], Filebeat[2] ou Logstash[3]. Poderia ainda ser acoplado uma ferramenta de visualização dos logs como Prometheus[4], Kibana[5] e Grafana[6].

[1]https://www.fluentd.org/
[2]https://www.elastic.co/pt/beats/filebeat
[3]https://www.elastic.co/pt/logstash/
[4]https://prometheus.io
[5]https://www.elastic.co/pt/kibana/
[6]https://grafana.com

4.3.3 Autenticação e Autorização

Outro controle importante, mas que, por restrições de tempo, não foi implementado foi a Autenticação e Autorização de usuários e APIs. Aqui, novamente tirando proveito da modularidade da arquitetura, é possível pensar em acoplar um gerenciador de identidade e acesso como o Keycloak[7].

4.3.4 Data Provider

A implementação do Data Provider foi centrada no componente Local Broker, sendo implementado como uma API REST para sua utilização, como ilustrado na Figura 19, sendo disponibilizados os seguintes endpoints:

POST /provider/register	Utilizado para registrar o Data Provider.
POST /dataset/send	Utilizado para enviar o conjunto de dados para publicação.
GET /dataset/pending/list	Utilizado para verificar os conjunto de dados que foram processados e estão pendentes de análise.
GET /dataset/pending/{id}	Utilizado para visualizar o conjunto de dados pendente de análise por seu ID.
POST /dataset/pending/update	Utilizado para aprovar ou cancelar os conjuntos de dados pendentes de análise.

[7] https://www.keycloak.org/

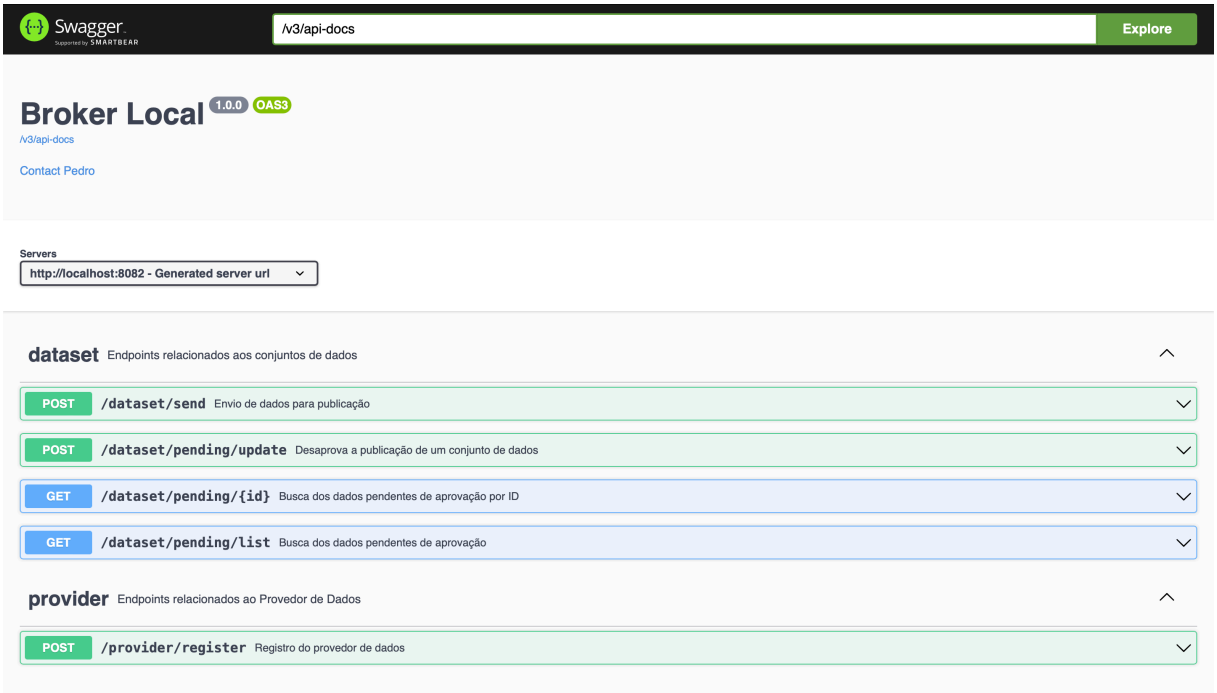


Figura 19: Tela principal do Swagger do Local Broker.

Os dados enviados do Local Broker ao Honest Broker são criptografados utilizando uma estratégia mista de criptografia simétrica e assimétrica. Na troca de mensagens, é aplicada a criptografia simétrica com chave de sessão K_s, gerada randomicamente para cada sessão de comunicação, que então segue criptografada com a chave pública do Honest Broker, de forma a garantir que os dados serão apenas descriptografados pelo HB, com sua chave privada.

O mapeamento entre pseudônimo local e o atributo identificador do indivíduo é persistido em um banco de dados, utilizando a tecnologia MongoDB[8]. O armazenamento dos conjuntos de dados original e (pseudo)anonimizado é realizado localmente em um diretório parametrizável, representando o *Staging Folder* da arquitetura. A implementação divide o *Staging Folder* em quatro subdiretórios: *forapproval* - destinado aos conjuntos de dados pendentes de análise; *tosend* - destinado aos conjuntos de dados pendentes de envio ao Honest Broker; *cancelled* - destinado aos conjuntos de dados que tiveram sua publicação cancelada/não aprovada; e *sent* - destinado aos conjuntos de dados que já foram enviados ao Honest Broker para publicação.

Por fim, o componente em seu estado atual de implementação é capaz de exercer suas principais funções na arquitetura. Através dele é possível realizar o registro do Data

[8]https://www.mongodb.com/home

Provider, iniciar o processo de anonimização em dois níveis, enviar conjuntos de dados para publicação e efetuar a análise dos conjuntos de dados por parte do Data Provider antes de sua publicação.

4.3.5 Honest Broker

A implementação realizada do Honest Broker, dispõe de uma API REST para sua utilização, como ilustrado na na Figura 20, sendo disponibilizados os seguintes endpoints:

GET /provider/register	Utilizado para solicitar a chave pública do Honest Broker.
POST /provider/register	Utilizado para registrar o Data Provider.
POST /dataset/send	Utilizado para enviar o conjunto de dados.

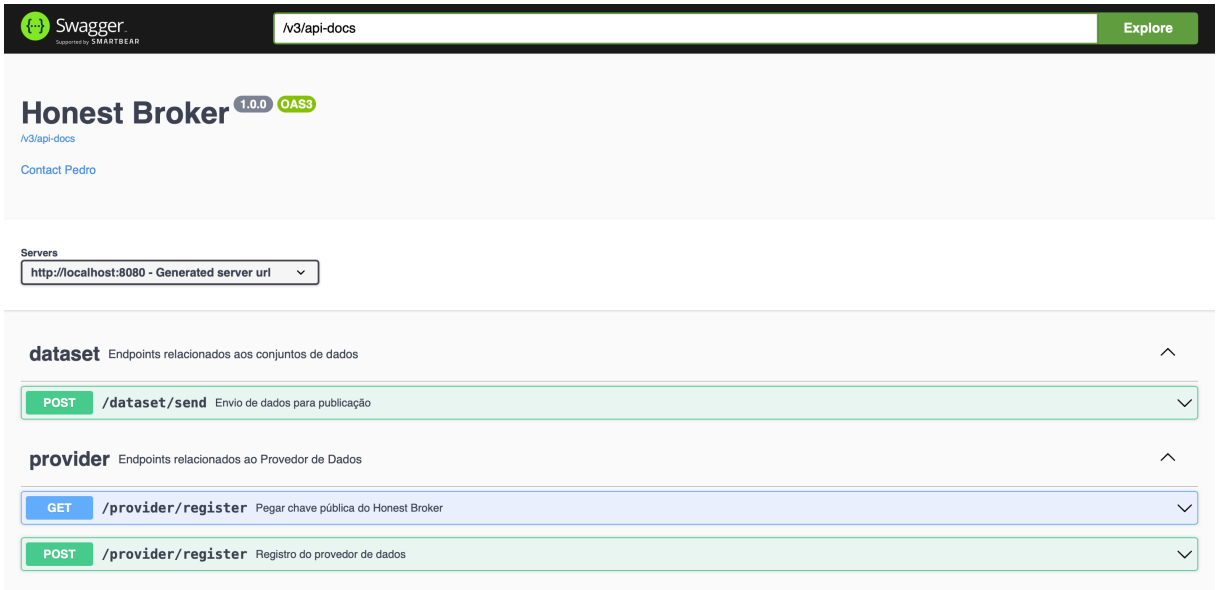


Figura 20: Tela principal do Swagger do Honest Broker.

Como comentado na Subseção 4.3.4, a comunicação entre os componentes Local Broker e Honest Broker é protegida por criptografia. Já a comunicação do Honest Broker com o Data Lake, uma vez que envolve dados já (pseudo)anonimizados e, portanto, não sensíveis, é realizada sem uso de criptografia.

De forma similar ao Local Broker, o mapeamento entre pseudônimo global e o pseudônimo local juntamente com sua origem e o registro dos Data Providers é persistido em

uma base de dados na tecnologia MongoDB. Também de forma similar ao Local Broker, os conjuntos de dados são armazenados localmente em um diretório parametrizável, representando a *Staging Folder* descrito na arquitetura. No entanto, aqui, este diretório é dividido em apenas dois subdiretórios: *tosend* - destinado aos conjuntos de dados pendentes de envio ao Data Lake para publicação; e *sent* - destinado aos conjuntos de dados que já foram enviados ao Data Lake.

Por fim, o componente em seu estado atual de implementação é capaz de exercer suas principais funções na arquitetura. Através dele é possível realizar o registro do Data Provider, executar a etapa final do processo de anonimização em dois níveis e enviar o conjunto de dados para publicação no Data Lake.

4.3.6 Data Lake

A implementação do Data Lake, dispõe de uma API REST para sua utilização, como ilustrado na Figura 21, sendo disponibilizados os seguintes endpoints:

POST **/template**	Utilizado para criar um novo *template*.
GET **/template/list**	Utilizado para verificar os *templates* disponíveis.
GET **/template/{id}**	Utilizado para visualizar todas as informações de um *template* por seu ID.
POST **/dataset/publish**	Utilizado para publicar o conjunto de dados.
GET **/dataset/*template*/{id}**	Utilizado para visualizar os dados publicados em um determinado *template* por seu ID.

De forma similar ao Local Broker e Honest Broker, os *templates* e os conjuntos de dados (pseudo)anonimizados são persistidos em uma base de dados na tecnologia MongoDB. Por fim, o componente em seu estado atual de implementação é capaz de exercer suas principais funções na arquitetura. Através dele é possível realizar a criação e visualização de *templates* e acessar aos conjuntos de dados publicados pelos Data Providers.

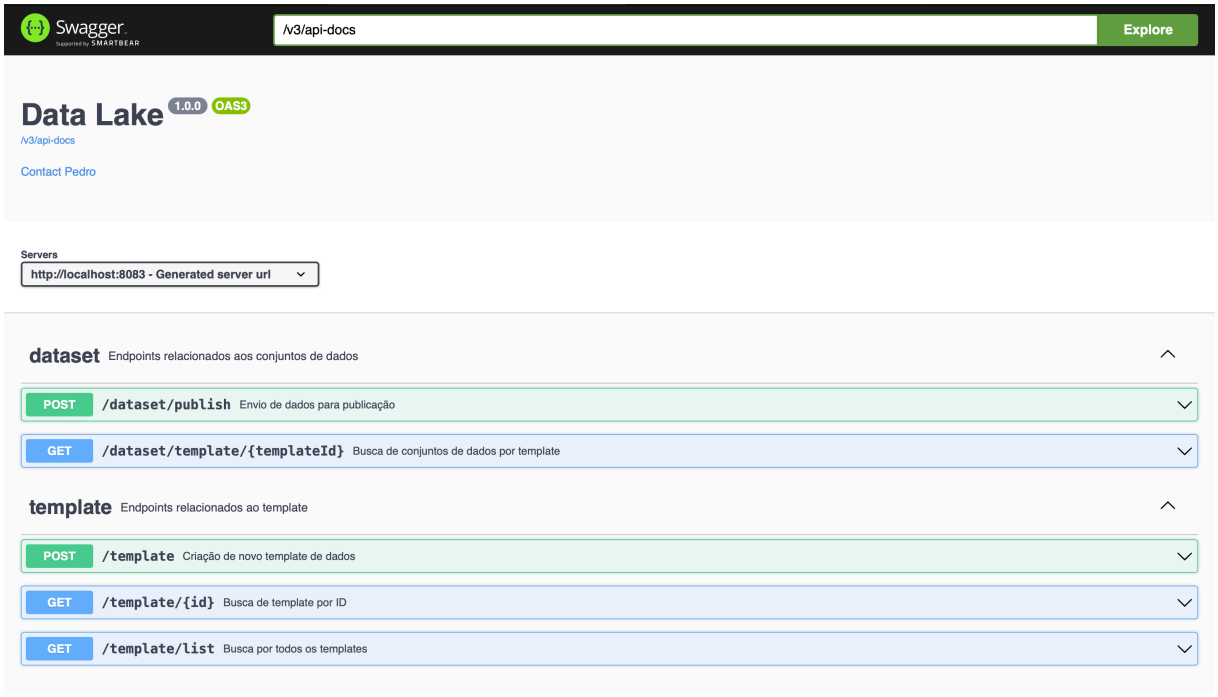


Figura 21: Tela principal do Swagger do Data Lake.

4.3.7 Fluxos

Esta seção descreve o estado atual da implementação dos principais fluxos da arquitetura proposta. No detalhamento de cada fluxo, são utilizadas as mesmas referências (A1..AN, B1..BN, etc) ilustradas na Figura 12.

A. Registro de Data Provider

O registro do Data Provider se inicia no Local Broker, onde o Data Provider realiza uma requisição POST ao endpoint **/provider/register** da API do Local Broker (A1). Nesta requisição, são enviados dados sobre o provedor, como: nome, email e número de telefone para contato. Ao receber esta requisição, o Local Broker inicia sua comunicação com o HB para dar seguimento ao processo de registro. Nesta comunicação com o Honest Broker, primeiro, o Local Broker solicita ao Honest Broker a sua chave pública, através de uma chamada GET ao endpoint **/provider/register** [A2.1].

Então, ao receber a solicitação, o Honest Broker codifica a sua chave pública no formato Base64 e a envia como resposta [A2.2]. Na geração do par de chaves do Honest Broker e demais operações criptográficas da implementação dos diferentes componentes, são utilizadas as APIs de criptografia do Bouncy Castle[9], como ilustrado na Listagem

[9]https://www.bouncycastle.org/index.html

4.1. Para fins de testes e validação, foi utilizada uma chave de 1024 bits, no entanto, para a utilização em produção este valor deve ser aumentado, por exemplo, para 2048 ou 4096. Além disto, o valor do expoente público utilizado é considerado fraco, sendo recomendado em produção utilizar um número primo grande. É importante ressaltar que a geração deste par de chaves ocorre uma única vez, na inicialização do Honest Broker, e são armazenadas, não podendo ser alteradas ou removidas, sob pena de inviabilizar a comunicação com Brokers Locais já registrados.

Listagem 4.1: Geração de par de chaves RSA com a utilização da biblioteca Bouncy Castle.

```
RSAKeyPairGenerator rsaGenerator = new RSAKeyPairGenerator();
rsaGenerator.init(new RSAKeyGenerationParameters(
        new BigInteger("1001", 16), // expoente público
        new SecureRandom(), // randômico
        1024, // força
        80)); // certeza

AsymmetricCipherKeyPair cipherKeyPair = rsaGenerator.generateKeyPair();
```

Do lado do Local Broker, ao receber a chave pública do HB, é realizada a decodificação Base64 e armazenada a chave em um arquivo de local parametrizável, evitando que seja preciso realizar uma nova solicitação a cada uso. Em seguida, o Local Broker codifica a sua chave pública em Base64, que será enviada junto com os dados de registro para o Honest Broker. Dando seguimento à preparação dos dados para o envio, o Local Broker gera uma chave AES de 256 bits, como ilustrado na Listagem 4.2. Esta chave simétrica é então utilizada para criptografar os dados de registro do provedor, como apresentado na Listagem 4.3. A chave simétrica AES é utilizada somente uma única vez, sendo gerada uma nova chave a cada nova comunicação.

Listagem 4.2: Geração de chave AES

```
KeyGenerator keyGenerator = KeyGenerator.getInstance("AES");
keyGenerator.init(256);
byte[] key = keyGenerator.generateKey().getEncoded();
```

A chave AES é então criptografada com a chave pública do Honest Broker, como ilustrado na Listagem 4.4. Então, o Local Broker envia uma requisição POST ao endpoint **/provider/register** do Honest Broker contendo a chave simétrica e os dados do provedor, ambos criptografados [A2.3].

Listagem 4.3: Criptografia de dados com chave AES

```
1  PaddedBufferedBlockCipher aes = new PaddedBufferedBlockCipher(new
       ↪ CBCBlockCipher(new AESEngine()));
2  aes.init(true, new KeyParameter(key));
3
4  byte[] cipherText = new byte[aes.getOutputSize(data.length())];
5  int outputLength = aes.processBytes(data.getBytes(), 0, data.length(),
       ↪ cipherText, 0);
6
7  try {
8      aes.doFinal(cipherText, outputLength);
9  } catch (InvalidCipherTextException e) {
10     throw new StandardException();
11 }
```

Ao receber a requisição, o Honest Broker descriptografa a chave AES e a utiliza para descriptografar os dados do provedor. Em seguida, é verificado se há algum provedor registrado com o mesmo nome. Se existir, é retornada então uma mensagem de erro [A2.4]. Caso contrário, as informações recebidas são armazenadas em uma *collection* do banco de dados e retornada uma mensagem de sucesso ao Local Broker [A2.4]. Sabe-se que o uso do nome para garantir a unicidade de provedores no sistema não é a solução ideal, mas sua implementação foi realizada para demonstrar a necessidade desta unicidade. Por fim, ao receber uma resposta de sucesso por parte do Honest Broker, o Local Broker está apto a utilizar os outros fluxos previstos.

Listagem 4.4: Ecriptação de dados com chave pública

```
1  Security.addProvider(new BouncyCastleProvider());
2
3  RSAEngine engine = new RSAEngine();
4  engine.init(true, getPublicKey());
5
6  byte[] hexEncodedCipher = engine.processBlock(data.getBytes(), 0,
       ↪ data.length());
```

B. Preparação dos Dados no Provedor

Para que o Data Provider possa preparar o conjunto de dados para envio é necessário a consulta do *template* desejado. Para visualizar os *templates* disponíveis são realizadas chamadas a API disponibilizada pelo Data Lake (B1). Em seguida, o provedor pode então enviar o conjunto de dados desejado para o Local Broker através do endpoint **/dataset/-send** (B2). Ao realizar a chamada o processo de (pseudo)anonimização é necessário obter o *template* completo, de acordo com o DID recebido do provedor, através da chamada ao endpoint **/template/{id}**, substituindo o parâmetro *id* pelo valor do DID (B3). Ao obter o *template* é verificado se todos os atributos definidos no *template* foram enviados, se todos os registros do conjunto de dados contém valor para todos os atributos, e se os valores recebidos de um atributo estão entre os valores permitidos (se aplicável).

C. (Pseudo)Anonimização dos Dados

Assim que o Local Broker recebe o conjunto de dados, se inicia o processo do primeiro nível de anonimização dos dados. Neste fluxo serão executadas duas ações, a anonimização dos dados e a geração do pseudônimo local. A anonimização dos dados no estado atual da implementação suporta dois modelos de privacidade, o k-anonimato e l-diversidade. A configuração para anonimização dos dados - os valores de k, l e o limite de supressão - são definidos no *template*, obtido no passo (B3) descrito anteriormente. O Local Broker faz chamadas à biblioteca do ARX A opção pelo uso da biblioteca do ARX deve-se ao fato de ser significativamente utilizada e citada na literatura, sendo amplamente validada. Ao receber os dados a serem anonimizados e o *template* (C1), estes são preparados e submetidos ao processo de (pseudo)anonimização. Caso os dados recebidos não sejam passíveis de anonimização, é retornado ao provedor uma mensagem a indicar que não é possível realizar a anonimização dos dados enviados de acordo com os parâmetros estabelecidos pelo *template* indicado. Ocorrendo este ou outro erro, o processamento dos dados é encerrado e não são armazenadas informações. Em caso de sucesso, o conjunto de dados segue para a geração dos pseudônimos.

O pseudônimo no Local Broker é gerado através da aplicação de uma função *hash* SHA256 ao atributo identificador junto a um *salt*. A utilização do *salt* na criação do pseudônimo local é extremamente significativa, dado que a arquitetura foi projetada para suportar mais de um Data Provider, e, portanto, a geração de pseudônimos diferentes para o mesmo indivíduo em diferentes provedores, garante que um provedor não seja capaz de identificar um indivíduo a partir dos dados de outro provedor, como discutido na Seção

4.1.2. A geração do pseudônimo local é realizada para cada um dos registros do conjunto de dados recebido, sendo armazenada em uma *collection* do banco de dados a ligação entre pseudônimo e atributo identificador, para quando for necessário a reidentificação de um indivíduo. Cabe aqui ressaltar que o processo de anonimização dupla impõe a necessidade de reidentificação também em dois passos. Assim, um ataque de reidentificação só terá sucesso se conseguir comprometer e reverter os pseudônimos em duas instituições (correspondentes ao Honest Broker e o Data Provider).

Seguindo o processamento dos dados, é aplicada a função *hash* SHA256 ao atributo identificador de cada registro, porém sem a utilização de um *salt*. Esse valor será enviado ao Honest Broker, que o utilizará para realizar a ligação dos dados anonimizados ao pseudônimo global.

Cabe aqui ressalvar que os passos (C2) e (C3), referentes à análise de risco de reidentificação de acordo com o R_{MAX} indicado no *template*, não foi implementada até a escrita deste documento. Por fim, o conjunto de dados anonimizado, o pseudônimo local e *hash* do atributo identificador dos registros são armazenados em um arquivo JSON no subdiretório *forapproval* do *Staging Folder* (C4). O conjunto de dados fica pendente de uma análise por parte do Data Provider. O provedor pode obter a lista de dados pendentes de envio através dos endpoints **/dataset/pending/list** e **/dataset/pending/{id}** do Local Broker (C5). O conjunto de dados somente será enviado ao Honest Broker mediante aprovação do provedor. O Data Provider pode aprovar ou cancelar a publicação destes dados através de uma chamada ao endpoint **/dataset/pending/update** no Local Broker (C6). Os conjuntos de dados aprovados são movidos do subdiretório *forapproval* para o subdiretório *tosend*, enquanto que os cancelados são movidos para o subdiretório *cancelled* do *Staging Folder*. O fluxo é finalizado com o envio dos conjuntos de dados aprovados de forma assíncrona através de uma tarefa agendada que é executada de tempo em tempo (C7).

D. Ligação e Publicação dos Dados Anonimizados

O fluxo de ligação dos dados anonimizados tem início quando o Honest Broker recebe os dados de um Local Broker. Primeiro é realizada a descriptografia dos dados, onde o HB utiliza de sua chave privada para descriptografar a chave AES enviada e então com esta chave é realizada a descriptografia dos dados. Então, o HB percorre todos os registros do conjunto de dados e faz a ligação dos dados (D1). Para ligar o pseudônimo local (LID) a um pseudônimo global (GID), o Honest Broker verifica a existência de um

pseudônimo global já criado para o *hash* do atributo identificador em seu banco de dados. Caso exista, o Honest Broker utiliza este pseudônimo global no mapeamento (D2). Caso contrário, o Honest Broker gera um novo pseudônimo global para representar globalmente o indivíduo e o adiciona no banco de dados, armazenando o *hash* do atributo identificador, o pseudônimo gerado, e uma lista contendo o pseudônimo local recebido e seu provedor de origem (D2).

A geração do pseudônimo global acontece da mesma forma que o local no Local Broker, aplicando uma função *hash* SHA256 ao *hash* do atributo identificador recebido junto a um *salt*. O *salt* é criado de forma randômica e armazenado na inicialização do Honest Broker e seu uso se faz necessário para que não seja possível inferir o pseudônimo global de um indivíduo se descoberto o *hash* de seu atributo identificador, visto que o *salt* é secreto. Com este processo do pseudônimo global realizado, fica garantido que, os dados anonimizados que tenham como origem um mesmo indivíduo, ainda que representado em diferentes provedores, seja mapeado para um único pseudônimo no Data Lake.

Em seguida, o Honest Broker remove o *hash* do atributo identificador e substitui o pseudônimo local pelo pseudônimo global correspondente, em cada registro do conjunto de dados, resultando no conjunto de dados que será publicado. Por fim, este novo conjunto de dados é armazenado em um arquivo JSON no subdiretório *tosend* do *Staging Folder* (D3).

De forma similar ao procedimento usado no Local Broker, o Honest Broker executa o envio dos conjuntos de dados ao Data Lake de forma assíncrona através de uma tarefa agendada que é executada periodicamente (D4). Esta tarefa verifica se há algum conjunto de dados no subdiretório *tosend*, enviando um por vez através de uma chamada ao endpoint **/dataset/publish** da API do Data Lake (D5). Alcançando o sucesso no envio dos dados, o Honest Broker move o arquivo do conjunto de dados enviado do subdiretório *tosend* para o subdiretório *send*. Caso ocorra alguma falha neste processo, é realizada uma nova tentativa de envio na execução subsequente da tarefa. Para encerrar o fluxo o conjunto de dados é armazenado no banco de dados no Data Lake (D6).

E-F. Gerenciamento de Usuários e Acesso ao Data Lake

Devido a restrições de tempo, os fluxos de Gerenciamento de Usuários (E) e Acesso ao Data Lake (F) não foram devidamente implementados como funcionalidades da plataforma. Assim, nos testes e validação, foram, quando possível, emulados via *script*.

5 Validação e Testes da Solução Proposta

Com o objetivo de validar os pressupostos de preservação da privacidade da arquitetura aqui proposta, foram desenhados e realizados quatro diferentes testes. O primeiro teste visa garantir que o processo de anonimização dos dados, usando a biblioteca do ARX, produz o mesmo resultado do software ARX. O segundo teste visa assegurar que os pseudônimos locais e global estão sendo gerados de forma apropriada e seguindo os requisitos de segurança da arquitetura. O terceiro teste tem como objetivo avaliar todo o processo de (pseudo)anonimização dos dados fim a fim, ou seja, desde a (pseudo)anonimização e envio dos dados dos diferentes Data Providers, até a sua recepção no Data Lake. Por fim, o quarto e último teste explora um cenário de reidentificação de um indivíduo a partir dos dados consolidados no Data Lake. Foi utilizado um conjunto de dados distinto para cada um dos cenários, visando evitar o enviesamento dos testes da arquitetura, provendo assim uma maior diversidade de informações. A execução de todos os testes foi realizada em uma única máquina com o sistema operacional macOS 12.6 com 8GB de memória e processador Apple M1.

5.1 Teste 1: Mecanismo de anonimização

Uma das principais funções da arquitetura é realizar a anonimização dos dados, processo este executado no Local Broker. Assim, com o objetivo de verificar o correto funcionamento deste processamento foi concebido este teste que consiste em certificar que os dados anonimizados pelo Local Broker, realizado através da biblioteca do ARX, produza o mesmo resultado da ferramenta do ARX em sua versão gráfica.

Para a execução do teste foi utilizada a ferramenta ARX na versão 3.9.0, ilustrada na Figura 22, e o Local Broker (LB), juntamente com os demais componentes da arquitetura, foi executado através de *containers* Docker. O LB foi implementado usando a biblioteca

do ARX também na versão 3.9.0. Ambos, o ARX e o *container* Docker, foram executados na mesma máquina.

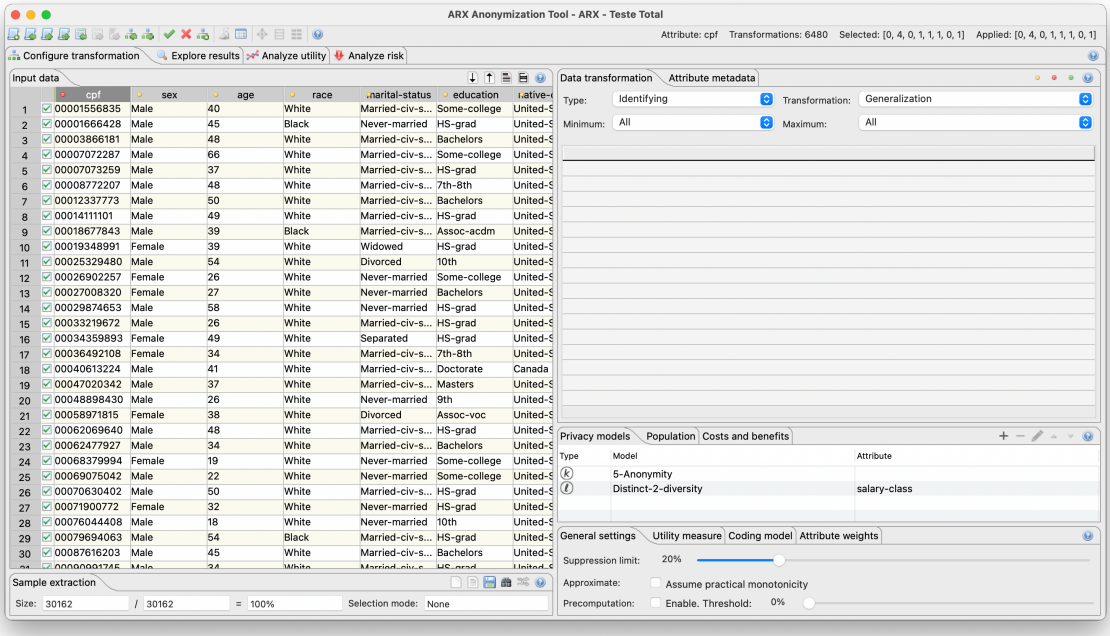


Figura 22: Tela inicial do ARX com o conjunto de dados e configurações carregadas.

Foi utilizado o conjunto de dados de exemplo ADULT (ver Anexo A), fornecido na distribuição da ferramenta ARX. Além dos dados, o exemplo traz ainda um arquivo com a definição das hierarquias de generalização necessárias, ambos em arquivos de formato *Comma-separated values* (CSV). O conjunto de dados contém 30.162 registros, onde cada registro é composto por nove atributos. Como não há, entre os atributos, nenhum identificador direto, a fim de exercitar todo o mecanismo de (pseudo)anonimização, foi adicionada a coluna CPF para servir como identificador direto dos indivíduos. Os valores para a coluna CPF foram gerados com o uso do pacote *cpf-generator* do Python, em que para cada registro era gerado um novo CPF. Com a adição desta nova coluna, o conjunto de dados passou a ser descrito por dez atributos, sendo um identificador, oito quase-identificadores e um sensível. Os atributos são listados na Tabela 2.

A ferramenta ARX e a arquitetura proposta, foram igualmente configuradas para utilizar os modelos de privacidade k-anonimato e l-diversidade combinados, sendo parametrizados com $k = 5$ e $l = 2$, respectivamente. Além disto, o limite máximo para a supressão de dados foi definido em 20%. No ARX, estas configurações de anonimização e a definição das hierarquias dos atributos quase-identificadores e sensíveis foram realizadas em sua interface gráfica, o conjunto de dados foi então submetido e o conjunto de dados

anonimizado foi obtido em formato CSV.

Tabela 2: Atributos do conjunto de dados utilizado no teste.

Atributo	Descrição
cpf	Identificador que representa o Cadastro de Pessoa Física (CPF) do indivíduo, sendo esta a coluna adicionada ao conjunto de dados original com valores gerados randomicamente.
sex	Quase-identificador que representa o sexo do indivíduo.
age	Quase-identificador que representa a idade do indivíduo.
race	Quase-identificador que representa a raça do indivíduo.
marital-status	Quase-identificador que representa o estado civil do indivíduo.
education	Quase-identificador que representa o nível de escolaridade do indivíduo.
native-country	Quase-identificador que representa o país de origem do indivíduo.
workclass	Quase-identificador que representa a classe de trabalho do indivíduo.
occupation	Quase-identificador que representa a ocupação/área profissional do indivíduo.
salary-class	Atributo sensível que representa a classe salarial do indivíduo.

Na arquitetura implementada, foi criado um *template* (ver Anexo D) com as configurações de anonimização e hierarquias dos atributos idênticas às usadas na ferramenta ARX. A seguir, o conjunto de dados original foi enviado ao Local Broker e obtido o conjunto de dados (pseudo)anonimizado antes que o mesmo fosse enviado ao Honest Broker. Importante recordar que na arquitetura os dados são enviados e retornados no formato *JavaScript Object Notation* (JSON) em uma estrutura própria, sendo necessária a transformação do conjunto de dados para o seu envio. Além disto, para possibilitar uma comparação direta entre os resultados no ARX e na arquitetura, a coluna com o pseudônimo local foi removida do conjunto de dados resultante, já que este atributo não é gerado na ferramenta ARX.

Para a comparação dos conjuntos de dados anonimizados, foi escrito um script em Python. De forma simplificada, os dois conjuntos de dados anonimizados resultantes foram carregados em estruturas do tipo *DataFrame* e aplicadas comparações disponíveis nos pacotes NumPy e Pandas - estes pacotes são amplamente utilizados e validados para a

análise de dados em Python. A Listagem 5.1 apresenta as linhas de código principais desta comparação, onde arxDf representa o *DataFrame* resultante da ferramenta ARX e archDf o resultante da arquitetura implementada. O resultado foi que os dois *DataFrames* são idênticos.

Listagem 5.1: Comparação dos conjuntos de dados resultantes da ferramenta e arquitetura implementada.

```
1 pandas.testing.assert_frame_equal(arxDf.reset_index(drop=True),
    ↪ archDf.reset_index(drop=True))
2 numpy.array_equal(arxDf.values, archDf.values)
```

Com base no resultado obtido, é possível concluir que o processo de anonimização dos dados, implementado pela arquitetura aqui proposta, produz um resultado idêntico ao produzido pela ferramenta ARX. Cabe relembrar que o ARX é tido como uma das principais ferramentas de anonimização de dados, tendo sido amplamente validada e utilizada em diversos trabalhos acadêmicos e projetos de pesquisa na área da privacidade dos dados.

5.2 Teste 2: Geração dos pseudônimos local e global com múltiplos Data Providers

Entre as decisões de implementação, uma das mais importantes é a aplicação de dois níveis de (pseudo)anonimização, processo que acontece no Local Broker e no Honest Broker, onde são gerados os pseudônimos local (LID) e global (GID), respectivamente. Com o objetivo de verificar o correto funcionamento da geração destes pseudônimos, este teste visa verificar se para mesmo indivíduo, representado em múltiplos provedores, é gerado um LID diferente em cada um dos provedores (ou Local Broker) e se, ao chegar no Honest Broker estes diferentes LIDs são corretamente mapeados para um mesmo GID.

A execução do teste contou com a instanciação de três *containers* Docker com a implementação do componente Local Broker, representando os múltiplos provedores em execução, nomeados de Provedor A, B e C. Foram ainda instanciados dois outros *containers*, um contendo as implementações para o Honest Broker e outro para o Data Lake. Todos os *containers* foram executados em uma única máquina. O conjunto de dados utilizado no teste foi o *ADMISSIONS* (ver Anexo B) na versão 1.0 de demonstração do MIMIC-IV, fornecido pelo *Medical Information Mart for Intensive Care (MIMIC)*. O MIMIC é um banco de dados que inclui dados desidentificados relacionados a saúde de pacientes das

unidades de cuidados críticos do *Beth Israel Deaconess Medical Center*, para os anos de 2008 a 2019. O conjunto de dados *ADMISSIONS* contém 275 registros, onde cada registro é descrito por 15 atributos.

Tabela 3: Atributos do conjunto de dados utilizado no teste 2.

Atributo	Descrição
index	Atributo insensível que representa a posição do registro no conjunto de dados originário.
subject_id	Identificador que representa o paciente.
hadm_id	Identificador que representa a admissão do paciente no hospital.
admittime	Quase-identificador que representa a data e hora da admissão do paciente no hospital.
dischtime	Quase-identificador que representa a data e hora que o do paciente recebeu alta do hospital.
deathtime	Identificador que representa a data e hora da morte do paciente no hospital, se aplicável.
admission_type	Atributo insensível que representa a classificação de urgência na admissão.
admission_location	Atributo insensível que representa o local do paciente antes da admissão.
discharge_location	Atributo insensível que representa a disposição do paciente após receber alta do hospital.
insurance	Atributo insensível que representa o seguro do paciente.
language	Identificador que representa o idioma do paciente.
marital_status	Quase-identificador que representa o estado civil do paciente.
ethnicity	Identificador que representa a etnia do paciente.
edregtime	Quase-identificador que representa a data e hora que o paciente foi registrado no departamento de emergência.
edouttime	Quase-identificador que representa a data e hora que o paciente recebeu alta do departamento de emergência.
hospital_expire_flag	Identificador que sinaliza a morte do paciente durante a internação.

Como o teste objetiva simular três Data Providers distintos, foi implementado um

script em linguagem de programação Python, para dividir o conjunto de dados em três partes. De forma geral, o script percorre sequencialmente o conjunto de dados original, adicionando cada registro lido em um dos três novos conjuntos de dados, num esquema de *round robin*. Assim, o primeiro registro é atribuído ao conjunto de dados A, o segundo ao B e o terceiro ao C, retornando de forma circular ao A e assim por diante, até que todo o conjunto de dados original seja completamente distribuído. Na criação destes novos conjuntos de dados foi introduzido artificialmente a cada registro um atributo denominado *index* que representa a posição do registro no conjunto de dados original. Este atributo foi adicionado exclusivamente para o teste, de forma a facilitar a verificação dos pseudônimos (LID e GID) gerados. Ao final do script foi verificada a distribuição dos indivíduos nos conjuntos de dados, a fim de garantir que um mesmo indivíduo foi distribuído nos diferentes conjuntos de dados. Assim, os conjuntos de dados utilizados no teste passaram a ser descritos por 16 atributos, sendo seis identificadores, cinco quase-identificadores e cinco insensíveis. Os atributos são listados na Tabela 3. Importante ressaltar que a classificação dos atributos foi realizada de forma arbitrária, visto que a anonimização não é o foco do teste e os dados já se encontram desidentificados.

No Data Lake, foi criado um *template* (ver Anexo E) com as configurações de (pseudo)anonimização e definição dos atributos e suas hierarquias. Os modelos de privacidade k-anonimato e l-diversidade foram parametrizados com os valores $k = 2$ e $l = 2$, respectivamente. Além disto, o limite máximo de supressão foi definido em 90%, um valor não recomendado em produção, porém como neste teste o importante é a geração dos LIDs e GIDs, um nível de supressão muito baixo poderia suprimir uma grande quantidade de registros, o que dificultaria a análise dos resultados. Já na execução do teste, foi realizado o envio dos conjuntos de dados A, B e C, no provedores A, B e C respectivamente. Posteriormente foi realizado todo o processo de aprovação dos dados e envio ao Honest Broker até que fosse disponibilizado no Data Lake.

A análise dos dados, baseou-se na seleção de um conjunto de indivíduos presente ao menos em dois dos três conjuntos de dados e no acompanhamento destes registros (usando a coluna *index*) ao longo de todo o processo. Por exemplo, foi selecionado o individuo com o *subject_id* 10004235 e com valores para coluna *index*: A [0, 156]; B [259]. Como primeira verificação, foi buscado no conjunto de dados (pseudo)anonimizado em cada um dos Brokers Locais, o LID associado ao individuo em cada um dos provedores. Como resultado da busca nos provedores A e B, foram encontrados os registros apresentados nas Listagens 5.2 e 5.3 nesta ordem, onde o primeiro atributo do registro original corresponde ao *index* e o segundo corresponde ao *subject_id*, enquanto que, no registro (pseudo)anonimizado,

o primeiro atributo corresponde ao LID e o segundo ao *index*.

Listagem 5.2: Individuo no Data Provider A.

```
Registro original:
0,10004235,24181354,2196-02-24 14:38:00,2196-03-04 14:02:00,,URGENT,TRANSFER
    ↪ FROM HOSPITAL,SKILLED NURSING
    ↪ FACILITY,Medicaid,ENGLISH,SINGLE,BLACK/AFRICAN AMERICAN,2196-02-24
    ↪ 12:15:00,2196-02-24 17:07:00,0
156,10004235,22187210,2196-06-20 21:11:00,2196-06-22 13:30:00,,DIRECT EMER.,
    ↪ PHYSICIAN REFERRAL,HOME HEALTH
    ↪ CARE,Medicaid,ENGLISH,SINGLE,BLACK/AFRICAN AMERICAN,,,0

Registro (pseudo)anonimizado no provedor:
6c2631083de00554a3ff3506052995580a03b09a3568b477237cdf2f4db9dc88,0,*,*,2196,
    ↪ 2196,*,URGENT,TRANSFER FROM HOSPITAL,SKILLED NURSING
    ↪ FACILITY,Medicaid,*,*,*,2196,2196,*
6c2631083de00554a3ff3506052995580a03b09a3568b477237cdf2f4db9dc88,156,*,*,*,*,*,
    ↪ DIRECT EMER.,PHYSICIAN REFERRAL,HOME HEALTH CARE,Medicaid,*,*,*,*,*,*
```

Através das Listagens 5.2 e 5.3 é possível notar que, para o mesmo indivíduo, foram gerados diferentes LIDs nos provedores A e B. Além disso, verifica-se que os múltiplos registros do mesmo indivíduo em um provedor dão origem sempre ao mesmo LID.

Listagem 5.3: Individuo no Data Provider B.

```
Registro original:
259,10004235,25970245,2196-06-14 08:30:00,2196-06-19 14:54:00,,SURGICAL SAME
    ↪ DAY ADMISSION,PHYSICIAN REFERRAL,HOME HEALTH
    ↪ CARE,Medicaid,ENGLISH,SINGLE,BLACK/AFRICAN AMERICAN,,,0

Registro (pseudo)anonimizado no provedor:
f765c338697058731bd7389dd7ebcfad7bea113d3788290ca1b591e66fb4036b,259,*,*,*,*,*,
    ↪ SURGICAL SAME DAY ADMISSION,PHYSICIAN REFERRAL,HOME HEALTH
    ↪ CARE,Medicaid,*,*,*,*,*,*
```

A seguir, foi verificado o mapeamento dos LIDs no Honest Broker para o mesmo GID. Novamente, foram utilizados os valores do campo *index* relacionados ao *subject_id* 10004235. Os registros obtidos do Data Lake, que já incluem o GID correspondente, são demonstrados na Listagem 5.4.

Listagem 5.4: Individuo no Data Lake.

```
1 Registros (pseudo)anonimizados no Data Lake:
2 e7a32aff94dfffb9c4c2041e5506ce19e2a248c837fa0990d964fe278b297f5b,0,*,*,2196,
    ↪ 2196,*,URGENT,TRANSFER FROM HOSPITAL,SKILLED NURSING
    ↪ FACILITY,Medicaid,*,*,*,2196,2196,*
3 e7a32aff94dfffb9c4c2041e5506ce19e2a248c837fa0990d964fe278b297f5b,156,*,*,*,*,*,
    ↪ DIRECT EMER.,PHYSICIAN REFERRAL,HOME HEALTH CARE,Medicaid,*,*,*,*,*,*
4 e7a32aff94dfffb9c4c2041e5506ce19e2a248c837fa0990d964fe278b297f5b,259,*,*,*,*,*,
    ↪ SURGICAL SAME DAY ADMISSION,PHYSICIAN REFERRAL,HOME HEALTH
    ↪ CARE,Medicaid,*,*,*,*,*,*
```

Na Listagem 5.4 é possível verificar que foi criado o mesmo GID (e7a32aff94dfffb9c4c2041e5506ce19e2a248c837fa0990d964fe278b297f5b) para os valores de *index* 0, 156 e 259: ou seja, para os registros correspondentes ao *subject_id* 10004235. Com isso, conclui-se que os registros de um mesmo indivíduo provenientes de diferentes provedores chegaram ao Data Lake com o mesmo pseudônimo global, permitindo a ligação entre dados (pseudo)anonimizados.

Por fim, com base nos resultados obtidos, é possível concluir que o processo de geração dos pseudônimos global e local a partir de múltiplos provedores, implementado pela arquitetura proposta, produz o resultado correto e esperado.

5.3 Teste 3: Geração de um conjunto de dados (pseudo)anonimizado fim a fim

Após validar o processo de anonimização e a geração dos pseudônimos locais (LID) e geral (GID), nas Seções 5.1 e 5.2 respectivamente, o próximo passo foi testar o fluxo completo para a alimentação de um conjunto de dados, conforme definido na arquitetura. Além de dados sobre o correto funcionamento da (pseudo)anonimização de um conjunto de dados fim a fim, este teste também permitiu obter dados iniciais sobre o desempenho da arquitetura e algumas de suas limitações.

De forma similar ao Teste 2, a execução do Teste 3 contou com a instanciação de três *containers* Docker com a implementação do componente Local Broker, representando múltiplos provedores em execução, nomeados de Provedor A, B e C. Foram ainda instanciados dois outros *containers*, um contendo a implementação do Honest Broker e outro do Data Lake. Todos os *containers* foram executados em uma única máquina. Porém, dife-

rentemente do Teste 2, neste Teste 3, por razões de desempenho e limitações da máquina, os provedores foram instanciados de forma alternada, mas sem prejuízo ao teste, uma vez que, numa situação real, é pouco provável o envio simultâneo por vários provedores.

O conjunto de dados utilizado no Teste 3 foi extraído da tabela *OBSERVATIONS* (ver Anexo C) do banco de dados COVID-19 na versão envolvendo 10.000 pacientes, gerado pela Synthea™ (Walonoski et al. 2017) e hospedados no SyntheticMass. A Synthea é uma Simulação de População de Pacientes Sintéticos, do inglês *Synthetic Patient Population Simulation*, usada para gerar dados sintéticos e realistas de pacientes e registros de saúde dentro do SyntheticMass. Embora envolva 10.000 pacientes, há diferentes observações relacionadas com um mesmo paciente e, portanto, a quantidade total de registros é de 1.659.750. Cada registro é descrito por 8 atributos, sendo dois atributos identificadores, dois quase-identificadores e quatro insensíveis. Detalhes sobre os atributos são apresentados na Tabela 4.

Tabela 4: Atributos do conjunto de dados utilizado no teste 3.

Atributo	Descrição
date	Quase-identificador que representa a data do encontro ou consulta.
patient	Identificador que representa o pseudônimo do paciente.
encounter	Atributo identificador que representa o pseudônimo identificador do encontro.
code	Quase-identificador que representa o código LOINC[1] referente ao encontro.
description	Atributo insensível que representa a descrição da observação realizada.
value	Atributo insensível que representa o valor observado ou medido.
unit	Atributo insensível que representa a unidade de medida aplicada.
type	Atributo insensível que representa o tipo do valor observado (*numeric*, *text*, etc).

Como o teste objetiva simular três Data Providers distintos, assim como no Teste 2, foi implementado um script em linguagem de programação Python, para dividir o conjunto de dados em três partes. Porém, diferente do script usado no Teste 2, o script usado aqui

[01] O *Logical Observation Identifiers, Names, and Codes (LOINC)* é um padrão internacional para a codificação de medições, observações e documentos de saúde. Com isso, a LOINC utilização do LOINC permite garantir a consistência semântica na integração entre diferentes sistemas <https://loinc.org>.

percorre sequencialmente o conjunto de dados original, agrupa cada conjunto de registros referente a um encontro e só então insere todo o conjunto de registros em um dos três novos conjuntos de dados, também num esquema de *round robin*. Assim, o primeiro conjunto de registros é atribuído ao conjunto de dados A, o segundo ao B e o terceiro ao C, retornando de forma circular ao A e assim por diante, até que todo o conjunto de dados original seja completamente distribuído.

No Data Lake, foi criado um *template* (ver Anexo F) com as configurações de (pseudo)anonimização e definição dos atributos e suas hierarquias. Os modelos de privacidade k-anonimato e l-diversidade foram parametrizados com os valores $k = 5$ e $l = 2$, respectivamente. Além disto, o limite máximo de supressão de dados foi definido em 30%. Cabe ressaltar que neste teste foi utilizado um valor de supressão mais próximo da realidade, embora ainda alto.

A fim de automatizar a execução do teste, foi desenvolvido um script na linguagem de programação Python para realizar as operações de envio, busca, visualização e aprovação de um conjunto de dados no Local Broker. O script é parametrizável, sendo possível definir qual o Provedor de destino das requisições e a quantidade de registros a serem enviados. Para analisar o desempenho da arquitetura implementada foram adicionados cálculos do tempo de execução em milissegundos para as operações de anonimização (através da biblioteca do ARX), geração de pseudônimo local (primeiro nível), geração de *hash* para o atributo identificador, armazenamento do conjunto de dados (pseudo)anonimizado, além de outras operações. Importante ressaltar que a adição do cálculo do tempo de execução das operações foi realizado de forma exclusiva para o teste e não fazem parte da arquitetura.

O script foi inicialmente desenhado para executar o teste de forma incremental, realizando envios dos primeiros 1.000, 10.000, 20.000, seguindo com incrementos de 10.000 até 100.000, então passando para 150.000, 200.000, seguindo com incrementos de 50.000 até aproximadamente 600.000 registros para cada um dos três provedores, ou seja, até o envio de todos os 1.659.750 registros. Após a execução e obtenção dos resultados, foi possível gerar o gráfico demonstrado na Figura 23 com a comparação dos tempos de execução de cada operação, de acordo com a sua carga.

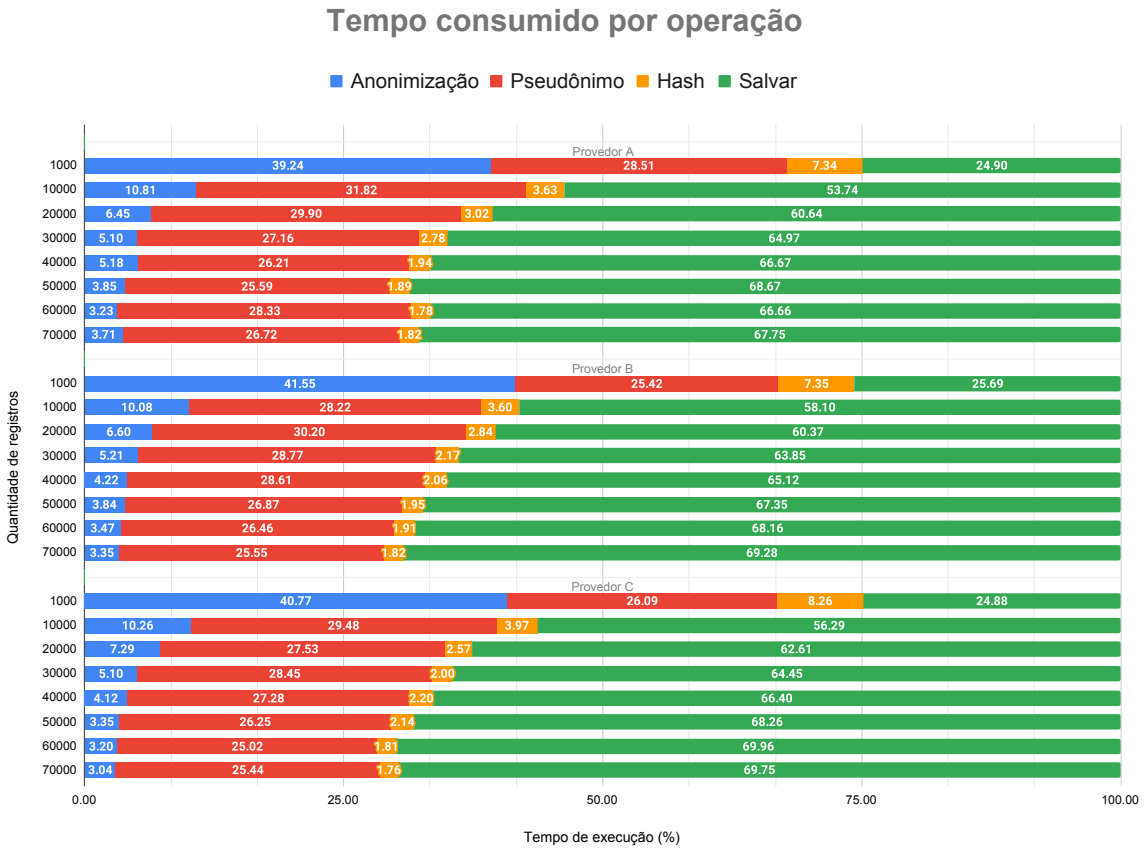


Figura 23: Tempo de execução das operações analisadas no teste.

Depois da análise dos resultados ilustrados na Figura 23, foi possível observar que:

1. A operação que consome menos tempo é a geração do *hash* do atributo identificador.

2. Os tempos de anonimização dos dados e geração do *hash* do atributo identificador não crescem tanto como a geração do pseudônimo local e armazenamento do conjunto de dados (pseudo)anonimizado com o aumento do número de registros.

3. A operação que consome mais tempo é o armazenamento dos dados, consumindo em média 60% do tempo total, sendo uma exceção apenas para a execução com 1.000 registros. Isto se dá devido ao fato da ação de escrita de dados em disco ser custosa e, assim, quanto mais registros (pseudo)anonimizados, maior o tamanho do arquivo a ser armazenado em disco. Por exemplo, no envio de 70.000 registros do Provedor C, foi armazenado em disco um arquivo com aproximadamente 40MB.

4. A operação de geração do pseudônimo local (LID), apesar de aplicar uma função de

hash com *salt*, ela não apresenta um tempo similar a geração de *hash* do atributo identificador. Isto ocorre devido ao fato de que a geração do pseudônimo aplica a função de *hash* somente quando ainda não há pseudônimo gerado para o atributo identificador no banco de dados. Caso exista, este valor é usado sem que seja necessário a aplicação da função de *hash*. Com isso, os valores acrescidos para a geração do pseudônimo, devem-se essencialmente às operações no banco de dados.

5. A Figura 23 apresenta tempos de execução somente para cargas até 70.000 registros, pois já nos testes com 80.000 registros, foi constatada uma limitação do componente MongoDB, utilizado como Sistema de Gerenciamento de Banco de Dados (SGDB) no Data Lake. Essa limitação está relacionada com o armazenamento de documentos com tamanho superior a 16Mb, segundo a documentação do MongoDB. Embora algumas soluções tenham sido pesquisadas, no geral, demandavam de alterações substanciais aos testes e até arquiteturais, tendo sido decidido por finalizar o Teste 3, relatar esta limitação e discutir estas possíveis soluções como trabalhos futuros.

De forma geral, este teste serviu para exercitar os diferentes componentes da arquitetura proposta e revelar limitações e possíveis melhorias em sua implementação. Como exemplo de melhoria, na primeira execução do teste, o tempo de execução da geração do *hash* do atributo identificador no Provedor A para 70.000 registros foi de 209.752ms, e após a análise e alteração, este tempo caiu para 891ms, uma redução de 208.861ms (99,58%). Por fim, com base nos resultados obtidos, é possível concluir que, apesar da limitação relatada a geração de um conjunto de dados anonimizado fim a fim, implementado pela arquitetura proposta, produz um resultado correto.

5.4 Teste 4: Reidentificação de um indivíduo

Com o uso dos pseudônimos local (LID) e global (GID) a arquitetura proposta possibilita a reidentificação de um indivíduo através de seu pseudônimo exposto no Data Lake, o GID. Cabe ressaltar que este procedimento é muito particular, sendo ativado apenas em casos que envolvam a segurança do paciente. Assim, com o objetivo de verificar a capacidade de reidentificação, este teste consiste em reidentificar um indivíduo a partir dos dados (pseudo)anonimizados no Data Lake. Isso implica em reverter o mapeamento do GID no Honest Broker e, em seguida, reverter o mapeamento do LID para o *subject_id* correspondente em cada Data Provider.

Este teste de reidentificação considera os resultados obtidos no Teste 2, incluindo a

mesma estrutura e componentes instanciados. Adicionalmente, uma vez que a arquitetura originalmente implementada não previa esta operação de reidentificação, foi implementado um endpoint na API do Honest Broker e outro na API do Local Broker, exclusivos para este teste de reidentificação, não sendo portanto, implementado mecanismos adicionais de controle de acesso.

Listagem 5.5: Registro do banco de dados do Honest Broker.

```
{
  "_id": "fd46029c3f49ed5738b64315130e1fd8ac5cc3bc1cb4a7714985f9775299073b",
  "pseudonymGlobal":
     ↪ "e9e1ae816ee85daccceae45c86b8cb566ff61b903382ece17820970c67cf9dea",
  "pseudonymsLocal": [
    {"pseudonym":
       ↪ "3e74eb3cf804f5022340f582074f85c658f02d87153d8cb47f99ac935a86d289",
      "sourceProvider": "ProvedorC" },
    {"pseudonym":
       ↪ "fa755a18ed68de43b30b8c5fa994f65cfa277c5dcc3bcaf6f6c1c618d2387d67",
      "sourceProvider": "ProvedorA" },
    {"pseudonym":
       ↪ "4c2aa334e0425f0e9f54938c48cca349099701e5af1d420993981ff58474af9b",
      "sourceProvider": "ProvedorB" }
  ]
}
```

Do lado do Honest Broker, foi criado o endpoint **/dataset/reidentiy** que recebe o GID como *query param*. Uma vez que o GID é uma informação pública, este dado não é criptografado. Ao receber o pseudônimo global, o Honest Broker busca, em seu banco de dados (exemplificado na Listagem 5.5), os LIDs e seus provedores de origem. A lista de LIDs e provedores resultante, caso exista, é percorrida e, a cada iteração, são obtidas as informações do provedor, como a URL de seu Local Broker e sua chave pública. Em seguida, o LID é criptografado com a chave pública de seu respectivo Local Broker que é então enviado através do endpoint de reidentificação. Em caso de sucesso, o *subject_id* recebido é descriptografado com a chave privada do Honest Broker e armazenado em uma lista. Após coletar as respostas (ou possível *timeout*) de todos os Brokers Locais da lista, é verificada a consistência das respostas, ou seja, se foi recebido ao menos um valor para o *subject_id* e se todos os valores recebidos são iguais, como ilustrado na Listagem 5.6. Confirmando a verificação, é retornado o identificador como resposta, caso contrário, é

enviada uma mensagem informando que não foram encontradas informações relacionadas ao GID. Esta verificação é para evitar que existam inconstâncias entre os provedores e acabe retornando um valor inválido no endpoint do Honest Broker.

Do lado do Local Broker, foi criado o endpoint **/dataset/reidentify** que recebe o LID, criptografado com a chave pública do provedor, através de *query param*. Assim, quando o endpoint é invocado, o Local Broker descriptografa o LID e faz uma busca pelo pseudônimo no banco de dados. Se encontrado, é então realizada a criptografia do identificador do indivíduo com a chave privada do Honest Broker e enviada como resposta. Caso contrário, é enviada uma mensagem indicando que não foi encontrada nenhuma informação relacionada ao LID enviado.

Listagem 5.6: Trecho de código da verificação dos identificadores recebidos pelos Data Providers.

```
1 if (!ids.isEmpty() && ids.stream().distinct().count() == 1) {
2     return ids.get(0);
3 } else {
4     return null;
5 }
```

Para o controle do teste, foi selecionado um conjunto de indivíduos presentes em, ao menos, dois dos três conjuntos de dados (A, B ou C). A validação do processo de reidentificação foi possível a partir da utilização da coluna *index* que permitiu associar GIDs e LIDs diretamente. Cabe ressaltar que esta coluna foi adicionada artificialmente aos conjuntos de dados apenas para este fim, não sendo gerada em produção. Os GIDs selecionados foram utilizados para a chamada do endpoint de reidentificação no Honest Broker. Por exemplo, foi selecionado o indivíduo com o *subject_id* 10039708 e com valores para a coluna *index*: A [102, 183, 222], B [220, 223, 247] e C [17, 101, 221, 224]. O GID obtido no Data Lake foi e9e1ae8...7cf9dea. Em seguida, foi enviada a requisição **/dataset/reidentify?gid=e9e1ae8...7cf9dea** para o Honest Broker. O mapeamento reverso foi realizado no Honest Broker e nos respectivos Brokers Locais com sucesso e um exemplo de resposta da requisição do Data Lake pode ser visualizado na Listagem 5.7, confirmando o sucesso da operação e indicando que o *subject_id* referente ao GID informado é 10039708.

Listagem 5.7: Resposta do endpoint de reidentificação do Honest Broker.

```
{
    "message": "10039708",
    "is_success": true
}
```

A Figura 24 exibe o fluxo de reidentificação do exemplo citado acima. Primeiro, no passo 1, o Data Lake solicitou ao Honest Broker a reidentificação do GID. Em seguida, no passo 2, o Honest Broker requisitou aos provedores que haviam LIDs ligados ao GID recebido a reidentificação destes LIDs. Com a resposta dos provedores, passo 3, o Honest Broker realizou as validações inerentes ao endpoint e respondeu ao Data Lake com o *subject_id* do indivíduo, passo 4.

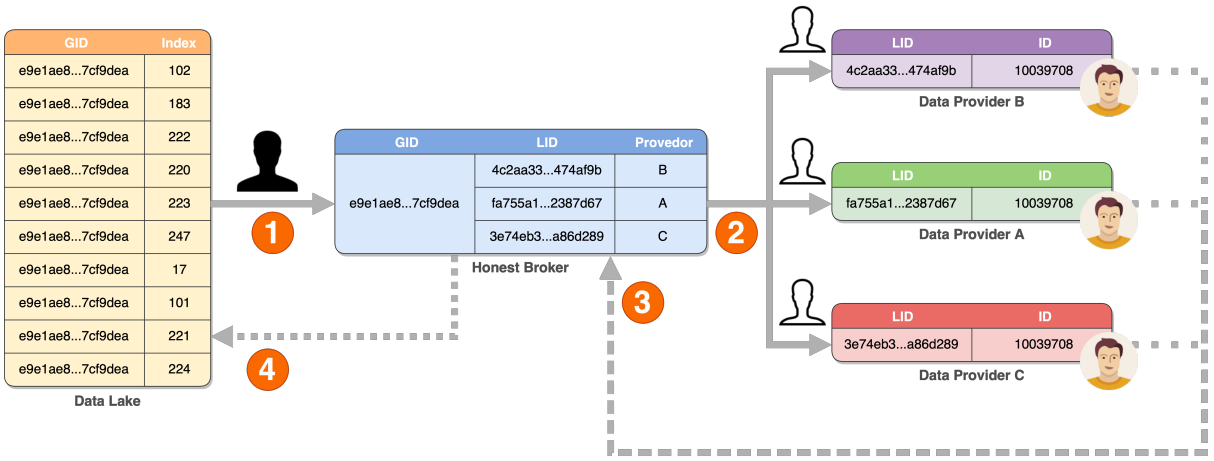


Figura 24: Ligação das informações nos componentes.

Finalmente, com base nos resultados obtidos, é possível concluir que a arquitetura permite a reidentificação correta de um indivíduo a partir dos dados (pseudo)anonimizados no Data Lake.

5.5 Considerações sobre a validação e testes

Com a execução dos quatro diferentes testes apresentados nesta seção foi possível validar os pressupostos da privacidade da arquitetura proposta neste trabalho. Os testes permitiram avaliar pontos importantes da arquitetura, além de demonstrar limitações e futuras melhorias da implementação da arquitetura proposta, sendo um processo importante para este trabalho.

O primeiro teste demonstrou e concluiu que o processo de anonimização dos dados na arquitetura implementada, utilizando a biblioteca do ARX, produz um resultado idêntico ao produzido pela ferramenta ARX. O segundo teste verificou o processo de geração dos pseudônimos local e global com múltiplos Data Providers, apresentando o passo a passo da geração e ligação destes pseudônimos e concluindo que o processo implementado na arquitetura proposta produz o resultado correto e esperado. O terceiro teste atestou e avaliou o funcionamento do processo de (pseudo)anonimização dos dados fim a fim, onde o teste exercitou os diferentes componentes da arquitetura proposta e revelou limitações e possíveis melhorias na sua implementação, além de concluir que apesar das limitação relatada no teste, a geração de um conjunto de dados (pseudo)anonimizado fim a fim produz um resultado correto. Por fim, o quarto e último teste explorou o cenário de reidentificação de um indivíduo a partir dos dados consolidados no Data Lake e concluiu que a arquitetura proposta permite a reidentificação correta de um indivíduo a partir dos dados (pseudo)anonimizados no Data Lake.

6 Considerações finais

Este trabalho teve como principal desafio compatibilizar a maximização na utilização dos dados com a garantia da privacidade dos titulares dos dados, sob a premissa do uso de técnicas e processos de desidentificação, com foco na (pseudo)anonimização, como solução viável. Embora contextualizado no âmbito do projeto SigSaúde e, portanto, voltado para dados em saúde, a discussão e resultados apresentados neste trabalho podem ser mapeados aos mais diferentes contextos atuais, pois, a cada dia, os dados adquirem maior valor e importância, tendo como consequência um aumento exponencial de sua coleta e utilização. Porém, a coleta, processamento e utilização indiscriminados têm gerado uma preocupação global e cada vez maior sobre o seu impacto na privacidade dos indivíduos a quem se refere esses dados. Evidência disto é a profusão de legislações específicas, em diversos países, criadas para proteger a privacidade de seus cidadãos.

O foco na área da saúde então justifica-se não apenas pelo fato de que o tema da privacidade e uso de dados pessoais, toma uma dimensão especial, uma vez que são coletados, processados e armazenados dados pessoais e sensíveis dos pacientes, mas também por evidenciar o desafio deste trabalho. Assim, por uma lado, manter estas informações seguras torna-se crucial para prática da saúde bem-sucedida, pois a divulgação ou vazamento destes dados sensíveis podem acarretar em danos sociais e até econômicos aos pacientes. Enquanto que, por outro lado, o uso secundário dos dados traz grandes benefícios para todo sistema de saúde, permitindo estudos clínicos mais abrangentes, desenvolvimento de novas técnicas e medicações, etc.

Uma primeira contribuição deste trabalho é a apresentação de um estudo do estado-da-arte em técnicas de (pseudo)anonimização e modelos de privacidade, podendo ser utilizado como referência para outros trabalhos da área. De forma a complementar esta contribuição, este trabalho também reúne informações sobre um conjunto relevante de ferramentas e projetos de desidentificação de dados. Deste estudo, vale aqui ressaltar que a maior parte das soluções estudadas apresentam apenas um nível de anonimização, pois consideram um contexto local, onde a sua execução é restrita a apenas um provedor

de dados. Apenas uma solução apresenta, de forma similar a proposta neste trabalho, a aplicação de (pseudo)anonimização em dois níveis. Além disto, a maioria das ferramentas apresenta apenas mecanismos para anonimização dos dados, sendo pouco abordado o tema da pseudonimização.

Uma outra contribuição importante deste trabalho é o desenho de uma arquitetura de (pseudo)anonimização que permite (pseudo)anonimizar dados em saúde e, consequentemente, viabilizar o seu compartilhamento e uso secundário. São três os pontos de destaque da arquitetura proposta: (i) os dados são sempre (pseudo)anonimizados na fonte, ou seja, ainda no Data Provider; (ii) permite juntar dados (pseudo)anonimizados de múltiplas fontes, uma vez que aplica dois níveis de (pseudo)anonimização dos dados, garantindo que um provedor não seja capaz de identificar um indivíduo seja a partir dos dados de outro provedor ou dos dados expostos no Data Lake; e (iii) o processo de (pseudo)anonimização a partir de múltiplas fontes é uniformizado, mantendo a definição dos parâmetros a serem utilizados de forma centralizada no Data Lake através dos *templates*.

Como contribuição final, foi implementada uma Prova de Conceito (PoC) da arquitetura proposta, a fim de exercitar as principais funcionalidades dos três elementos principais (Data Provider, Local Broker e Honest Broker) e dos fluxos definidos na arquitetura. A PoC foi submetida a um conjunto de testes e validações realizados, obtendo resultados encorajadores e, principalmente, demonstrando a viabilidade da solução proposta. Os testes demonstraram que é possível reunir conjuntos de dados (pseudo)anonimizados, de diferentes fontes, em um Data Lake centralizado, de forma segura e mantendo a privacidade dos indivíduos. Como exercício adicional, foi validada a capacidade de reidentificação da arquitetura a partir dos dados publicados no Data Lake. Cabe relembrar que o processo de reidentificação faz todo o sentido no contexto do trabalho, particularmente em situações que envolvam a segurança ou melhoria da condição de saúde dos titulares dos dados (pseudo)anonimizados.

Por fim, o trabalho cumpre seu objetivo geral ao entregar o desenho e a prototipagem de uma arquitetura para o compartilhamento de dados de saúde para uso secundário com a preservação da privacidade dos indivíduos inicialmente relacionados aos dados. Além de facilitar a integração com sistemas de terceiros ao fornecer interfaces API REST simples e padronizadas.

6.1 Limitações e trabalhos futuros

No tempo de implementação da arquitetura, deu-se prioridade aos blocos fundamentais para o seu funcionamento. Com isso, alguns componentes e funcionalidades importantes, mas não essenciais à demonstração do funcionamento da arquitetura, não foram implementados nesta fase, mas devem ser objeto de estudo em trabalhos futuros.

O Teste 3, descrito na Seção 5.3, revelou que o armazenamento dos conjuntos de dados (pseudo)anonimizados é limitado a um tamanho de 16 MB. Esta é uma limitação da tecnologia de banco de dados utilizada (MongoDB), que limita o tamanho de documento BSON em 16 MB, conforme foi implementado. Entre as possíveis medidas de mitigação do problema, estão: (i) a alteração da forma de armazenamento no MongoDB, passando a salvar os conjuntos de dados em lotes ou como arquivos de maneira particionada utilizando a funcionalidade GridFS do MongoDB; e (ii) o uso do armazenamento em disco, embora isso possa causar uma queda significativa de desempenho geral. Neste sentido, vislumbra-se como trabalho futuro a realização de um estudo mais aprofundado de soluções de armazenamento para Data Lake que permitam lidar com conjuntos de dados maiores.

O tempo de gravação do conjunto de dados (pseudo)anonimizado no Local Broker, evidenciado no Teste 3, mostrou-se um gargalo no envio dos dados fim a fim, consumindo, em média, 60% do tempo total. Alternativas para diminuir o impacto deste processo devem ser avaliadas, incluindo, a alteração da lógica implementada para que o processo de gravação do conjunto de dados (pseudo)anonimizado ocorra de forma assíncrona, não sendo necessária a sua finalização para que seja confirmado o envio e processamento dos dados ao provedor. Importante ressaltar que ao gravar os dados assincronamente, deve ser previsto um mecanismo para alertar o provedor em caso de erro, como por exemplo o encerramento inesperado do Local Broker, pois isso levaria à perda de dados.

O uso da biblioteca do ARX para o processo de anonimização dos dados na implementação do Local Broker se mostrou consistente e eficiente nos testes e validações realizados no trabalho. Porém, apesar do ARX ser muito mais escalável em comparação com outras soluções, têm a sua utilização restrita a conjuntos de dados de pequeno e médio porte, ou seja, contendo poucos milhões de registros e com um máximo de 50 atributos quase-identificadores. Sendo assim, é interessante pensar em outras alternativas para o processo de anonimização da arquitetura implementada, inclusive, sendo uma melhoria futura interessante, a implementação própria de um motor de anonimização.

Um outro ponto que merece atenção e futuras melhorias diz respeito a autenticação

e o controle de acesso de usuários e APIs. Estes controles não foram implementados na PoC realizada, tendo sido simulados nos testes, através de scripts. Felizmente, devido à modularidade da arquitetura, pode ser facilmente acoplado um gerenciador de identidade e acesso externo, como o Keycloak[1].

Um outro ponto de melhoria que certamente impactará na usabilidade e adoção da solução proposta diz respeito à disponibilização de uma interface gráfica do usuário no Local Broker e no Data Lake, permitindo a configuração e utilização dos componentes, como por exemplo nas operações de criação de um *template*, envio de conjuntos de dados para (pseudo)anonimização e consulta dos dados publicados.

Na arquitetura implementada, ao criar um *template* é possível definir um limite máximo de risco, porém, durante o processo de anonimização este valor não é considerado. Assim, vislumbra-se como um trabalho futuro interessante a implementação do componente Analisador de Risco de Reidentificação no Local Broker, permitindo ajustar o risco dos dados (pseudo)anonimizados ao risco máximo indicado no *template*.

O armazenamento e geração de logs de forma persistente e estruturada não foi implementado. Diante disto, entende-se como trabalho futuro a integração com ferramentas de serviços de log como Fluentd, Filebeat ou Logstash, e até mesmo de visualização dos logs como Prometheus, Kibana ou Grafana. Isto permitirá auditar, analisar e monitorar as operações executadas e as possíveis exceções lançadas.

Um trabalho futuro promissor é a análise e estudo sobre a segurança da arquitetura. A arquitetura como um todo deve ser analisada e discutida no âmbito da segurança, verificando a segurança dos componentes e de suas comunicações. Cenários e testes de ataques também devem ser explorados.

Por fim, cabe ressaltar que as limitações aqui discutidas não comprometem o funcionamento dos componentes principais da arquitetura e também não impactam diretamente sobre os resultados obtidos nos testes e validações. No entanto, devem ser devidamente tratadas antes da utilização da solução em ambiente de produção.

[1] https://www.keycloak.org/

Referências

Aamot et al. 2013 AAMOT, H. et al. Pseudonymization of patient identifiers for translational research. *BMC medical informatics and decision making*, BioMed Central, v. 13, n. 1, p. 1–15, 2013. Disponível em: <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-75>.

Aircloak 2022 AIRCLOAK. *Aircloak: Peace of Mind – Immediate Insights*. 2022. Disponível em: <https://aircloak.com/>.

AlSalamah, Cámara e Kelly 2018 ALSALAMAH, A. K.; CÁMARA, J. M. S.; KELLY, S. Applying virtualization and containerization techniques in cybersecurity education. In: *Proceedings of the 34th Information Systems Education Conference, ISECON*. [s.n.], 2018. p. 1–14. Disponível em: <https://www.researchgate.net/profile/Arwa-Alsalamah/publication/337929211_Applying_Virtualization_and_Containerization_Techniques_in_Cybersecurity_Education/links/5e3e610192851c7f7f2603c7/Applying-Virtualization-and-Containerization-Techniques-in-Cybersecurity-Education.pdf>.

Amnesia 2022 AMNESIA. *Amnesia Anonymization Tool*. 2022. Disponível em: <https://amnesia.openaire.eu/>.

Anonimatron 2022 ANONIMATRON. *Anonimatron. Providing GDPR compliance since 2010.* 2022. Disponível em: <https://realrolfje.github.io/anonimatron/>.

ARGUS 2022 ARGUS μ. *μ-ARGUS - Research*. 2022. Disponível em: <https://research.cbs.nl/casc/mu.htm>.

Balusamy e Muthusundari 2014 BALUSAMY, M.; MUTHUSUNDARI, S. Data anonymization through generalization using map reduce on cloud. In: *Proceedings of IEEE International Conference on Computer Communication and Systems ICCCS14*. [s.n.], 2014. p. 039–042. Disponível em: <https://ieeexplore.ieee.org/document/7068164>.

Bellavista e Zanni 2017 BELLAVISTA, P.; ZANNI, A. Feasibility of fog computing deployment based on docker containerization over raspberrypi. In: *Proceedings of the 18th international conference on distributed computing and networking*. [s.n.], 2017. p. 1–10. Disponível em: <https://dl.acm.org/doi/10.1145/3007748.3007777>.

BizDataX 2022 BIZDATAX. *BizDataX: Data Masking for Managing Test ...* 2022. Disponível em: <https://bizdatax.com/>.

Boeck et al. 2021 BOECK, K. D. et al. Dataset anonymization with purpose: a resource allocation use case. In: IEEE. *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*. 2021. p. 202–210. Disponível em: <https://ieeexplore.ieee.org/document/9644331/>.

Brito e Machado 2017 BRITO, F. T.; MACHADO, J. C. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. *Jornadas de A tualização em Informática*, 2017. Disponível em: <https://www.researchgate.net/publication/318726149_Preservacao_de_Privacidade_de_Dados_Fundamentos_Tecnicas_e_Aplicacoes>.

Butler et al. 2018 BUTLER, J. et al. *D8.03 "Proof-of-Principle" study: SOP HARMONY Anonymization Procedure*. [S.l.], 2018. Disponível em: <https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bf59eb2f&appId=PPGMS>.

Byun et al. 2007 BYUN, J.-W. et al. Efficient k-anonymization using clustering techniques. In: SPRINGER. *International Conference on Database Systems for Advanced Applications*. 2007. p. 188–200. Disponível em: <https://doi.org/10.1007/978-3-540-71703-4_18>.

Cesário et al. 2022 CESáRIO, H. et al. Arquitetura para gerenciamento de dispositivos através de assistentes virtuais comandados por voz. In: *Anais do VI Workshop de Computação Urbana*. Porto Alegre, RS, Brasil: SBC, 2022. p. 1–14. ISSN 2595-2706.

CloverDX 2022 CLOVERDX. *CloverDX | Solve demanding, real-world data challenges*. 2022. Disponível em: <https://www.cloverdx.com/>.

Crutzen, Peters e Mondschein 2019 CRUTZEN, R.; PETERS, G.-J. Y.; MONDSCHEIN, C. Why and how we should care about the general data protection regulation. *Psychology & health*, Taylor & Francis, v. 34, n. 11, p. 1347–1357, 2019. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/08870446.2019.1606222>.

Dai et al. 2009 DAI, C. et al. Tiamat: a tool for interactive analysis of microdata anonymization techniques. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 2, n. 2, p. 1618–1621, 2009. Disponível em: <https://dl.acm.org/doi/10.14778/1687553.1687607>.

Dallas 2022 DALLAS, T. U. of Texas at. *UTD Anonymization Toolbox*. 2022. Disponível em: <http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php>.

Docbyte 2022 DOCBYTE. *Intelligent Document Processing Solution Anonymization*. 2022. Disponível em: <https://www.docbyte.com/solutions/anonymization/>.

Domingo-Ferrer, Martínez e Sánchez 2022 DOMINGO-FERRER, J.; MARTÍNEZ, S.; SáNCHEZ, D. Decentralized k-anonymization of trajectories via privacy-preserving tit-for-tat. *Comput. Commun.*, Elsevier Science Publishers B. V., NLD, v. 190, n. C, p. 57–68, jun 2022. ISSN 0140-3664. Disponível em: <https://doi.org/10.1016/j.comcom.2022.04.011>.

Domingo-Ferrer e Soria-Comas 2015 DOMINGO-FERRER, J.; SORIA-COMAS, J. From t-closeness to differential privacy and vice versa in data anonymization. *Knowledge-Based Systems*, v. 74, p. 151–158, 2015. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705114004031>.

Dubagunta, van Son e Magimai.-Doss 2022 DUBAGUNTA, S. P.; van Son, R. J.; MAGIMAI.-DOSS, M. Adjustable deterministic pseudonymization of speech. *Computer Speech & Language*, v. 72, p. 101284, 2022. ISSN 0885-2308. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0885230821000851>.

esito 2022 ESITO. *g9 Anonymizer - Database anonymization tool*. 2022. Disponível em: <https://www.esito.no/en/products/anonymizer/>.

Fernandes et al. 2020 FERNANDES, R. et al. S3as: uma solução de autenticação e autorização através de aplicativos de smartphones. *Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação*, v. 3, n. 1, 2020. ISSN 2446-7634.

Fernandes et al. 2019 FERNANDES, R. et al. SAAS: Uma Solução de Autenticação para Aplicativos de Smartphones. In: *4o Workshop Regional de Segurança da Informação e de Sistemas Computacionais*. Alegrete-RS, Brasil: [s.n.], 2019.

Ferrao et al. 2019 FERRAO, I. et al. Urnas eletrônicas no brasil: linha do tempo, evolução e falhas e desafios de segurança. *Revista Brasileira de Computação Aplicada*, v. 11, n. 2, p. 1–12, maio 2019.

Filho et al. 2020 FILHO, I. B. et al. Development of a health dashboard for an electronic health record system. In: IEEE. *2020 20th International Conference on Computational Science and Its Applications (ICCSA)*. 2020. p. 16–22. Disponível em: <https://ieeexplore.ieee.org/document/9257530>.

Fiorenza et al. 2021 FIORENZA, M. et al. Representação e aplicação de políticas de segurança em firewalls de redes híbridas. In: *Anais do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Porto Alegre, RS, Brasil: SBC, 2021. p. 490–503. ISSN 2177-9384.

Fredj, Lammari e Comyn-Wattiau 2015 FREDJ, F. B.; LAMMARI, N.; COMYN-WATTIAU, I. Abstracting anonymization techniques: a prerequisite for selecting a generalization algorithm. *Procedia computer science*, Elsevier, v. 60, p. 206–215, 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050915022474>.

Fung et al. 2010 FUNG, B. C. et al. *Introduction to privacy-preserving data publishing: Concepts and techniques*. Chapman and Hall/CRC, 2010. ISBN 9780367383756. Disponível em: <https://doi.org/10.1201/9781420091502>.

Fung, Wang e Yu 2005 FUNG, B. C.; WANG, K.; YU, P. S. Top-down specialization for information and privacy preservation. In: IEEE. *21st international conference on data engineering (ICDE'05)*. 2005. p. 205–216. Disponível em: <https://doi.org/10.1109/ICDE.2005.143>.

Gates e Matthews 2014 GATES, C.; MATTHEWS, P. Data is the new currency. In: *Proceedings of the 2014 New Security Paradigms Workshop*. [s.n.], 2014. p. 105–116. Disponível em: <https://dl.acm.org/doi/10.1145/2683467.2683477>.

Gentili, Hajian e Castillo 2017 GENTILI, M.; HAJIAN, S.; CASTILLO, C. A case study of anonymization of medical surveys. In: *Proceedings of the 2017 International Conference on Digital Health*. [s.n.], 2017. p. 77–81. Disponível em: <https://dl.acm.org/doi/10.1145/3079452.3079490>.

Gkoulalas-Divanis e Loukides 2012 GKOULALAS-DIVANIS, A.; LOUKIDES, G. Utility-guided clustering-based transaction data anonymization. *Trans. Data Priv.*, v. 5, n. 1, p. 223–251, 2012. Disponível em: <http://www.tdp.cat/issues11/tdp.a083a11.pdf>.

Gramener 2022 GRAMENER. *Gramener: Data Science and AI Company*. 2022. Disponível em: <https://gramener.com/>.

Groseth et al. 2019 GROSETH, L. A. et al. *Anonymization as a Service*. Tese (Bachelor thesis) — Oslo Metropolitan University - OsloMet, 2019. Disponível em: <https://oslomet-arx-as-a-service.github.io/resources/Anonymization_as_a_Service_Thesis.pdf>.

Han, Yu e Yu 2008 HAN, J.; YU, H.; YU, J. An improved l-diversity model for numerical sensitive attributes. In: IEEE. *2008 Third International Conference on Communications and Networking in China*. 2008. p. 938–943. Disponível em: <https://doi.org/10.1109/CHINACOM.2008.4685178>.

HARMONY 2022 HARMONY. *Home - HARMONY Healthcare Alliance — harmony-alliance.eu*. 2022. Acessado em: 23-08-2022. Disponível em: <https://www.harmony-alliance.eu/>.

Hore et al. 2021 HORE, B. et al. Contrained generalization for data anonymization-a systematic search based approach. *arXiv*, 2021. Disponível em: <https://arxiv.org/abs/2108.04897>.

Jakob et al. 2020 JAKOB, C. E. et al. Design and evaluation of a data anonymization pipeline to promote open science on covid-19. *Scientific data*, Nature Publishing Group, v. 7, n. 1, p. 1–10, 2020. Disponível em: <https://www.nature.com/articles/s41597-020-00773-y>.

James et al. 2021 JAMES, S. et al. Synthetic data use: exploring use cases to optimise data utility. *Discover Artificial Intelligence*, Springer, v. 1, n. 1, p. 1–13, 2021. Disponível em: <https://link.springer.com/article/10.1007/s44163-021-00016-y>.

Johnson et al. 2020 JOHNSON, N. et al. Chorus: a programming framework for building scalable differential privacy mechanisms. In: IEEE. *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*. 2020. p. 535–551. Disponível em: <https://ieeexplore.ieee.org/document/9230409>.

Journal 2022 JOURNAL, H. *Healthcare Data Breach Statistics - latest data for 2022*. 2022. Disponível em: <https://www.hipaajournal.com/healthcare-data-breach-statistics/>.

Jr, Machado e Monteiro 2014 JR, E. C. B.; MACHADO, J. C.; MONTEIRO, J. M. Estratégias para proteção da privacidade de dados armazenados na nuvem. *(In Proceedings) Simpósio Brasileiro de Banco de Dados (29. : 2014 out. 6-9 : Curitiba, PR). Tópicos em Gerenciamento de Dados e Informações*, UFPR; PUC-PR, v. 1, 2014. ISBN 978-85-7669-290-4. Disponível em: <https://www.inf.ufpr.br/sbbd-sbsc2014/sbbd/proceedings/artigos/pdfs/14.pdf>.

Jyothi e Rao 2017 JYOTHI, M.; RAO, M. Preserving the privacy of sensitive data using data anonymization. *International Journal of Applied Engineering Research*, v. 12, n. 8, p. 1639–1663, 2017. Disponível em: <http://www.ripublication.com/ijaer17/ijaerv12n8_25.pdf>.

Júnior et al. 2022 JúNIOR, O. P. et al. *Guia: LGPD e uso Secundário de Dados de Saúde*. [S.l.], 2022. Disponível em: <https://baptistaluz.com.br/wp-content/uploads/2022/07/Bluz_220726_PD_AYIP_DadosSaude_V3.pdf>.

Kreutz et al. 2020 KREUTZ, D. et al. Auth4app: Protocols for identification and authentication using mobile applications. In: *Anais do XX Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*. Porto Alegre, RS, Brasil: SBC, 2020. p. 422–435. ISSN 0000-0000.

Kulkarni e Bedekar 2022 KULKARNI, S.; BEDEKAR, M. Perception of privacy in a data driven world. *International Journal for Modern Trends in Science and Technology*, v. 8, n. 04, p. 380–388, 2022. Disponível em: <http://www.ijmtst.com/volume8/issue04/64.IJMTST0804182.pdf>.

Lee et al. 2017 LEE, H. et al. Utility-preserving anonymization for health data publishing. *BMC medical informatics and decision making*, Springer, v. 17, n. 1, p. 1–12, 2017. Disponível em: <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0499-0>.

LeFevre, DeWitt e Ramakrishnan 2005 LEFEVRE, K.; DEWITT, D. J.; RAMAKRISHNAN, R. Incognito: Efficient full-domain k-anonymity. In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. [s.n.], 2005. p. 49–60. Disponível em: <https://doi.org/10.1145/1066157.1066164>.

LeFevre, DeWitt e Ramakrishnan 2006 LEFEVRE, K.; DEWITT, D. J.; RAMAKRISHNAN, R. Mondrian multidimensional k-anonymity. In: IEEE. *22nd International conference on data engineering (ICDE'06)*. 2006. p. 25–25. Disponível em: <https://doi.org/10.1109/ICDE.2006.101>.

Li, Li e Venkatasubramanian 2006 LI, N.; LI, T.; VENKATASUBRAMANIAN, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: IEEE. *2007 IEEE 23rd international conference on data engineering*. 2006. p. 106–115. Disponível em: <https://ieeexplore.ieee.org/document/4221659>.

Li, Li e Venkatasubramanian 2007 LI, N.; LI, T.; VENKATASUBRAMANIAN, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In: *2007 IEEE 23rd International Conference on Data Engineering*. [s.n.], 2007. p. 106–115. Disponível em: <https://doi.org/10.1109/ICDE.2007.367856>.

Li e Lai 2017 LI, Z.; LAI, T. H. δ-privacy: Bounding privacy leaks in privacy preserving data mining. In: *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Springer, 2017. p. 124–142. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-67816-0_8>.

Liu, Feng e Zhu 2022 LIU, X.; FENG, X.; ZHU, Y. Transactional data anonymization for privacy and information preservation via disassociation and local suppression. *Symmetry*, MDPI, v. 14, n. 3, p. 472, 2022. Disponível em: <https://www.mdpi.com/2073-8994/14/3/472>.

Loukides, Gkoulalas-Divanis e Malin 2011 LOUKIDES, G.; GKOULALAS-DIVANIS, A.; MALIN, B. Coat: Constraint-based anonymization of transactions. *Knowledge*

and Information Systems, Springer, v. 28, n. 2, p. 251–282, 2011. Disponível em: <https://doi.org/10.1007/s10115-010-0354-4>.

Machanavajjhala et al. 2006 MACHANAVAJJHALA, A. et al. L-diversity: privacy beyond k-anonymity. In: *22nd International Conference on Data Engineering (ICDE'06)*. [s.n.], 2006. p. 24–24. Disponível em: <https://ieeexplore.ieee.org/abstract/document/1617392>.

Maier 2013 MAIER, J. Anonymity: Formalisation of privacy–k-anonymity. In: *Seminar paper, Technische Universität München, Munich*. [s.n.], 2013. Disponível em: <https://www.net.in.tum.de/fileadmin/TUM/NET/NET-2013-08-1/NET-2013-08-1_06.pdf>.

Marques e Bernardino 2020 MARQUES, J. F.; BERNARDINO, J. Analysis of data anonymization techniques. In: *KEOD*. [s.n.], 2020. p. 235–241. Disponível em: <https://www.scitepress.org/Papers/2020/101423/101423.pdf>.

Medicine 2022 MEDICINE, N. L. of. *NLM-Scrubber*. 2022. Disponível em: <https://lhncbc.nlm.nih.gov/scrubber/>.

Medková 2020 MEDKOVÁ, J. High-degree noise addition method for the k-degree anonymization algorithm. In: IEEE. *2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)*. 2020. p. 1–6. Disponível em: <https://ieeexplore.ieee.org/document/9322670>.

Mimoto., Basu. e Kiyomoto. 2016 MIMOTO., T.; BASU., A.; KIYOMOTO., S. Towards practical k-anonymization: Correlation-based construction of generalization hierarchy. In: INSTICC. *Proceedings of the 13th International Joint Conference on e-Business and Telecommunications - SECRYPT, (ICETE 2016)*. SciTePress, 2016. p. 411–418. ISBN 978-989-758-196-0. ISSN 2184-2825. Disponível em: <https://www.scitepress.org/Link.aspx?doi=10.5220/0005963804110418>.

Moor, Claerhout e Meyer 2003 MOOR, G. D.; CLAERHOUT, B.; MEYER, F. D. Privacy enhancing techniques - the key to secure communication and management of clinical and genomic data. *Methods of information in medicine*, Schattauer GmbH, v. 42, n. 02, p. 148–153, 2003. Disponível em: <https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0038-1634326>.

Murthy et al. 2019 MURTHY, S. et al. A comparative study of data anonymization techniques. In: IEEE. *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. 2019. p. 306–309. Disponível em: <https://ieeexplore.ieee.org/document/8819477>.

Nelson 2015 NELSON, G. S. Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification. In: *SAS global forum proceedings*. [s.n.], 2015. p. 1–23. Disponível em: <https://support.sas.com/resources/papers/proceedings15/1884-2015.pdf>.

Neumann et al. 2019 NEUMANN, G. K. et al. Pseudonymization risk analysis in distributed systems. *Journal of Internet Services and Applications*, SpringerOpen, v. 10,

n. 1, p. 1–16, 2019. Disponível em: <https://jisajournal.springeropen.com/articles/10.1186/s13174-018-0098-z>.

OLIVEIRA, MADEIRA e MONTEIRO 2020 OLIVEIRA, E. d.; MADEIRA, H. d. S.; MONTEIRO, P. A. M. *A lei geral de proteção de dados pessoais e a anonimização de dados: uma aplicação da técnica em uma base de dados real.* Dissertação (Trabalho de Conclusão de Curso) — Faculdade de Tecnologia de São Caetano do Sul, 2020. Curso de Segurança da Informação. Disponível em: <http://ric.cps.sp.gov.br/handle/123456789/5258>.

Oliveira 2020 OLIVEIRA, F. N. S. C. d. *Gestão de riscos no direito fundamental à privacidade de dados pessoais no Processo Judicial Eletrônico/Diário de Justiça Eletrônico.* Dissertação (Dissertação (Mestrado Profissional em Computação Aplicada)) — Universidade de Brasília (UnB), 2020. Disponível em: <https://repositorio.unb.br/handle/10482/39152>.

Oliveira et al. 2021 OLIVEIRA, I. et al. dh-aes-p4: On-premise encryption and in-band key-exchange in p4 fully programmable data planes. In: *2021 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN).* [S.l.: s.n.], 2021. p. 148–153.

OpenAIRE 2022 OpenAIRE. *OpenAIRE.* 2022. Acessado em: 23-08-2022. Disponível em: <https://www.openaire.eu/>.

Patki, Wedge e Veeramachaneni 2016 PATKI, N.; WEDGE, R.; VEERAMACHANENI, K. The synthetic data vault. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA).* [s.n.], 2016. p. 399–410. Disponível em: <https://ieeexplore.ieee.org/document/7796926>.

PINHO 2020 PINHO, F. A. *Anonimização de bases de dados empresariais de acordo com a nova Regulamentação Europeia de Proteção de Dados. 2017.* Tese (Doutorado) — Dissertação (Mestrado em Segurança Informática), Departamento de Ciência de Computadores, Faculdade de Ciências, Universidade do Porto., 2020. Disponível em: <https://cracs.fc.up.pt/sites/default/files/MSI_Dissertacao_FINAL.pdf>.

Poulis et al. 2014 POULIS, G. et al. Secreta: A system for evaluating and comparing relational and transaction anonymization algorithms. *Conference Proceedings*, OpenProceedings. org, University of Konstanz, 2014. Disponível em: <http://users.uop.gr/~poulis/SECRETA/PDF/EDBT2014.pdf>.

Poulis et al. 2013 POULIS, G. et al. Anonymizing data with relational and transaction attributes. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* [S.l.], 2013. p. 353–369.

Prasser et al. 2016 PRASSER, F. et al. Lightning: Utility-driven anonymization of high-dimensional data. *Trans. Data Priv.*, v. 9, n. 2, p. 161–185, 2016. Disponível em: <https://dl.acm.org/doi/10.5555/2993206.2993209>.

Prasser et al. 2020 PRASSER, F. et al. Flexible data anonymization using arx—current status and challenges ahead. *Software: Practice and Experience*, v. 50, n. 7, p. 1277–1304, 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2812>.

Raghunathan 2013 RAGHUNATHAN, B. *The complete book of data anonymization: from planning to implementation*. CRC Press, 2013. ISBN 9781439877302. Disponível em: <https://doi.org/10.1201/b13097>.

Rajendran, Jayabalan e Rana 2017 RAJENDRAN, K.; JAYABALAN, M.; RANA, M. E. A study on k-anonymity, l-diversity, and t-closeness techniques. *IJCSNS*, v. 17, n. 12, p. 172, 2017. Disponível em: <http://www.ijirst.org/articles/IJIRSTV6I6015.pdf>.

Rodrigues et al. 2019 RODRIGUES, D. O. et al. *Computação Urbana da Teoria à Prática: Fundamentos, Aplicações e Desafios*. [S.l.], 2019.

SDV 2022 SDV. *The synthetic data vault. put synthetic data to work!* 2022. Https://sdv.dev/. Accessed: 2022-07-20.

Sei et al. 2019 SEI, Y. et al. Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness. *IEEE Transactions on Dependable and Secure Computing*, v. 16, n. 4, p. 580–593, 2019. Disponível em: <https://ieeexplore.ieee.org/document/7913692>.

Senavirathne e Torra 2019 SENAVIRATHNE, N.; TORRA, V. Integral privacy compliant statistics computation. In: PÉREZ-SOLÀ, C. et al. (Ed.). *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Cham: Springer International Publishing, 2019. p. 22–38. ISBN 978-3-030-31500-9. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-31500-9_2>.

Shin e Kim 2021 SHIN, S.-Y.; KIM, H.-S. Data pseudonymization in a range that does not affect data quality: Correlation with the degree of participation of clinicians. *J Korean Med Sci*, The Korean Academy of Medical Sciences, v. 36, n. 44, 2021. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/34783216/>.

SigSaúde 2018 SIGSAÚDE. *Portal do SigSaúde*. 2018. Disponível em: <https://sigsaude.imd.ufrn.br/sigsaude/>. Acesso em: 10 jan 2022.

Silva, Basso e Moraes 2017 SILVA, H. de O.; BASSO, T.; MORAES, R. L. de O. Privacy and data mining: evaluating the impact of data anonymization on classification algorithms. In: IEEE. *2017 13th European Dependable Computing Conference (EDCC)*. 2017. p. 111–116. Disponível em: <https://ieeexplore.ieee.org/document/8123561>.

Simi, Nayaki e Elayidom 2017 SIMI, M. S.; NAYAKI, K. S.; ELAYIDOM, M. S. An extensive study on data anonymization algorithms based on k-anonymity. *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, v. 225, p. 012279, aug 2017. Disponível em: <https://iopscience.iop.org/article/10.1088/1757-899X/225/1/012279>.

Singapore 2018 SINGAPORE, P. D. P. C. *Guide to basic data anonymisation techniques*. 2018. Disponível em: <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf>.

Soria-Comas et al. 2015 SORIA-COMAS, J. et al. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, v. 27, n. 11, p. 3098–3110, 2015. Disponível em: <https://ieeexplore.ieee.org/document/7110607>.

Stenersen 2020 STENERSEN, H. W. *Anonymization of Health Data*. Dissertação (Mestrado) — University of Oslo, 2020. Disponível em: <https://www.duo.uio.no/bitstream/handle/10852/79902/13/Anonymization-of-Health-Data.pdf>.

Stepanova e Jechel 2021 STEPANOVA, O.; JECHEL, P. *The Privacy, Data Protection and cybersecurity law review: Germany*. 2021. Disponível em: <https://thelawreviews.co.uk/title/the-privacy-data-protection-and-cybersecurity-law-review/germany>.

Sweeney 1997 SWEENEY, L. Guaranteeing anonymity when sharing medical data, the datafly system. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the AMIA Annual Fall Symposium*. 1997. p. 51. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233452/>.

Sweeney 2002 SWEENEY, L. K-anonymity: A model for protecting privacy. World Scientific Publishing Co., Inc., USA, v. 10, n. 5, p. 557–570, oct 2002. ISSN 0218-4885. Disponível em: <https://dl.acm.org/doi/10.1142/S0218488502001648>.

Templ, Kowarik e Meindl 2015 TEMPL, M.; KOWARIK, A.; MEINDL, B. Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, v. 67, n. 4, p. 1–36, 2015. Disponível em: <https://www.jstatsoft.org/article/view/v067i04>.

Terrovitis, Mamoulis e Kalnis 2011 TERROVITIS, M.; MAMOULIS, N.; KALNIS, P. Local and global recoding methods for anonymizing set-valued data. *The VLDB Journal*, Springer, v. 20, n. 1, p. 83–106, 2011. Disponível em: <https://doi.org/10.1007/s00778-010-0192-8>.

Tomás, Rasteiro e Bernardino 2022 TOMÁS, J.; RASTEIRO, D.; BERNARDINO, J. Data anonymization: An experimental evaluation using open-source tools. *Future Internet*, Multidisciplinary Digital Publishing Institute, v. 14, n. 6, p. 167, 2022. Disponível em: <https://www.mdpi.com/1999-5903/14/6/167>.

Virupaksha e Dondeti 2021 VIRUPAKSHA, S.; DONDETI, V. Anonymized noise addition in subspaces for privacy preserved data mining in high dimensional continuous data. *Peer-to-Peer Networking and Applications*, Springer, v. 14, n. 3, p. 1608–1628, 2021. Disponível em: <https://link.springer.com/article/10.1007/s12083-021-01080-y>.

Vovk, Piho e Ross 2021 VOVK, O.; PIHO, G.; ROSS, P. Evaluation of anonymization tools for health data. In: SPRINGER. *International Conference on Model and Data Engineering*. 2021. p. 302–313. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-87657-9_23>.

Walonoski et al. 2017 WALONOSKI, J. et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, v. 25, n. 3, p. 230–238, 08 2017. ISSN 1527-974X. Disponível em: <https://doi.org/10.1093/jamia/ocx079>.

Xiao e Tao 2006 XIAO, X.; TAO, Y. Anatomy: Simple and effective privacy preservation. In: *Proceedings of the 32nd international conference on Very large data bases*. [s.n.], 2006. p. 139–150. Disponível em: <https://www.cse.cuhk.edu.hk/~taoyf/paper/vldb06.pdf>.

Xiao, Wang e Gehrke 2009 XIAO, X.; WANG, G.; GEHRKE, J. Interactive anonymization of sensitive data. In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. [s.n.], 2009. p. 1051–1054. Disponível em: <https://dl.acm.org/doi/10.1145/1559845.1559979>.

Zuo et al. 2021 ZUO, Z. et al. Data anonymization for pervasive health care: Systematic literature mapping study. *JMIR medical informatics*, JMIR Publications Inc., Toronto, Canada, v. 9, n. 10, p. e29871, 2021. Disponível em: <https://medinform.jmir.org/2021/10/e29871>.

ANEXO A – Conjunto de Dados *ADULT*

Na Tabela 5 são apresentados os primeiros 13 registros do conjunto de dados *ADULT*, do conjunto de exemplos da ferramenta ARX, disponível no repositório do github através do endereço <https://github.com/arx-deidentifier/arx/blob/master/data/adult.csv>. O conjunto de dados *ADULT* foi utilizado no teste de validação do mecanismo de anonimização, descrito na Seção 5.1 do Capítulo 5.

Tabela 5: Registros iniciais do conjunto de dados *ADULT*.

sex	age	race	marital-status	education	native-country	workclass	occupation	salary-class
Male	39	White	Never-married	Bachelors	United-States	State-gov	Adm-clerical	=50K
Male	50	White	Married-civ-spouse	Bachelors	United-States	Self-emp-not-inc	Exec-managerial	=50K
Male	38	White	Divorced	HS-grad	United-States	Private	Handlers-cleaners	=50K
Male	53	Black	Married-civ-spouse	11th	United-States	Private	Handlers-cleaners	=50K
Female	28	Black	Married-civ-spouse	Bachelors	Cuba	Private	Prof-specialty	=50K
Female	37	White	Married-civ-spouse	Masters	United-States	Private	Exec-managerial	=50K
Female	49	Black	Married-spouse-absent	9th	Jamaica	Private	Other-service	=50K
Male	52	White	Married-civ-spouse	HS-grad	United-States	Self-emp-not-inc	Exec-managerial	50K
Female	31	White	Never-married	Masters	United-States	Private	Prof-specialty	50K
Male	42	White	Married-civ-spouse	Bachelors	United-States	Private	Exec-managerial	50K
Male	37	Black	Married-civ-spouse	Some-college	United-States	Private	Exec-managerial	50K
Male	30	Asian-Pac-Islander	Married-civ-spouse	Bachelors	India	State-gov	Prof-specialty	50K
Female	23	White	Never-married	Bachelors	United-States	Private	Adm-clerical	=50K

ANEXO B – Conjunto de Dados *ADMISSIONS*

Na Listagem B.1 são apresentados os primeiros 6 registros do conjunto de dados *ADMISSIONS*, da versão de demonstração do MIMIC-IV, disponível em seu site através do endereço <https://physionet.org/content/mimic-iv-demo/1.0/core/admissions.csv.gz>. O conjunto de dados *ADMISSIONS* foi utilizado no teste de validação da geração dos pseudônimos local e global com múltiplos provedores de dados, descrito na Seção 5.2 do Capítulo 5.

Listagem B.1: Registros iniciais do conjunto de dados *ADMISSIONS*.

```
1 subject_id,hadm_id,admittime,dischtime,deathtime,admission_type,
    ↪ admission_location,discharge_location,insurance,language,marital_status,
    ↪ ethnicity,edregtime,edouttime,hospital_expire_flag
2 10004235,24181354,2196-02-24 14:38:00,2196-03-04 14:02:00,,URGENT,TRANSFER
    ↪ FROM HOSPITAL,SKILLED NURSING
    ↪ FACILITY,Medicaid,ENGLISH,SINGLE,BLACK/AFRICAN AMERICAN,2196-02-24
    ↪ 12:15:00,2196-02-24 17:07:00,0
3 10009628,25926192,2153-09-17 17:08:00,2153-09-25 13:20:00,,URGENT,TRANSFER
    ↪ FROM HOSPITAL,HOME HEALTH CARE,Medicaid,?,MARRIED,HISPANIC/LATINO,,,0
4 10018081,23983182,2134-08-18 02:02:00,2134-08-23 19:35:00,,URGENT,TRANSFER
    ↪ FROM HOSPITAL,SKILLED NURSING
    ↪ FACILITY,Medicare,ENGLISH,MARRIED,WHITE,2134-08-17 16:24:00,2134-08-18
    ↪ 03:15:00,0
5 10006053,22942076,2111-11-13 23:39:00,2111-11-15 17:20:00,2111-11-15
    ↪ 17:20:00,URGENT,TRANSFER FROM
    ↪ HOSPITAL,DIED,Medicaid,ENGLISH,,UNKNOWN,,,1
6 10031404,21606243,2113-08-04 18:46:00,2113-08-06 20:57:00,,URGENT,TRANSFER
    ↪ FROM HOSPITAL,HOME,Other,ENGLISH,WIDOWED,WHITE,,,0
7 10005817,20626031,2132-12-12 01:43:00,2132-12-20 15:04:00,,URGENT,TRANSFER
    ↪ FROM HOSPITAL,HOME HEALTH CARE,Medicare,ENGLISH,MARRIED,WHITE,,,0
```

ANEXO C – Conjunto de Dados *OBSERVATIONS*

Na Listagem C.1 são apresentados os primeiros 7 registros do conjunto de dados *OBSERVATIONS* do COVID-19 na versão 10K, gerados pela Synthea™, disponível em seu site através do endereço <https://synthea.mitre.org/downloads>. O conjunto de dados *OBSERVATIONS* foi utilizado no teste de geração de um conjunto de dados (pseudo)anonimizado fim a fim, descrito na Seção 5.3 do Capítulo 5.

Listagem C.1: Registros iniciais do conjunto de dados *OBSERVATIONS*.

```
1 DATE,PATIENT,ENCOUNTER,CODE,DESCRIPTION,VALUE,UNITS,TYPE
2 2019-08-01,f0f3bc8d-ef38-49ce-a2bd-dfdda982b271,
    ↪ 6a74fdef-2287-44bf-b9e7-18012376faca,8302-2,Body Height,82.7,cm,numeric
3 2019-08-01,f0f3bc8d-ef38-49ce-a2bd-dfdda982b271,
    ↪ 6a74fdef-2287-44bf-b9e7-18012376faca,72514-3,Pain severity - 0-10
    ↪ verbal numeric rating [Score] - Reported,2.0,{score},numeric
4 2019-08-01,f0f3bc8d-ef38-49ce-a2bd-dfdda982b271,
    ↪ 6a74fdef-2287-44bf-b9e7-18012376faca,29463-7,Body Weight,12.6,kg,numeric
5 2019-08-01,f0f3bc8d-ef38-49ce-a2bd-dfdda982b271,
    ↪ 6a74fdef-2287-44bf-b9e7-18012376faca,77606-2,Weight-for-length Per age
    ↪ and sex,86.1,%,numeric
6 2019-08-01,f0f3bc8d-ef38-49ce-a2bd-dfdda982b271,
    ↪ 6a74fdef-2287-44bf-b9e7-18012376faca,9843-4,Head Occipital-frontal
    ↪ circumference,46.9,cm,numeric
7 2019-08-01,f0f3bc8d-ef38-49ce-a2bd-dfdda982b271,
    ↪ 6a74fdef-2287-44bf-b9e7-18012376faca,8462-4,Diastolic Blood
    ↪ Pressure,76.0,mm[Hg],numeric
8 2019-08-01,f0f3bc8d-ef38-49ce-a2bd-dfdda982b271,
    ↪ 6a74fdef-2287-44bf-b9e7-18012376faca,8480-6,Systolic Blood
    ↪ Pressure,107.0,mm[Hg],numeric
```

ANEXO D – Template para o conjunto de dados *ADULT*

Na Listagem D.1 é apresentado o template para o conjunto de dados *ADULT* usado no Teste 1, descrito na Seção 5.1 do Capítulo 5, em formato JSON.

Listagem D.1: Template criado para o Teste 1.

```
{
    "name": "Teste 1 - ARX",
    "description": "Template de teste da arquitetura",
    "organization": "ppgti",
    "contact": {"email": "teste@teste.com" },
    "privacy": {
        "kanonimity_value": 5,
        "ldiversity_value": 2,
        "max_risk_value": 0.3,
        "suppression_value": 0.2,
        "principal_identifying": "cpf",
        "attributes": [
            {"name": "cpf", "type": "IDENTIFYING" },
            {"name": "sex", "type": "QUASI_IDENTIFYING",
              "hierarchy": [["Male", "*" ],["Female", "*" ]]},
            {"name": "age", "type": "QUASI_IDENTIFYING",
              "hierarchy": [["1", "0-4", "0-9", "0-19", "*" ],
                            ["2", "0-4", "0-9", "0-19", "*" ],
                            ["3", "0-4", "0-9", "0-19", "*" ],
                            ["4", "0-4", "0-9", "0-19", "*" ],
                            ["5", "0-4", "0-9", "0-19", "*" ],
                            ["6", "5-9", "0-9", "0-19", "*" ],
                            ["7", "5-9", "0-9", "0-19", "*" ],
                            ["8", "5-9", "0-9", "0-19", "*" ],
```

```
                ["9", "5-9", "0-9", "0-19", "*" ],
                ["10", "5-9", "0-9", "0-19", "*" ],
                ["11", "10-14", "10-19", "0-19", "*" ],
                ["12", "10-14", "10-19", "0-19", "*" ],
                ["13", "10-14", "10-19", "0-19", "*" ],
                ["14", "10-14", "10-19", "0-19", "*" ],
                ["15", "10-14", "10-19", "0-19", "*" ],
                ["16", "15-19", "10-19", "0-19", "*" ],
                ["17", "15-19", "10-19", "0-19", "*" ],
                ["18", "15-19", "10-19", "0-19", "*" ],
                ["19", "15-19", "10-19", "0-19", "*" ],
                ["20", "15-19", "10-19", "0-19", "*" ],
                ["21", "20-24", "20-29", "20-39", "*" ],
                ["22", "20-24", "20-29", "20-39", "*" ],
                ["23", "20-24", "20-29", "20-39", "*" ],
                ["24", "20-24", "20-29", "20-39", "*" ],
                ["25", "20-24", "20-29", "20-39", "*" ],
                ["26", "25-29", "20-29", "20-39", "*" ],
                ["27", "25-29", "20-29", "20-39", "*" ],
                ["28", "25-29", "20-29", "20-39", "*" ],
                ["29", "25-29", "20-29", "20-39", "*" ],
                ["30", "25-29", "20-29", "20-39", "*" ],
                ["31", "30-34", "30-39", "20-39", "*" ],
                ["32", "30-34", "30-39", "20-39", "*" ],
                ["33", "30-34", "30-39", "20-39", "*" ],
                ["34", "30-34", "30-39", "20-39", "*" ],
                ["35", "30-34", "30-39", "20-39", "*" ],
                ["36", "35-39", "30-39", "20-39", "*" ],
                ["37", "35-39", "30-39", "20-39", "*" ],
                ["38", "35-39", "30-39", "20-39", "*" ],
                ["39", "35-39", "30-39", "20-39", "*" ],
                ["40", "35-39", "30-39", "20-39", "*" ],
                ["41", "40-44", "40-49", "40-59", "*" ],
                ["42", "40-44", "40-49", "40-59", "*" ],
                ["43", "40-44", "40-49", "40-59", "*" ],
                ["44", "40-44", "40-49", "40-59", "*" ],
                ["45", "40-44", "40-49", "40-59", "*" ],
                ["46", "45-49", "40-49", "40-59", "*" ],
```

```
                ["47", "45-49", "40-49", "40-59", "*" ],
                ["48", "45-49", "40-49", "40-59", "*" ],
                ["49", "45-49", "40-49", "40-59", "*" ],
                ["50", "45-49", "40-49", "40-59", "*" ],
                ["51", "50-54", "50-59", "40-59", "*" ],
                ["52", "50-54", "50-59", "40-59", "*" ],
                ["53", "50-54", "50-59", "40-59", "*" ],
                ["54", "50-54", "50-59", "40-59", "*" ],
                ["55", "50-54", "50-59", "40-59", "*" ],
                ["56", "55-59", "50-59", "40-59", "*" ],
                ["57", "55-59", "50-59", "40-59", "*" ],
                ["58", "55-59", "50-59", "40-59", "*" ],
                ["59", "55-59", "50-59", "40-59", "*" ],
                ["60", "55-59", "50-59", "40-59", "*" ],
                ["61", "60-64", "60-69", "60-79", "*" ],
                ["62", "60-64", "60-69", "60-79", "*" ],
                ["63", "60-64", "60-69", "60-79", "*" ],
                ["64", "60-64", "60-69", "60-79", "*" ],
                ["65", "60-64", "60-69", "60-79", "*" ],
                ["66", "65-69", "60-69", "60-79", "*" ],
                ["67", "65-69", "60-69", "60-79", "*" ],
                ["68", "65-69", "60-69", "60-79", "*" ],
                ["69", "65-69", "60-69", "60-79", "*" ],
                ["70", "65-69", "60-69", "60-79", "*" ],
                ["71", "70-74", "70-79", "60-79", "*" ],
                ["72", "70-74", "70-79", "60-79", "*" ],
                ["73", "70-74", "70-79", "60-79", "*" ],
                ["74", "70-74", "70-79", "60-79", "*" ],
                ["75", "70-74", "70-79", "60-79", "*" ],
                ["76", "75-79", "70-79", "60-79", "*" ],
                ["77", "75-79", "70-79", "60-79", "*" ],
                ["78", "75-79", "70-79", "60-79", "*" ],
                ["79", "75-79", "70-79", "60-79", "*" ],
                ["80", "75-79", "70-79", "60-79", "*" ],
                ["81", "80-84", "80-89", "80-99", "*" ],
                ["82", "80-84", "80-89", "80-99", "*" ],
                ["83", "80-84", "80-89", "80-99", "*" ],
                ["84", "80-84", "80-89", "80-99", "*" ],
```

```
                    ["85", "80-84", "80-89", "80-99", "*" ],
                    ["86", "85-89", "80-89", "80-99", "*" ],
                    ["87", "85-89", "80-89", "80-99", "*" ],
                    ["88", "85-89", "80-89", "80-99", "*" ],
                    ["89", "85-89", "80-89", "80-99", "*" ],
                    ["90", "85-89", "80-89", "80-99", "*" ],
                    ["91", "90-94", "90-99", "80-99", "*" ],
                    ["92", "90-94", "90-99", "80-99", "*" ],
                    ["93", "90-94", "90-99", "80-99", "*" ],
                    ["94", "90-94", "90-99", "80-99", "*" ],
                    ["95", "90-94", "90-99", "80-99", "*" ],
                    ["96", "95-99", "90-99", "80-99", "*" ],
                    ["97", "95-99", "90-99", "80-99", "*" ],
                    ["98", "95-99", "90-99", "80-99", "*" ],
                    ["99", "95-99", "90-99", "80-99", "*" ],
                    ["100", "95-99", "90-99", "80-99", "*" ]]},
          {"name": "race", "type": "QUASI_IDENTIFYING",
            "hierarchy": [["White", "*" ],["Asian-Pac-Islander", "*" ],
                    ["Amer-Indian-Eskimo", "*" ],["Other", "*" ],
                    ["Black", "*" ]]},
          {"name": "marital-status", "type": "QUASI_IDENTIFYING",
            "hierarchy": [["Married-civ-spouse", "spouse present", "*" ],
                    ["Divorced", "spouse not present", "*" ],
                    ["Never-married", "spouse not present", "*" ],
                    ["Separated", "spouse not present", "*" ],
                    ["Widowed", "spouse not present", "*" ],
                    ["Married-spouse-absent", "spouse not present", "*" ],
                    ["Married-AF-spouse", "spouse present", "*" ]]},
          {"name": "education", "type": "QUASI_IDENTIFYING",
            "hierarchy": [["Bachelors", "Undergraduate", "Higher education",
                ↪ "*" ],
                    ["Some-college", "Undergraduate", "Higher education",
                       ↪ "*" ],
                    ["11th", "High School", "Secondary education", "*" ],
                    ["HS-grad", "High School", "Secondary education", "*"
                       ↪ ],
                    ["Prof-school", "Professional Education", "Higher
                       ↪ education", "*" ],
```

```
                        ["Assoc-acdm", "Professional Education", "Higher
                            ↪ education", "*" ],
                        ["Assoc-voc", "Professional Education", "Higher
                            ↪ education", "*" ],
                        ["9th", "High School", "Secondary education", "*" ],
                        ["7th-8th", "High School", "Secondary education", "*"
                            ↪ ],
                        ["12th", "High School", "Secondary education", "*" ],
                        ["Masters", "Graduate", "Higher education", "*" ],
                        ["1st-4th", "Primary School", "Primary education",
                            ↪ "*" ],
                        ["10th", "High School", "Secondary education", "*" ],
                        ["Doctorate", "Graduate", "Higher education", "*" ],
                        ["5th-6th", "Primary School", "Primary education",
                            ↪ "*" ],
                        ["Preschool", "Primary School", "Primary education",
                            ↪ "*" ]]},
          {"name": "native-country", "type": "QUASI_IDENTIFYING",
            "hierarchy": [["United-States", "North America", "*" ],
                        ["Cambodia", "Asia", "*" ],
                        ["England", "Europe", "*" ],
                        ["Puerto-Rico", "North America", "*" ],
                        ["Canada", "North America", "*" ],
                        ["Germany", "Europe", "*" ],
                        ["Outlying-US(Guam-USVI-etc)", "North America", "*" ],
                        ["India", "Asia", "*" ],
                        ["Japan", "Asia", "*" ],
                        ["Greece", "Europe", "*" ],
                        ["South", "Africa", "*" ],
                        ["China", "Asia", "*" ],
                        ["Cuba", "North America", "*" ],
                        ["Iran", "Asia", "*" ],
                        ["Honduras", "North America", "*" ],
                        ["Philippines", "Asia", "*" ],
                        ["Italy", "Europe", "*" ],
                        ["Poland", "Europe", "*" ],
                        ["Jamaica", "North America", "*" ],
                        ["Vietnam", "Asia", "*" ],
```

```
                    ["Mexico", "North America", "*" ],
                    ["Portugal", "Europe", "*" ],
                    ["Ireland", "Europe", "*" ],
                    ["France", "Europe", "*" ],
                    ["Dominican-Republic", "North America", "*" ],
                    ["Laos", "Asia", "*" ],
                    ["Ecuador", "South America", "*" ],
                    ["Taiwan", "Asia", "*" ],
                    ["Haiti", "North America", "*" ],
                    ["Columbia", "South America", "*" ],
                    ["Hungary", "Europe", "*" ],
                    ["Guatemala", "North America", "*" ],
                    ["Nicaragua", "South America", "*" ],
                    ["Scotland", "Europe", "*" ],
                    ["Thailand", "Asia", "*" ],
                    ["Yugoslavia", "Europe", "*" ],
                    ["El-Salvador", "North America", "*" ],
                    ["Trinadad&Tobago", "South America", "*" ],
                    ["Peru", "South America", "*" ],
                    ["Hong", "Asia", "*" ],
                    ["Holand-Netherlands", "Europe", "*" ]]},
          {"name": "workclass", "type": "QUASI_IDENTIFYING",
            "hierarchy": [["Private", "Non-Government", "*" ],
                    ["Self-emp-not-inc", "Non-Government", "*" ],
                    ["Self-emp-inc", "Non-Government", "*" ],
                    ["Federal-gov", "Government", "*" ],
                    ["Local-gov", "Government", "*" ],
                    ["State-gov", "Government", "*" ],
                    ["Without-pay", "Unemployed", "*" ],
                    ["Never-worked", "Unemployed", "*" ]]},
          {"name": "occupation", "type": "QUASI_IDENTIFYING",
            "hierarchy": [["Tech-support", "Technical", "*" ],
                    ["Craft-repair", "Technical", "*" ],
                    ["Other-service", "Other", "*" ],
                    ["Sales", "Nontechnical", "*" ],
                    ["Exec-managerial", "Nontechnical", "*" ],
                    ["Prof-specialty", "Technical", "*" ],
                    ["Handlers-cleaners", "Nontechnical", "*" ],
```

```
                            ["Machine-op-inspct", "Technical", "*" ],
                            ["Adm-clerical", "Other", "*" ],
                            ["Farming-fishing", "Other", "*" ],
                            ["Transport-moving", "Other", "*" ],
                            ["Priv-house-serv", "Other", "*" ],
                            ["Protective-serv", "Other", "*" ],
                            ["Armed-Forces", "Other", "*" ]]},
            {"name": "salary-class", "type": "SENSITIVE" }]
    }
}
```

ANEXO E – Template para o conjunto de dados *ADMISSIONS*

Na Listagem E.1 é apresentado o template para o conjunto de dados *ADMISSIONS* usado no Teste 2, descrito na Seção 5.2 do Capítulo 5, em formato JSON.

Listagem E.1: Template criado para o Teste 2.

```
{
    "name": "Teste 2 - Geração dos pseudônimos local e global com múltiplos
        ↪ provedores de dados",
    "description": "Template de teste da arquitetura",
    "organization": "ppgti",
    "contact": {"email": "teste@teste.com" },
    "privacy": {
        "kanonimity_value": 2,
        "ldiversity_value": 2,
        "max_risk_value": 0.3,
        "suppression_value": 0.9,
        "principal_identifying": "subject_id",
        "attributes": [
            {"name": "index", "type": "INSENSITIVE" },
            {"name": "subject_id", "type": "IDENTIFYING" },
            {"name": "hadm_id", "type": "IDENTIFYING" },
            {"name": "admittime", "type": "QUASI_IDENTIFYING",
              "builder": {"type": "DATE", "date_format": "yyyy-MM-dd HH:mm:ss",
                          "granularities": ["DAY_MONTH_YEAR", "MONTH_YEAR",
                              ↪ "YEAR" ]}},
            {"name": "dischtime", "type": "QUASI_IDENTIFYING",
              "builder": {"type": "DATE", "date_format": "yyyy-MM-dd HH:mm:ss",
                          "granularities": ["DAY_MONTH_YEAR", "MONTH_YEAR",
                              ↪ "YEAR" ]}},
```

```
            {"name": "deathtime", "type": "IDENTIFYING" },
            {"name": "admission_type", "type": "INSENSITIVE" },
            {"name": "admission_location", "type": "INSENSITIVE" },
            {"name": "discharge_location", "type": "INSENSITIVE" },
            {"name": "insurance", "type": "INSENSITIVE" },
            {"name": "language", "type": "IDENTIFYING" },
            {"name": "marital_status", "type": "QUASI_IDENTIFYING",
              "builder": {"type": "ORDER_BASED",
                          "levels": [{"level": 0,
                                      "groups": [{"grouping": 2, "label":
                                          ↪ "nao-casado" },
                                                 {"grouping": 3, "label":
                                                     ↪ "casado" }]}]},
              "allowed_values": ["", "SINGLE", "WIDOWED", "MARRIED", "DIVORCED" ]
                  ↪ },
            {"name": "ethnicity", "type": "IDENTIFYING" },
            {"name": "edregtime", "type": "QUASI_IDENTIFYING",
              "builder": {"type": "DATE", "date_format": "yyyy-MM-dd HH:mm:ss",
                          "granularities": ["DAY_MONTH_YEAR", "MONTH_YEAR",
                              ↪ "YEAR" ]}},
            {"name": "edouttime", "type": "QUASI_IDENTIFYING",
              "builder": {"type": "DATE", "date_format": "yyyy-MM-dd HH:mm:ss",
                          "granularities": ["DAY_MONTH_YEAR", "MONTH_YEAR",
                              ↪ "YEAR" ]}},
            {"name": "hospital_expire_flag", "type": "IDENTIFYING" }]
    }
}
```

ANEXO F – Template para o conjunto de dados *OBSERVATIONS*

Na Listagem F.1 é apresentado o template para o conjunto de dados *OBSERVATIONS* usado no Teste 3, descrito na Seção 5.3 do Capítulo 5, em formato JSON.

Listagem F.1: Template criado para o Teste 3.

```
{
    "name": "Teste 3 - Geração de um conjunto de dados anonimizado fim a fim",
    "description": "Template de teste da arquitetura",
    "organization": "ppgti",
    "contact": {"email": "teste@teste.com" },
    "privacy": {
        "kanonimity_value": 5,
        "ldiversity_value": 2,
        "max_risk_value": 0.3,
        "suppression_value": 0.3,
        "principal_identifying": "PATIENT",
        "attributes": [
            {"name": "DATE", "type": "QUASI_IDENTIFYING",
              "builder": {"type": "DATE", "date_format": "yyyy-MM-dd",
                          "granularities": ["MONTH_YEAR", "YEAR" ]}},
            {"name": "PATIENT", "type": "IDENTIFYING" },
            {"name": "ENCOUNTER", "type": "IDENTIFYING" },
            {"name": "CODE", "type": "QUASI_IDENTIFYING" },
            {"name": "DESCRIPTION", "type": "INSENSITIVE" },
            {"name": "VALUE", "type": "INSENSITIVE" },
            {"name": "UNITS", "type": "INSENSITIVE" },
            {"name": "TYPE", "type": "INSENSITIVE" }]
    }
}
```